

# Learning Representations with End-to-End Models for Improved Remaining Useful Life Prognostic

Alaaeddine Chaoub<sup>1</sup>, Alexandre Voisin<sup>2</sup>, Christophe Cerisara<sup>3</sup>, and Benoît Iung<sup>4</sup>

<sup>1,3</sup> *Université de Lorraine, CNRS, INRIA, LORIA, France*  
*alaaeddine.chaoub@loria.fr*  
*christophe.cerisara@loria.fr*

<sup>2,4</sup> *Université de Lorraine, CNRS, CRAN, France*  
*alexandre.voisin@univ-lorraine.fr*  
*benoit.iung@univ-lorraine.fr*

## ABSTRACT

Remaining Useful Life (RUL) of equipment is defined as the duration between the current time and the time when it no longer performs its intended function. An accurate and reliable prognostic of the remaining useful life provides decision-makers with valuable information to adopt an appropriate maintenance strategy to maximize equipment utilization and avoid costly breakdowns. In this work, we propose an end-to-end deep learning model based on multi-layer perceptron and long short-term memory layers (LSTM) to predict the RUL. After normalization of all data, inputs are fed directly to an MLP layers for feature learning, then to an LSTM layer to capture temporal dependencies, and finally to other MLP layers for RUL prognostic. The proposed architecture is tested on the NASA commercial modular aero-propulsion system simulation (C-MAPSS) dataset. Despite its simplicity with respect to other recently proposed models, the model developed outperforms them with a significant decrease in the competition score and in the root mean square error score between the predicted and the gold value of the RUL. In this paper, we will discuss how the proposed end-to-end model is able to achieve such good results and compare it to other deep learning and state-of-the-art methods.

## 1. INTRODUCTION

Maintenance is a crucial and costly activity: studies show that depending on the industry, between 15 and 70 percent of total production costs originate from maintenance activities (Krupitzer et al., 2020). Nevertheless, maintenance has always been a main factor in a company's ability to be competitive in performance and deliver pricely and high quality

Alaaeddine Chaoub et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

product. To that end, Prognostics and Health Management (PHM) is receiving a lot of attention in recent years thanks to its ability to drive maintenance in a more optimal way. Prognostic or remaining useful life estimation, on top of diagnosis and fault detection, remains the core topic in PHM, as it provides PHM the ability to anticipate fault and provide relevant information on failure time to maintenance decision-makers.

Numerous prognostic algorithms for RUL estimation have been reported in the literature. The approaches for this challenging task can roughly be classified into two main approaches: Model-based approaches (Cauchi, Macek, & Abate, 2017) (Yuan, Jiang, & Liu, 2013), and Data driven approaches (W. Zhang, Yang, & Wang, 2019). The focus of this work is on RUL prognostic based on a data driven approach, more precisely deep learning models.

A classical way to tackle prognostics, due to the variety of PHM problems and cases, relies on a first stage of feature selection or engineering before introducing them into machine learning (ML) models. One of the main feature of deep learning (DL) models is their ability to handle a large number of inputs and leverage complex correlation patterns among them. Surprisingly, when screening the literature, DL models used for prognostic are often preceded by such a feature selection stage. The originality of this work is to propose an end-to-end DL model that start with an MLP in the first layers for predicting the RUL using raw normalized input directly, an approach that reduce feature selection expenditure and deal with complex datasets with multiple operating conditions. We apply the model on the publicly available data sets C-MAPSS (Saxena, Goebel, Simon, & Eklund, 2008) that describes the operational history of simulated aircraft turbofan engines. The result shows that, despite its simplicity, our model performs better than state-of-the-art approaches on this dataset.

The rest of this paper is structured as follows. Section 2 introduces related works on RUL prognostic. Section 3 describes the proposed RUL prognostic architecture. Section 4 highlights the effectiveness of the proposed method by comparing the results with other popular methods. Finally, conclusions and discussion are provided in section 5.

## 2. RELATED WORK

RUL prognostic methods based on artificial intelligence approaches are attracting increasing attention, due to their ability to model highly nonlinear, complex and multidimensional systems. A number of deep learning (DL) techniques have been deployed in order to learn the mapping from monitored system data to their associated RUL. In this section, we focus on recent state-of-the-art work applying DL model to the C-MAPSS Dataset (Saxena et al., 2008). For a general overview, (L. Zhang et al., 2019) review deep learning approaches applied to Prognostics and Health Management.

