Remaining Useful Life Estimation Using a Recurrent Variational Autoencoder *

Nahuel Costa[1][0000−0002−9189−2192] and Luciano Sánchez[1][0000−0002−2446−1915]

University of Oviedo, Gijón, Asturias, Spain

Abstract. A new framework for the assessment of Engine Health Monitoring (EHM) data in aircraft is proposed. Traditionally, prognostics and health management systems rely on prior knowledge of the degradation of certain components along with professional expert opinion to predict the Remaining Useful Life (RUL). In order to avoid reliance on this process while still providing an accurate diagnosis, a data-driven approach using a novel recurrent version of a VAE is introduced. The latent space learned by this model, trained with the historical data recorded by the sensors embedded in these engines, is used to visually evaluate the deterioration progress of the engines. High prognostic accuracy in estimating the RUL is achieved by building a simple classifier on top of the learned features of the VAE. The superiority of the proposed method is compared with other popular and state-of-the-art approaches using Rolls Royce Turbofan engine data. The results of this study suggest that the proposed data-driven prognostic and explainable framework offers a new and promising approach.

Keywords: Remaining Useful Life · Prognostics and Health Management · Recurrent Neural Networks · Variational Autoencoder

1 Introduction

Engineering maintenance and prognostics are a must in modern aircraft engines. Data is routinely collected from the engine to monitor and prevent it from operating in undesirable conditions. The knowledge of the system built into the engine and aircraft is configured to trigger alerts that highlight the need for pilot action, maintenance action or directly shut the engine down if a significant condition is encountered. Over the years, the number of variables and data collected has increased substantially which on the one hand is positive in terms of making more accurate diagnoses, but at the same time increases the difficulty to reach them since traditional strategies such as corrective maintenance of failures and scheduled preventive maintenance are becoming less capable of meeting the growing industrial demand for efficiency and reliability. On the other hand, smart

* Partially supported by the Ministry of Economy, Industry and Competitiveness ("Ministerio de Economía, Industria y Competitividad") of Spain/FEDER under grants TIN2017-84804-R and PID2020-112726-RB.

Prognostics and Health Management (PHM) technologies are showing promising capabilities for application in industries [14]. Remaining useful life (RUL) is a key metric in this regard and can be estimated from the historical data that sensors record on each trip, which is very important to improve maintenance schedules and avoid engineering, safety and reliability failures and, as a consequence, determine engine deterioration, increase engine flight time and reduce maintenance costs.

Accurate diagnosis can be achieved with model-based approaches if the degradation of the complex system is accurately modeled, some examples are Weibull distribution [1] or Eyring model [4]. The main limitation of these approaches is that they require extensive prior knowledge about the physical systems that is usually not available in practice. This is precisely why data-driven approaches have been gaining popularity in recent years, as they are able to model degradation characteristics based solely on historical sensor data from which the underlying correlations and causalities in the collected data can be modeled. In other words, knowledge can be generated from the collected data with little prior prognostic experience to infer valuable system information, such as RUL.

In this paper we propose a new Deep Learning approach for RUL estimation, based on a visual diagnosis capable of assessing the evolution of RUL. To this end, we present a novel recurrent version of a Variational Autoencoder (VAE). Aircraft data is captured over time, comprising a time series, however, VAE research is very much oriented to the field of images, and not so much to that of time series, although some work is beginning to emerge [7]. To the best of our knowledge, this is the first contribution in which a recurrent VAE is used for RUL estimation.

Besides, despite achieving very promising results, most Machine Learning models focus their efforts on predicting a number or a label, leaving aside how they got there [3]. By making use of the internal representation learned by the VAE we can elaborate a strong explanatory component, since a simple prediction may not be informative enough to determine engine deterioration, thus giving insight into the state of an engine with a simple look to a self-explanatory map.

2 RUL Estimation

In the last decade, the relationship between the use of monitored system data and the RUL of engines has gained the attention of data-driven prognostic models. Especially, the use of neural networks has had a great impact given that they have the advantage of learning to model highly nonlinear, complex and multidimensional systems without experience in the physical behavior of the system. In this sense, there are works such as [12], where the authors applied multilayer perceptrons (MLP) for estimating the RUL of laboratory-tested bearings. In addition, some researchers have integrated fuzzy logic to capture more information for EHM [5] [9]. It is also worth mentioning works like [11] and [6] where Gradient Boosting trees and Support Vectors are applied for engine RUL prediction.

More recently, in this field, as in many other areas such as image or speech recognition, the application of Deep Learning models has been gaining ground over the years for RUL estimation as the raw data obtained from machine health monitoring share a high dimensionality similar to that of image processing studies. There are clearly two trends: the application of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). In the first group, works as [8] explore RUL estimation using different configurations of CNNs minimizing prior knowledge about prognostics and signal processing. Regarding RNN, the most common architecture that can be found in the literature are LSTMs [10] and in the last years, Bidirectional LSMTs [13] are beginning to gain importance as their ability to make full use of the sensor date sequence in bi-direction seems to have promising results.

Again, as discussed above even if these models achieve good results, in the end, these systems will be used by people outside Machine Learning and what they will be looking for is a high-quality interpretation of the data. One possible way to provide this is to establish a Representation Learning approach since, unlike others, the performance of models following this approach depends directly on internal representations, which in turn can be leveraged in favor of a better understanding of the problem itself. In this direction, we propose applying a VAE.

VAEs are designed to reconstruct the input data while at the same time learn a compressed representation of it, the so-called latent space. That compression depends on a probability distribution, causing the data to be organized in a continuous space, i.e. two nearby points in the latent space should give similar contents when reconstructed. This also means that similar data are located close together in the latent space, forming different clusters depending on their underlying nature.

The framework we propose relies then on a new recurrent VAE to model the time complexity of the engine data. The fact that it has a recurrent input offers the possibility to feed the model each time a new data sample becomes available, allowing online training. The VAE learns different degradation stages to the point of being able to place in the latent space, which can be understood as a two-dimensional map, aircraft with similar RUL values in the vicinity. This is used to, given data of a new aircraft, place it on the map according to the RUL it has, thus offering a simple and intuitive diagnosis.

3 Model architecture

Variational autoencoders consist of 3 parts: an encoder network, a decoder network and a latent space. The encoder learns to compress the data to the latent space from which the decoder can generate new samples. The latent space is described by the parameters learned by the encoder that initialize the probability distribution to which the data belong, so that the decoder can not only reconstruct the input data, as conventional autoencoders do, but can also generate new samples from that distribution. The loss function introduces the Kullback-Leibler

divergence, which measures how much one probability distribution diverges from another, to learn the above-mentioned parameters. The reconstruction error between the original input and the output of the decoder is also included. In the end, all together allows the model to produce a latent space in which similar data will be located close to each other and also enables new data to be sampled from points that do not belong to the original data, thus having a generative model.

Given the way the model works, the workflow followed for this problem is quite simple: a VAE is trained with data from Turbofan engines to learn a simplified representation. Thus, the learned encoder acts as a feature extractor by projecting onto the latent space the data according to its properties, which are different stages of deterioration in the engines. This section explains how this extraction, can be exploited to create the diagnostic map we are pursuing. Emphasis is also placed on the recurrent architecture proposed to deal with the time series, as well as how the latent features give rise to perform other tasks such as classification or prediction.