Recurrent neural networks are often used for problems involving time series data, because of their ability to process information over time. (Zheng, Ristovski, Farahat, & Gupta, 2017) proposed a model based on multiple LSTM layers followed by a feed forward neural network that maps the input features to the predicted RUL, which is a standard deep learning architecture to deal with sequence data. (Huang, Huang, & Li, 2019) used a similar architecture but with Bidirectional LSTM cells in order to capture relevant information from both directions over time.

Convolutional neural networks are also often used when it comes to dealing with time series data thanks to their ability to model correlations in a temporal window around every time frame. (Li, Zhao, Zhang, & Zio, 2020) proposed an architecture based on multi-scale convolution kernels to capture information at different scales in the network, which helps to learn temporal features from different sequence sizes.

Trying to combine the advantages of different techniques, a hybrid architecture was proposed in (Al-Dulaimi, Zabihi, Asif, & Mohammadi, 2019), integrating a deep LSTM and a deep CNN followed by a Multi-Layer Perceptron (MLP) to improve the prognostic performance.

Semi supervised learning was employed in (Hou, Xu, Zhou, Yang, & Fu, 2020), where deep convolutional generative adversarial networks are used. The generator is an auto-encoder that tries to reconstruct the input signals, while the discriminator tries to distinguish the true data from the false ones. After this pre-training phase, the encoded features are used as input to an LSTM/MLP model for RUL prognostic.

These related works and other publications suggest that various deep learning architectures have been proposed and tested for RUL prognostic. Nevertheless, we argue that most of them may suffer from two potential issues. First, including a

first stage of feature selection may harm the subsequent modeling process because it may genuinely discard relevant information and weak signals that may be hidden and overlooked by experts. Second, simple but well-designed neural networks often prove to match the performances of more complex deep learning architectures, whose hyper-parameters are more difficult to tune, eventually requiring many time-consuming and energy-hungry experiments, which presents a technical barrier for industrial applications.

## 3. PROPOSED MODEL ARCHITECTURE

To overcome the two aforementioned drawbacks, the proposed model has an MLP-LSTM-MLP architecture trained in an end-to-end manner for RUL prediction. This architecture has also been proposed in (An, Li, Wang, & Jiang, 2020) and has shown promising results for diagnostic applications.

Long Short Term Memory networks (LSTM) address the gradient vanishing problems in Vanilla recurrent networks by introducing new gates that allow for better control of gradient flow, and better preservation of long-term dependencies, which is needed in applications like RUL prognostic. However, LSTM cells are designed to capture time dependencies but they do not have the capacity to handle complex feature processing, which has led other works in the literature to perform this task manually before the learning phase. Conversely, MLP are well fitted to perform such a task. We thus propose to feed all of the raw inputs into an MLP before the LSTM layers. The MLP will be in charge of processing the raw inputs and learning a good representation of each time frame, while the LSTM shall capture the dependencies through time of frame sequences. Then, a final regression head, composed of another MLP, predicts the RUL from these temporally smoothed representations.

Figure 1 shows the proposed architecture: each input vector  $x_t$  is processed by a first MLP with 3 layers, and the resulting sequence of feature vectors is processed by a single LSTM layer. The output of each LSTM cell is finally passed to another MLP with 3 layers that outputs a scalar  $y_t$  that represents the predicted RUL. The weights of the features-MLPs are shared across all time steps, which is convenient when working with variable length sequences.

## 4. EXPERIMENTS

### 4.1. The C-MAPSS Dataset overview

The commercial modular aero-propulsion system simulation (C-MAPSS) is a turbofan engine simulation environment from NASA that provides access to health, control, and engine parameters through a graphical user interface (GUI). The C-MAPSS dataset (Saxena et al., 2008) is generated using the simulation program by monitoring the degradation of multiple Turbofan engines.

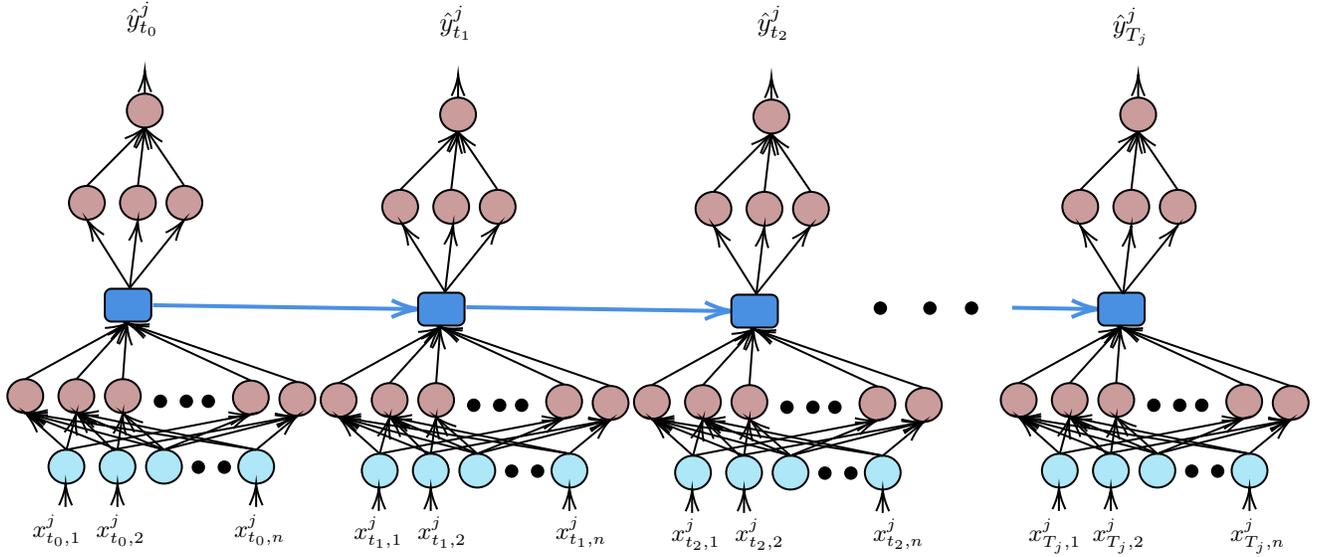


Figure 1. Architecture of the proposed model: it takes as input a complete sequence  $j$  of raw sensor values, encoded as a tensor  $x^j$  composed of  $T_j$  time frames with  $n$ -dimensional observations each. At training time, this sequence ranges from the first observed time frame  $t_0$  to the last  $T_j$  just before the Turbofan  $j$  halts. At test time, a single forward pass is performed and the RUL  $\hat{y}_t$  is predicted at every time step given the previous observations. To simplify the diagram, only one layer has been drawn for the MLPs.

The data set is divided into four sub-data sets (from FD001 to FD004) with varying number of operating conditions and fault modes (see Table (1)). Each sub-data set is further divided into training and test subsets. The training set is composed of input time series which are assumed to go on until failure. In the test set, time series are truncated arbitrarily and the objective is to estimate the number of remaining operational cycles before the system failure occurs.

The Turbofan Engine Degradation Simulation data sets (Table (1)) are widely used by academics and scholars to test prognostic algorithms.

Table 1. The C-MAPSS Datasets

DATASET	FD001	FD002	FD003	FD004	
Nb Train trajectories	100	260	100	249	
Nb Test trajectories	100	259	100	249	
Nb Operating conditions	1	6	1	6	
Nb Fault modes	1	1	2	2	
trajectories length distribution	Max	360	378	525	543
	Mean	206	206	247	245
	Min	128	128	145	128

The data contains multivariate time series, which correspond to 24 sensors measurements taken at each operating cycle of a particular simulated turbofan engine.

## 4.2. Corpus preparation

A gold RUL value for every cycle (or equivalently, time frame) is computed on the training set. This is achieved by assuming that the RUL decreases linearly over time.

In practice, the degradation of a turbofan engine may be considered as negligible at the beginning of its use, and increases as the component approaches the end of its life. To better model the changes in the remaining useful life as function of time with respect to the non linearity of the degradation, we adopt a strategy that is often used in related works, and model the RUL with a piece-wise linear function shown in Figure (2), which limits the maximum RUL to a constant value and then begins linear degradation after a certain time (Heimes, 2008). The most common maximal RUL values used among the works in the literature are 125 and 130 (Al-Dulaimi et al., 2019) (Zheng et al., 2017). In the following, we choose a maximum RUL of 130.

The gold RUL is defined in Eq. 1 as:

$$Gold\ RUL = \begin{cases} 130 & \text{if } True\ RUL \geq 130 \\ True\ RUL & \text{if } True\ RUL < 130 \end{cases} \quad (1)$$

This maximum value to predict shall facilitate training the model, because it reduces the range of useful values to predict. However, it may also prevent the model from being able to predict very long RUL values, but we expect this case to rarely occur in the test set.