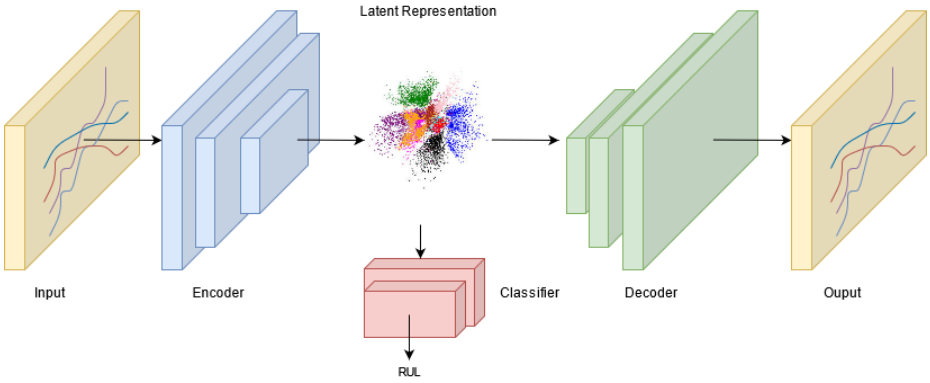

Fig. 1: Network structure of the proposed method. The blue and green blocks are the encoder and decoder respectively and the red blocks refer to the linear classifier.

Encoder as a feature extractor In a VAE the training process is regularized to avoid overfitting and to ensure that the latent space has good properties that enable the generative process. To obtain these properties, the encoder must be able to map the data in the latent space in such a way that similar data are close. This allows not only that the decoder can reconstruct the data efficiently but also that it can generate new data from a point in the latent space that does not correspond to the encoding of any training data.

A VAE, given an input, tries to find a latent vector that is capable of describing it and at the same time has the instructions to generate it again. The

process can be described as: $p(x) = \int p(x|z)p(z)dz$. Given that the integral of this formula is intractable due to the continuous domain of z, the variational inference is needed via the lower bound of the log likelihood, $\mathcal{L}_{vae}$,

$$\mathcal{L}_{vae} = E_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})). \tag{1}$$

The first term is the reconstruction of x that tends to make the coding-decoding scheme as efficient as possible by maximizing the log-likelihood $\log p_\theta(\mathbf{x}|\mathbf{z})$ with sampling from $q_\phi(\mathbf{z}|\mathbf{x})$, modeled by a neural network (the encoder) whose output is the parameters of a multivariate Gaussian: a mean and a diagonal covariance matrix (the latent space). That is to say, the main goal of the encoder is to map the input data into a lower-dimensional space, acting as a feature extractor. The second term tends to regularise the organization of the latent space by causing the distributions returned by the encoder to approach a standard normal. It regularises the latent variables (represented by z) by minimizing the KL divergence between the variational approximation and the prior distribution of z.

Based on the representation learned by the encoder, the data, x, is sampled from the conditional probability distribution p(x—z). For generative purposes, this regularization in the latent space is very effective in facilitating random sampling and interpolation for the creation of new data. This is why VAEs are understood to be generative models and precisely it is the most widespread application in the literature. Nevertheless, we place our efforts not on generating new aircraft data but on diagnosing it instead, therefore after training the decoder is not used anymore.

As stated in the introduction, most recent studies make use of recurrent networks to model the time complexity of historical aircraft data. Among the different types of RNNs that can be found, LSTM networks are the most prominent. These networks process data from backward to forward preserving the information of the past through hidden states. However, Bidirectional LSTM networks are in high demand because they provide not only information about the past but also about the future: data is first processed from past to future and then from the future to the past, thus preserving the information from both periods. This is very valuable because the network knows what the data may look like in its future stages, which helps it to understand what kind of information to predict (different stages of engine degradation).

All things considered, we decide to implement the VAE with Bidirectional LSTM networks. In this way, the encoder approximates the Gaussian distribution $p_\theta(\mathbf{z})$ by feeding the output into two linear modules to estimate its mean and covariance. The compression of the input data results in a two-dimensional latent space dominated by the axis represented by the mean and the variance of the approximated distribution. Figure 1 shows the pipeline followed for applying this framework for RUL estimation: the input and output are the same and in the middle it is expected that engines are grouped in different clusters in the latent space according to their features, depicting a simpler representation of their nature. Furthermore, we add a classifier on top of the learned features in

order to explicitly report which RUL is the one that best represents each engine that is fed to the model.

Diagnostic map The diagnostic tool introduced in this study is a color-coded map that displays the actual state of the engine and the speed of change from healthy to deteriorated. Once the VAE is trained with the engine´s data, every input can be compressed into the latent space in terms of the mean and variance of the learned approximated distribution. This information can be projected onto a map as shown in Figure 2. Every projected point is a sample from the dataset and is colored according to their corresponding RUL: aircraft with low RUL values are painted in red while aircraft with high RUL values are painted in green. It can be seen there is a clear progression in the colors between points since units with no signs of deterioration are located in the upper part of the map (greater values of RUL) while the most deteriorated ones are located in the lower part (lower values of RUL). This representation can be leveraged later on: when new unseen units are used as inputs, the encoder will place them according to their features, giving information about their RUL depending on the proximity to other nearby points that are labeled. That is why it is considered explainable, because the method itself explains the status of each engine. The following section provides further details on the interpretation of this map.

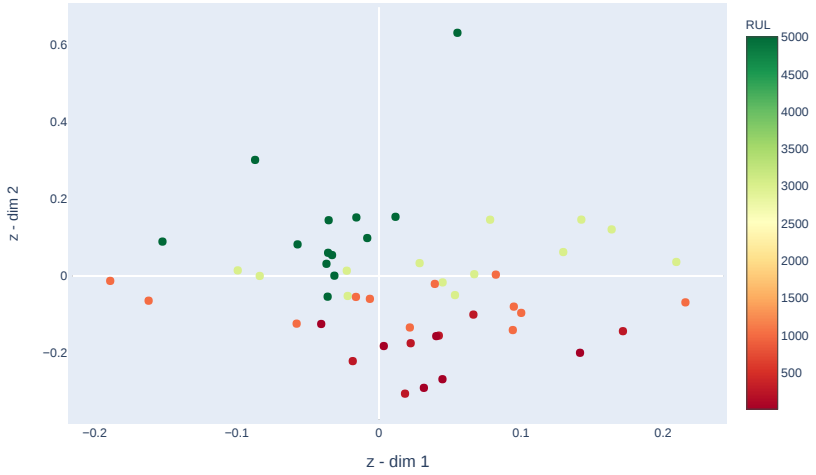

Fig. 2: Latent representation learned by the encoder.

4 Experimental Study

RUL estimation is an actual engineering problem posed by Rolls-Royce therefore, in this paper, the proposed method is evaluated on EHM data from Rolls-Royce Turbofan engines. This data contains multi-variate temporal information obtained from several built-in sensors. Each time step represents the values of the variables collected for a single flight. Each engine unit may start with different starting conditions all of which, albeit, are considered to be healthy. The expected behavior of an aero-engine is to degrade as time progresses, however, the fact that this degradation is not linear hampers the estimation of the health state, that is the RUL. Anticipating the breakpoint before failure is key to prevent potential problems in the future, thus expanding the lifetime of the engines. Particularly, information is available on the state of the turbine and compressor of the aircraft at the end of the measurement, which allows us to create a dataset that reflects the states through which an aircraft may go through. Based on this, we aim is to generate valuable knowledge that will allow us to estimate the RUL of a new aircraft given its flight history.

4.1 Data Pre-Processing

As usual, data capture has its limitations, as a result, in order to prepare a consistent dataset it has to be subjected to a purging and pre-processing process. To begin with, there are quite a few time steps where data is missing, therefore as long as these rows can be dealt with, an imputation is applied based on the column average, i.e. the average of the values captured by the sensor.

Natural factors such as wind, number of passengers or a change of trajectory may cause noticeable peaks in each signal which ultimately leads to noise. As a consequence, it is necessary to apply a smoothing that minimizes the effects of this noise and prioritizes the trend of each signal. For this purpose, an exponential smoothing with an alpha value of 0.4 is applied.