Another constraint on data comes from the choice of a deep

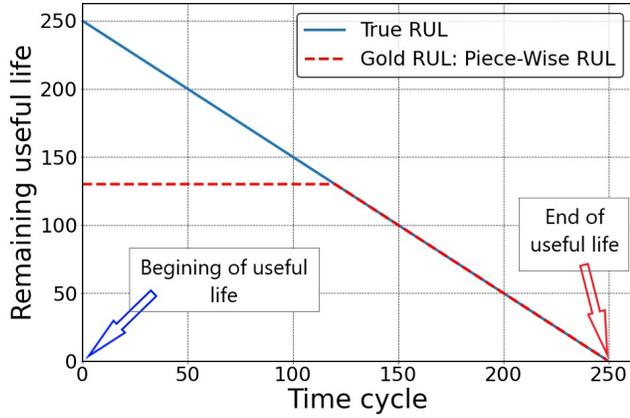


Figure 2. True RUL vs Gold RUL of one trajectory with a length = 250 (Piece-wise maximum RUL is 130 time cycles).

learning model, which trains better when all inputs and outputs are normalized. Hence, we normalize both inputs and outputs in the range  $[0, 1]$  using Equation 2 for the inputs, and similarly for the RUL values:

$$x_{t,i}^j = \frac{v_{t,i}^j - \min_{t,j}(v_{t,i}^j)}{\max_{t,j}(v_{t,i}^j) - \min_{t,j}(v_{t,i}^j)} \quad (2)$$

Where  $v_{t,i}^j$  is the value of the  $i$ th sensor at time  $t$  from engine  $j$ , and  $x_{t,i}^j$  is the corresponding normalized value. In order to compute the min and max values,  $t$  and  $j$  range across the entire data set, i.e. all trajectories of all turbofans are used. Normalization helps to stabilize training of the network parameters, speeds up convergence of gradient descent and reduces the risk of getting stuck in local optima.

The normalized data is directly fed to the network, without any feature engineering or selection. Therefore, no prior expertise on Turbofans or signal processing is required for the proposed method.

### 4.3. Performance metrics

In the PHM context, it is generally desirable to predict failures as early as possible. Therefore, the scoring function that will be used to evaluate the performances of the models penalizes more the errors that predict a RUL too late than too early, and it is given by Eq. 3 as proposed in the original C-MAPSS evaluation campaign (Saxena et al., 2008):

$$S = \sum_{j=1}^N \sum_{t=1}^{T_j} s_t^j, \text{ Where:} \quad (3)$$

$$s_t^j = \begin{cases} e^{-\frac{d_t^j}{13}} - 1 & \text{for } d_t^j < 0 \\ e^{\frac{d_t^j}{10}} - 1 & \text{for } d_t^j \geq 0 \end{cases}$$

Where  $N, T_j$  are respectively the number, length of the trajec-

tories, and  $d_t^j = \hat{y}_t^j - y_t^j$  (Predicted RUL - Gold RUL)

For the sake of comparability with other literature results, we will use the scoring function of the challenge plus the Root Mean Square Error (RMSE) (Eq. (4)):

$$RMSE = \sqrt{\frac{1}{N} \sum_{j=1}^N \frac{1}{T_j} \sum_{t=1}^{T_j} d_t^j{}^2} \quad (4)$$

Figure (3) shows how both metrics penalize errors in detail. However, the main objective is to achieve the smallest value possible for both.

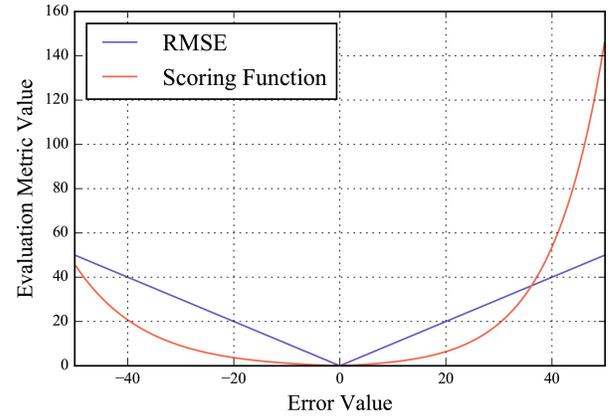


Figure 3. Comparison between the scoring function and RMSE with respect to different error values.