Anomalies, mainly points with spurious values, must be carefully removed because as a final step before passing the data to the network, a normalization is applied to the whole dataset to ease the network to deal with reasonable values and these anomalies can significantly affect the range of the data.

Series length: the useful life of a new aircraft is expected to be at least five thousand cycles. Nonetheless, among the aircraft for which information is available, there are varying lengths due to the fact that there are measurements for different life stages, therefore there are units that may end in eighty cycles, while others may end in seven thousand. It is known that aircraft with less than one thousand cycles had failures linked to different problems other than turbine and compressor, and so it is decided to get rid of them because they are random failures that would add nothing but noise to the dataset. Each aero-engine goes to the workshop every one thousand flights (cycles) and the condition of the turbine and compressor are recorded independently by the mechanics. Precisely, this time window is used to determine the length of the series to be received by the network as input. For this purpose, as one thousand cycles are almost

intractable for a neural net, each sequence is transformed by taking each point as the average of the next ten, thus reducing the size of the sequence ten times while maintaining the morphology of the signal.

Finally, it is not so much the values that each signal takes that matter, but rather the increases or decreases that they undergo. As an example, even if two units have different pressures in the turbine, if the pressure has been increasing over the cycles in both units, it means that both have suffered wear in this component, therefore it is more informative to take the derivative of the signal instead of the signal itself. Figure 3 reflects the transformation carried out by this pre-processing for a random aircraft.

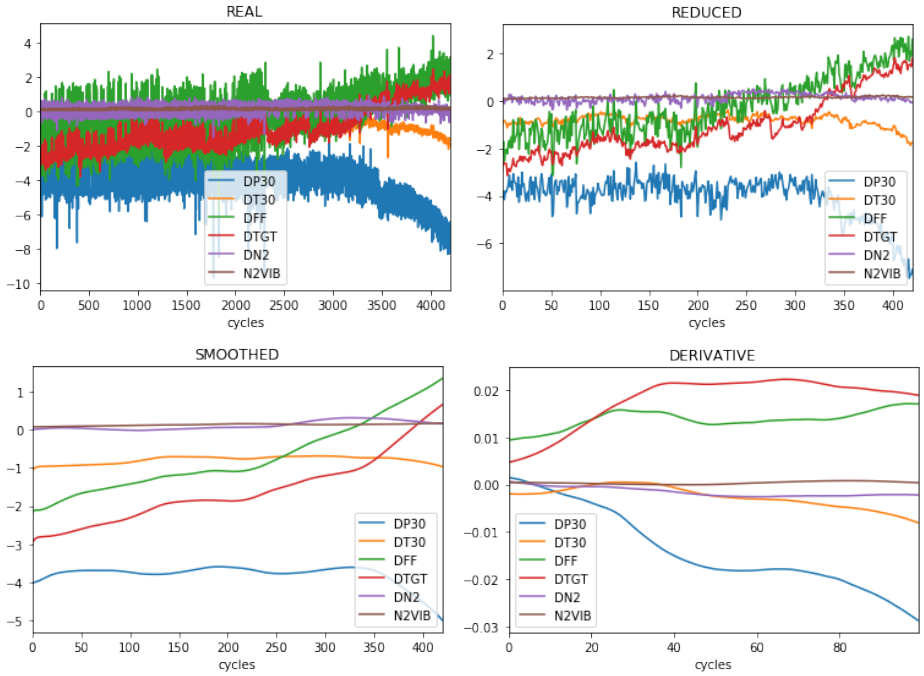

Fig. 3: Every aircraft in the dataset undergoes a transformation that mainly involves the reduction of the number of time steps, noise reduction and calculation of the derivatives of the signals.