### 4.4. Model Training and evaluation

In order to choose the hyper-parameters of the model, we split the training set into a training subset and a validation subset, based on the ID of the engines. The original test set is reserved for final evaluation.

Our model does not use neither fixed length sequences nor truncation nor padding, and each training sample is a full time series of one turbofan engine from its first cycle until failure. Henceforth, different samples have variable sequence length. We used 75% of the turbofans run to failure trajectories as training subset, and 25% as validation subset.

Both metrics may be used as loss functions for training. Preliminary experiments show that both the score and the RMSE give similar results. So we decided to work with the RMSE because the training process is faster.

Hyper-parameters have been tuned manually with a few trials and errors on the validation set. The hyper-parameters to be optimized are the learning rate, the number of layers in the input and output MLPs, the number of LSTM layers, the number of neurons/cells in each layer, the activation functions, the dropout percentages and the optimizer. The best hyper-parameters found for the proposed model are listed in

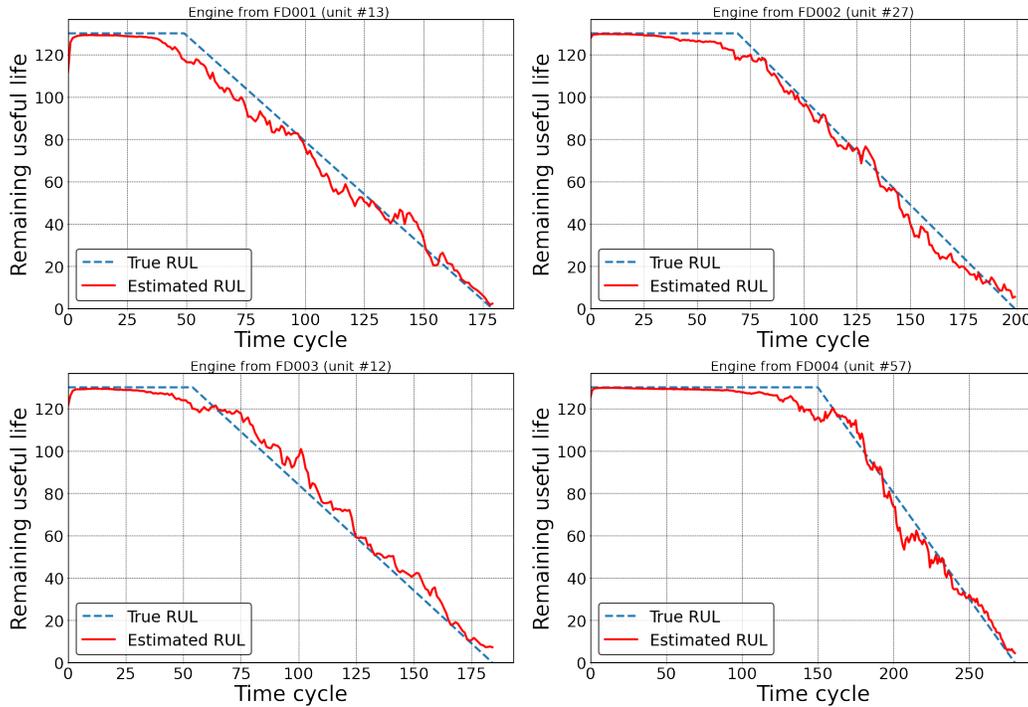


Figure 4. Different examples of RUL prognostic for engine units from the validation samples of each data set.

Table (2).

Table 2. Hyper-parameters of the proposed model

Hyper-parameter	Value
Learning Rate	0.0001
Number of MLP layers before LSTM	3
Number of neurons in MLP layers	100/50/50
Number of LSTM layers	1
Number of LSTM cells	60
Number of MLP layers after LSTM	3
Number of neurons in MLP layers	60/30/1
Activation function for MLP layers	Tanh()
Batch size	5
Dropout percentage	0%

#### 4.5. Results and Discussion

Because of random initialization, the optimized model parameters values may vary across different training runs. We thus evaluate the model’s performances across 10 runs, and the mean values and standard deviations are given in Table (3) for the four data sets.

Figure (4) shows the predicted RUL vs the gold value of the RUL for four trajectories from the validation subset. We see that the proposed model can follow degradation patterns even in complex data sets as FD002 and FD004 with 6 operating conditions. Thanks to our end to end learning approach, the

Table 3. prognostic performance of the proposed model.