4.2 Illustrative Example

To illustrate how the proposed model works, an example that can be understood as a visual fleet diagnosis is presented below. During training, the network learns different degradation patterns which leads the encoder to project the units into the latent space according to their degradation, maintaining coherence in the

distances between healthy and compromised engines as described in Figure 2. This projection is reused as a basis to find out, given undiagnosed units, how their deterioration evolves as the number of trips (cycles) increases. Figure 4 pictures this idea: 8 airplanes have been chosen to project their state into the latent space in four different time steps: t=0 would correspond to feeding the network with the data corresponding to the cycles from 0...1000, t=1000 from 1000 to 2000 and so on until t=3000. Leaving the latent projection obtained in train gives us some insight into the progression of the health status of these units: The latent projection of samples s2, s3, s5 and s8 during all the time steps shown remain over the upper left quadrant, next to other aircraft with similar characteristics, RUL around five thousand and so with no signs of near degradation. On the contrary, there is a clear progression in samples s1, s3, s6 and s7, which move slightly downward and to the right at t=1000 and t=2000 to finally at t=3000 be placed together with units close to their end of life (low values of RUL), thus obtaining an accurate and explainable diagnosis beyond a possible label indicating the predicted health.

In the presented figure only four time steps have been selected to show the update of the engine status according to the data input, however, it is remarkable that once passed the barrier of the first thousand cycles in every posterior trip this update can be done thanks to the fact that we are using recurrent networks and this is where the interest really lies because in the end this can be used as a diagnostic tool: As long as the engine projection remains in the healthy range, its state will be considered positive; on the contrary, if the projection moves towards the red zone, this can be a clear sign of deterioration, information that will be used by the mechanics to make a decision regarding its monitoring, either to make it more exhaustive or to take the aircraft to the workshop for a more complete check-up, just to name a few alternatives. This translates into an extension of the life of these engines by being able to anticipate the break-point where severe deterioration may occur.

4.3 Numerical Results

In this section, we demonstrate that our framework can compete with other modern approaches for time series for the case at hand. It is important to note that RUL estimation is typically a prediction problem but the RUL information available to us are labels that determine the health status of the engine. These labels can be classified into 5 groups from best to worst condition: "Good", "Good To Normal", "Normal", "Normal To High" and "High". It is also understood that each tag corresponds to an approximate number of RUL, being Good$\approx$5000 cycles, Good to Normal$\approx$3000 cycles, Normal$\approx$1000 cycles Normal To High$\approx$250 cycles and High$\approx$10 cycles. For this reason, we chose to compare our method with models widely used in the literature: MLP, CNN, LSTM and Bidirectional LSTM, but changing their last layer so that instead of predicting a number they predict a class and thus become a classification problem like the one we have.

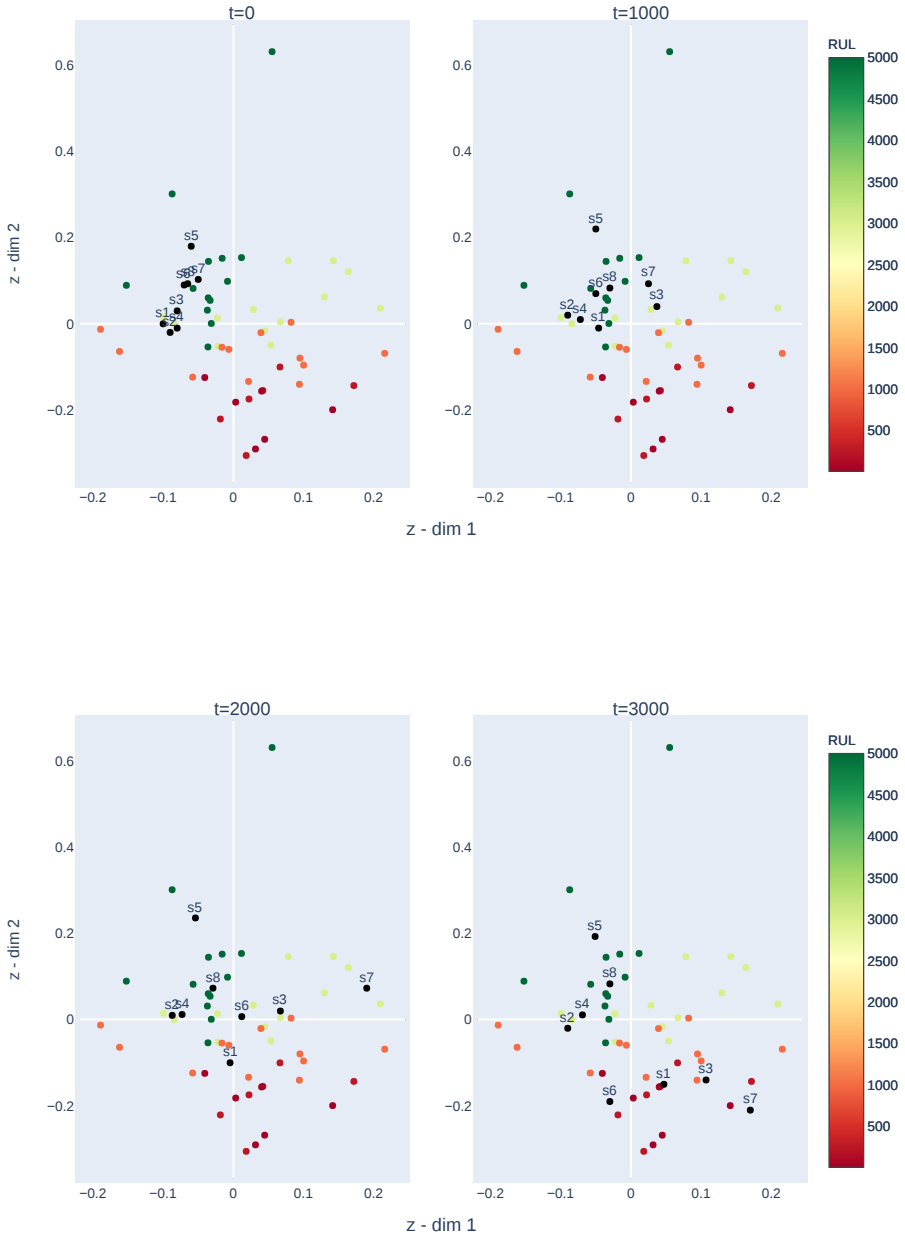

Fig. 4: RUL evolution in a few selected aircraft. As the cycles progress, the fastest degrading aircraft are placed in the zone occupied by aircraft with similar deteriorations.

Table 1 shows the performance of the different models for each class in terms of accuracy. Each entry in the table is the number of times an engine in a class was recognized by each model for the appropriate class. In addition, to illustrate the performance of each method, the ranking calculated by the Friedman method (ranking by range) for each sub-dataset and the resulting averaged ranking are included.

It can be seen that the best classifier in terms of accuracy is ours, labeled as RVAE. To extend the comparison between the different methods, post-hoc tests were carried out to detect significant differences in pairs between all the classifiers as recommended in [2]. If the significance test yields a p-value lower than a predefined threshold (usually 0.05), then the difference is considered significant, therefore one model is declared superior to another. As a result, our method is the only one that yields a p-value of less than 0.05 (0.0357) for the comparison with MLP. This means that the other methods are not statistically superior to MLP, which makes us value the performance of our method. Additionally, it should be noted that the baseline methods we present do not include any representation of the data, but simply predict the class to which each sample belongs, while interpretability of black-box models, like the one we provide, can present predicted information in a more illustrative way than just a numerical or categorical result.

Table 1: Accuracy of the different classifiers, 5 types of RUL.