DATASET	FD001	FD002	FD003	FD004
RMSE	13.26 ± 0.57	12.49 ± 0.28	13.11 ± 1.28	13.97 ± 0.48
SCORE	284.88 ± 42.32	571.4 ± 37.45	352.39 ± 179.96	1252.32 ± 104.97

MLP that precedes the LSTM automatically learns a representation of the input data that is relevant to the task of RUL prediction. Figure (5) shows the normalized raw input signals of unit #13 from the FD004 data set, where no clear trend can be seen because of the high variance in the data, which is partly due to the operating conditions that vary from cycle to cycle. Figure (6) shows the output signals of the first MLP, where noticeable degradation pattern have been learned from the normalized inputs and can be observed. Feeding this learned representation to the rest of the model is more efficient than handcrafting features that require expertise and time. This first representation learning stage is particularly useful when dealing with complex data sets where no clear trend is seen, and also when inputs have a large number of dimensions.

After this first MLP, the role of the LSTM layer is to capture temporal patterns and dependencies in the time series. Figure (7) shows the signal at the output of the LSTM. We can see that this part of the model minimizes the variance of the

Table 4. Performance comparison of related methods with our proposed model on the C-MAPSS benchmark.

Models	DATA SETS								Average	
	FD001		FD002		FD003		FD004			
	RMSE	Score	RMSE	Score	RMSE	Score	RMSE	Score	RMSE	Score
DA-CNN (Song et al., 2020)	11.78	229.48	16.95	1842.38	11.56	257.11	18.23	2317.32	14.63	1161.57
DCGAN (Hou et al., 2020)	<b>10.71</b>	<b>174</b>	19.49	2982	<b>11.48</b>	273	19.71	3874	15.34	1825.75
MS-DCNN (Li et al., 2020)	11.44	196.22	19.35	3747	11.67	<b>241.89</b>	22.22	4844	16.17	2257.27
HDNN (Al-Dulaimi et al., 2019)	13.017	245	15.24	1282.42	12.22	287.72	18.15	1527.42	14.65	835.64
LSTM (Pasa et al., 2019)	16.5	444	18.1	942	15.9	718	17.2	1487	16.92	897.75
<b>Proposed LSTM without the first MLP</b>	14.31	337.86	17.44	1716.11	15.53	1356.36	18.86	2111.05	16.53	1380.34
<b>Proposed LSTM with the first MLP</b>	13.26	284.88	<b>12.49</b>	<b>571.4</b>	13.11	352.39	<b>13.97</b>	<b>1252.32</b>	<b>13.20</b>	<b>615.24</b>

learned features across time cycles giving a smoother signals that can be used by the final MLP for RUL estimation.

#### 4.6. Comparison with related works

We evaluate in Table (4) our proposed model by comparing its performances with the most recent methods published in the literature that give the best results on the C-MAPSS data set to the best of our knowledge.

Although the previously published models are performing well on the first and third data sets (FD001, FD003), with only one operating condition, they perform poorly on the other subsets that have up to 6 operating conditions, except for the approaches proposed in (Al-Dulaimi et al., 2019) and in (Pasa et al., 2019).

The proposed end-to-end architecture outperforms all other models in complex data sets (FD002 and FD004) as well as on the global results averaged over all datasets. It improves by more than 18% for the RMSE and 39% for the Score on FD002, and 18% for the RMSE and 15 % for the Score on the FD004 data set, as compared to literature results.

We explain these good results as a consequence of adding a first representation learning MLP before the LSTM and training both of them in an end-to-end manner. Indeed, we can see from the last two rows of Table (4) that the results improved significantly after adding the first MLP to the architecture.

The outputs of the first MLP shows that this part is removing a large part of the variability of the sensor signals that is due to varying operating conditions (Figure (6)). This greatly facilitates the work of the LSTM that can focus on temporal smoothing, and then of the final MLP, which role is to achieve prediction. This idea of facilitating the work of the LSTM can also be achieved by feature engineering as proposed in (Pasa

et al., 2019), where they normalized input according to the operating conditions, the results are relatively good, but this approach can not be performed when some or all of the operating conditions are not known, unlike the proposed approach in this work.

The clear decomposition in our model of these three roles is the key to the increased robustness to variable input signals and better final performances. This can be also observed in Table (4), where all competing models suffer from a large variability of their performance between the FD001 and FD003 subsets on the one hand, and the FD002 and FD004 subsets on the other hand, while a significantly lower difference in the results between the 4 subsets can be observed with our proposed model.

#### 5. CONCLUSION AND FUTURE WORK

In this paper, we presented an end-to-end deep learning approach for RUL estimation from multivariate time-series signals. The proposed method has been tested on the public C-MAPSS data set where the goal is to predict the RUL of commercial aero-engine units. Comparisons with several state-of-the-art approaches have been conducted. The results show that our proposed neural architecture gives the best scores when compared to other approaches applied to the same data sets, especially on complex ones with different operating conditions. Furthermore, it exhibits a more consistent behaviour across the four datasets.

In future work, we plan to explore the performances of the proposed model on other more realistic data sets with different input size, where we may not be able to use the same approach as used in this paper, such as feeding the entire run to failure trajectories to the model due to their long length.

## FUNDING

This work is part of the project AI-PROFICIENT which has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No 957391. The Grid5000 computing resources have been used to partly train and evaluate the proposed models.

## REFERENCES

- Al-Dulaimi, A., Zabihi, S., Asif, A., & Mohammadi, A. (2019). Hybrid deep neural network model for remaining useful life estimation. In *Icassp 2019-2019 IEEE international conference on acoustics speech and signal processing (icassp)* (pp. 3872–3876).
- An, Z., Li, S., Wang, J., & Jiang, X. (2020). A novel bearing intelligent fault diagnosis framework under time-varying working conditions using recurrent neural network. *ISA Transactions*, *100*, 155-170. doi: <https://doi.org/10.1016/j.isatra.2019.11.010>
- Cauchi, N., Macek, K., & Abate, A. (2017). Model-based predictive maintenance in building automation systems with user discomfort. *Energy*, *138*, 306–315.
- Heimes, F. O. (2008). Recurrent neural networks for remaining useful life estimation. In *2008 international conference on prognostics and health management* (p. 1-6). doi: 10.1109/PHM.2008.4711422
- Hou, G., Xu, S., Zhou, N., Yang, L., & Fu, Q. (2020). Remaining useful life estimation using deep convolutional generative adversarial networks based on an autoencoder scheme. *Computational Intelligence and Neuroscience*, *2020*.
- Huang, C.-G., Huang, H.-Z., & Li, Y.-F. (2019). A bidirectional lstm prognostics method under multiple operational conditions. *IEEE Transactions on Industrial Electronics*, *66*(11), 8792–8802.
- Krupitzer, C., Wagenhals, T., Züfle, M., Lesch, V., Schäfer, D., Mozaffarin, A., ... Kounev, S. (2020). *A survey on predictive maintenance for industry 4.0*.
- Li, H., Zhao, W., Zhang, Y., & Zio, E. (2020). Remaining useful life prediction using multi-scale deep convolutional neural network. *Applied Soft Computing*, *89*, 106113.
- Pasa, G. D., Medeiros, I., & Yoneyama, T. (2019). Operating condition-invariant neural network-based prognostics methods applied on turbofan aircraft engines..
- Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008). Damage propagation modeling for aircraft engine run-to-failure simulation. In *2008 international conference on prognostics and health management* (pp. 1–9).
- Song, Y., Gao, S., Li, Y., Jia, L., Li, Q., & Pang, F. (2020). Distributed attention-based temporal convolutional network for remaining useful life prediction. *IEEE Internet of Things Journal*.
- Yuan, Y., Jiang, X., & Liu, X. (2013). Predictive maintenance of shield tunnels. *Tunnelling and Underground Space Technology*, *38*, 69–86.
- Zhang, L., Lin, J., Liu, B., Zhang, Z., Yan, X., & Wei, M. (2019). A review on deep learning applications in prognostics and health management. *IEEE Access*, *7*, 162415–162438.
- Zhang, W., Yang, D., & Wang, H. (2019). Data-driven methods for predictive maintenance of industrial equipment: a survey. *IEEE Systems Journal*, *13*(3), 2213–2227.
- Zheng, S., Ristovski, K., Farahat, A., & Gupta, C. (2017). Long short-term memory network for remaining useful life estimation. In *2017 IEEE international conference on prognostics and health management (icphm)* (pp. 88–95).