	Accuracy				
	MLP	CNN	LSTM	BiLSTM	RVAE
High(50)	0.166(5)	0.330(4)	0.500(2.5)	0.500(2.5)	0.666(1)
NormalToHigh(250)	0.500(4.5)	0.666(1.5)	0.666(1.5)	0.500(4.5)	0.666(1.5)
Normal(1000)	0.833(3)	0.100(1)	0.750(4)	0.916(2)	0.666(5)
GoodToNormal(3000)	0.692(3.5)	0.615(5)	0.692(3.5)	0.846(1.5)	0.846(1.5)
Good(5000)	0.692(5)	0.846(4)	1.000(1.5)	0.923(3)	1.000(1.5)
Summary Results					
Accuracy	0.64	0.74	0.76	0.78	0.80
Average rank	4.2	3.1	2.6	2.7	2.1

5 Concluding Remarks

We have introduced a recurrent VAE architecture based on Bidirectional LSTMs to create a graphical map that describes the condition of engine fleets. The diagnostic tool learns a 2D representation of engine data with different degradation stages to, given a new engine, project its representation near engines with similar deterioration patterns. This allows providing an efficient diagnostic tool on

the state of health of the engines without prior knowledge of its physical nature. In addition, the lightness of the model and its recurrent nature would allow incorporating the model as a diagnostic system on any hardware with limited computational capabilities and at the same time updating the learned patterns as more data becomes available, thus coping with the non-stationarity of the data distribution.

References

1. Ali, J.B., Chebel-Morello, B., Saidi, L., Malinowski, S., Fnaiech, F.: Accurate bearing remaining useful life prediction based on weibull distribution and artificial neural network. Mechanical Systems and Signal Processing **56**, 150–172 (2015)
2. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine learning research **7**(Jan), 1–30 (2006)
3. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
4. Jouin, M., Gouriveau, R., Hissel, D., Péra, M.C., Zerhouni, N.: Degradations analysis and aging modeling for health assessment and prognostics of pemfc. Reliability Engineering & System Safety **148**, 78–95 (2016)
5. Khawaja, T., Vachtsevanos, G., Wu, B.: Reasoning about uncertainty in prognosis: a confidence prediction neural network approach. In: NAFIPS 2005-2005 Annual Meeting of the North American Fuzzy Information Processing Society. pp. 7–12. IEEE (2005)
6. Khelif, R., Chebel-Morello, B., Malinowski, S., Laajili, E., Fnaiech, F., Zerhouni, N.: Direct remaining useful life estimation based on support vector regression. IEEE Transactions on industrial electronics **64**(3), 2276–2285 (2016)
7. Li, L., Yan, J., Wang, H., Jin, Y.: Anomaly detection of time series with smoothness-inducing sequential variational auto-encoder. IEEE transactions on neural networks and learning systems (2020)
8. Li, X., Ding, Q., Sun, J.Q.: Remaining useful life estimation in prognostics using deep convolution neural networks. Reliability Engineering & System Safety **172**, 1–11 (2018)
9. Martínez, A., Sánchez, L., Couso, I.: Engine health monitoring for engine fleets using fuzzy radviz. In: 2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). pp. 1–8. IEEE (2013)
10. Miao, H., Li, B., Sun, C., Liu, J.: Joint learning of degradation assessment and rul prediction for aeroengines via dual-task deep lstm networks. IEEE Transactions on Industrial Informatics **15**(9), 5023–5032 (2019)
11. Singh, S.K., Kumar, S., Dwivedi, J.: A novel soft computing method for engine rul prediction. Multimedia Tools and Applications **78**(4), 4065–4087 (2019)
12. Tian, Z.: An artificial neural network method for remaining useful life prediction of equipment subject to condition monitoring. Journal of Intelligent Manufacturing **23**(2), 227–237 (2012)
13. Zhang, A., Wang, H., Li, S., Cui, Y., Liu, Z., Yang, G., Hu, J.: Transfer learning with deep recurrent neural networks for remaining useful life estimation. Applied Sciences **8**(12), 2416 (2018)
14. Zhao, Z., Liang, B., Wang, X., Lu, W.: Remaining useful life prediction of aircraft engine based on degradation pattern learning. Reliability Engineering & System Safety **164**, 74–83 (2017)