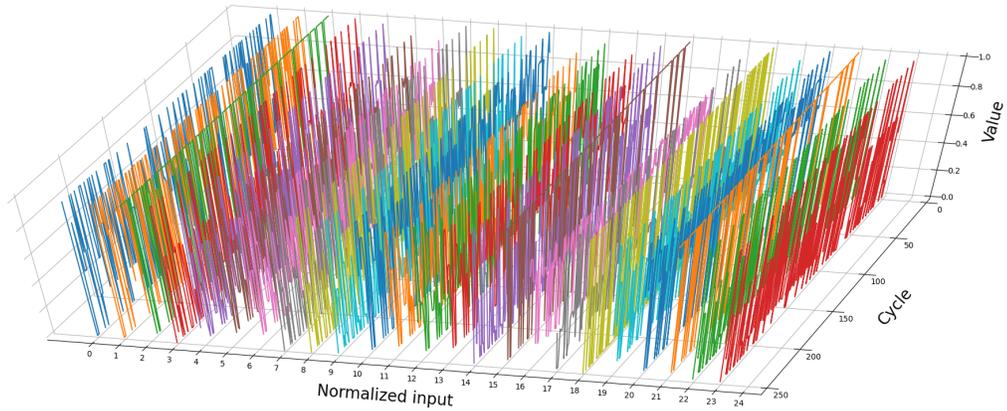


Figure 5. Normalized input signals that are fed directly into the model;  $n = 24$  sensor measurements of the turbofan unit #13 from the beginning of its life until its failure; this engine data was taken from the 4<sup>th</sup> data set (FD004) that contains 6 operating conditions and 2 fault modes; we clearly see that these normalized signals do not directly provide visible and interpretable clues for RUL estimation.

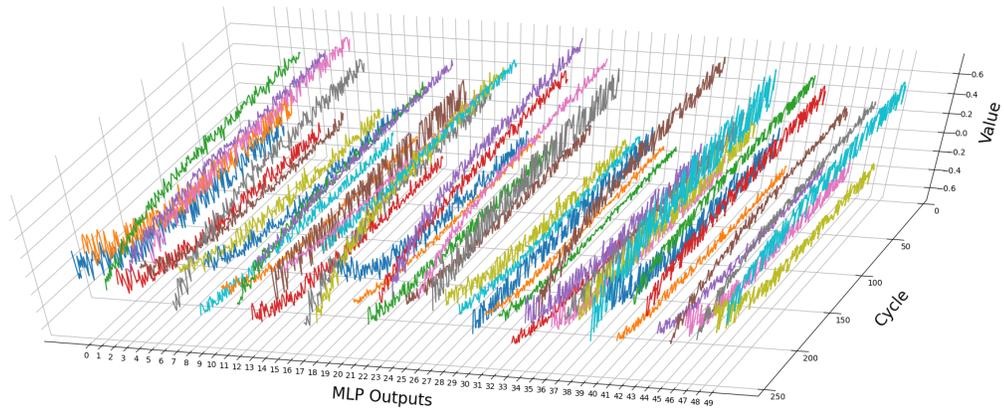


Figure 6. This plot presents the 50 features learned by the MLP for unit #13; we can observe trending degradation representations that have been learned from the normalized input signals. Since the first MLP is not time dependent, the learned features exhibit a relatively large variance across time cycles.

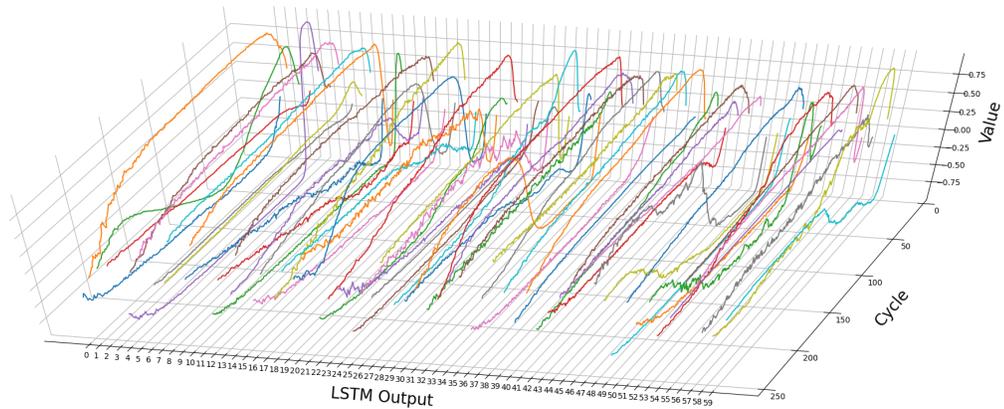


Figure 7. The outputs of the LSTM for unit #13 present much smoother signals, due to the LSTM’s ability to leverage recurrent connection from prior time steps.