Soft Computing
https://doi.org/10.1007/s00500-023-08322-6

APPLICATION OF SOFT COMPUTING

Stacking-based ensemble learning for remaining useful life estimation

Begum Ay Ture[1] · Akhan Akbulut[2] · Abdul Halim Zaim[1] · Cagatay Catal[3]

Accepted: 18 April 2023

Abstract
Excessive and untimely maintenance prompts economic losses and unnecessary workload. Therefore, predictive maintenance models are developed to estimate the right time for maintenance. In this study, predictive models that estimate the remaining useful life of turbofan engines have been developed using deep learning algorithms on NASA's turbofan engine degradation simulation dataset. Before equipment failure, the proposed model presents an estimated timeline for maintenance. The experimental studies demonstrated that the stacking ensemble learning and the convolutional neural network (CNN) methods are superior to the other investigated methods. While the convolution neural network (CNN) method was superior to the other investigated methods with an accuracy of 93.93%, the stacking ensemble learning method provided the best result with an accuracy of 95.72%.

Keywords Remaining useful life · Ensemble learning · Deep learning · Stacking ensemble learning

1 Introduction

The way maintenance decisions are made within the context of Industry 4.0 has significantly changed in recent years. The importance of matters such as planning maintenance, calculating the failure point of equipment, and inspecting machinery inside the factory has increased. In different industries, it is often necessary to turn off a machine as soon as possible to prevent further damage in case of failure (Lei et al. 2016).

However, such interruptions during the production phase can cause significant time and economic losses if they occur at the wrong time. Hence, employers that aim to build an innovative factory environment have started integrating maintenance applications into their systems. Monitoring machine conditions is a fundamental part of business operations. Also, predicting which machine will fail or complete its remaining useful life is significant for maintenance planning.

Maintenance strategies fall under two main categories: proactive and reactive. In proactive maintenance, the main goal is to preserve the system running at the best level and minimize financial losses, while reactive maintenance deals with post-failure situations and financial losses. Maintenance strategies have evolved from error and correction practices (i.e., diagnostics) to prediction and prevention methodologies (i.e., prognostics) (Lee et al. 2014). Prognostics can be described as a remaining useful life (RUL) estimation process of the system or a subsystem (Soni et al. 2020). RUL is expressed in units of time (e.g., cycles or hours) and provides a threshold value that helps in setting safe operating limits. Thus, it is considered that the system will work smoothly until the RUL threshold value. Estimating the RUL can also help extend the lifespan of components by upgrading worn-out machinery or changing operational parameters that have an impact on the entire system. Maintenance planning is significant for the continuity of equipment and system health. Prognostic knowledge is the main factor for condition-based maintenance (CBM). With the help of the RUL estimation, decision makers can avoid unnecessary maintenance costs

✉ Cagatay Catal
ccatal@qu.edu.qa

Begum Ay Ture
begum.ay@istanbulticaret.edu.tr

Akhan Akbulut
a.akbulut@iku.edu.tr

Abdul Halim Zaim
azaim@ticaret.edu.tr

1 Department of Computer Engineering, Istanbul Commerce University, 34840 Istanbul, Turkey

2 Department of Computer Engineering, Istanbul Kültür University, 34536 Istanbul, Turkey

3 Department of Computer Science and Engineering, Qatar University, Doha 2713, Qatar

Published online: 21 May 2023

and perform on-site and timely maintenance. The RUL estimates are also crucial for safety in aircraft systems and maintenance planning. It is also necessary to perform highly accurate estimates and prepare relevant maintenance plans based on these accurate predictions. Maintenance is generally grouped under two main topics, as shown in Fig. 1. Corrective maintenance deals with bringing the system to its previous faultless state after the failure. Preventive maintenance focuses on maintenance planning at regular intervals before the system fails, preventing sudden systemic breakdowns and unnecessary financial losses.

The subject of RUL estimation is a sub-field of prognosis and health management (PHM) and has become more popular recently to enhance performance and prevent financial losses in the work environment. The RUL of an asset is described as the length from the current time to the end of its useful life (Si et al. 2011). The RUL estimation is a process that determines the useful life remaining before the machine loses its ability to operate (Lee et al. 2014). RUL prediction offers a number of key benefits, including early prevention, cost savings, and increased operational effectiveness.

There are three main classes of RUL prediction methods: (i) data-driven methods, (ii) physics model-based methods, and (iii) methods that combine data-driven and physics model-based methods (Soni et al. 2020). Historical system data and fault records are critical for the data-driven method. Another approach, which is called the model-based approach, is a method that requires expert knowledge, and historical data on the environment is not required for the estimation. As Isermann (1984) introduced in his study, there are various model-based approaches in the literature. Finally, the hybrid approach combines the use of model-based and data-based methodologies. In this paper, we present a data-driven approach to estimating the RUL of turbofan jet engines. The turbofan engine degradation simulation dataset that was produced with NASA's C-MAPSS simulator was used in this research. Chen et al. (2017) pointed out that if the past failure samples are limited, the support vector machine method might be appropriate. If the number of samples is abundant, it is appropriate to adopt a similarity-based methodology.

The following two tasks are mostly used in predictive maintenance studies: classification and regression. When the problem is addressed using a regression algorithm, a value called the RUL(Li and He 2015) is predicted. In this research, ensemble learning was performed using different machine learning techniques. Li et al. (2006) showed that supervised learning algorithms can be adopted as the main predictors for ensemble learning. In this research, the stacking ensemble learning model was developed by applying all the regression algorithms. Long short-term memory (LSTM) and convolutional neural network (CNN) models were investigated in this study.

The objectives of this study are listed as follows:

1. To evaluate the performance of the stacking ensemble learning approach for the remaining useful life estimation problem,
2. To benchmark the performance of the stacking ensemble learning with other widely used techniques such as convolutional neural network,
3. To present a highly accurate prediction model in order to guide future RUL estimation studies.

The following sections are organized as follows: Sect. 2 presents the related work. Section 3 shows the methodology. Section 4 explains the experimental results. Section 5 concludes the paper.

2 Related work

Mathew et al. (2017) compared machine learning algorithms in estimating the RUL of turbofan jet engines. They showed that the random forest algorithm provided the best performance. Ahsan et al. (2017) developed the autoregressive model using NASA's turbofan engine dataset. In some studies, experiments were performed using machine learning methods for both the classification and the regression tasks. Mosallam (2014) carried out classification and regression-based RUL calculations and described his Bayesian-based approach. Hu et al. (2012) applied ensemble learning instead of traditional data-driven methods to achieve the most optimal estimation result. As Polikar and Polikar (2006) shows, there are many theoretical and practical reasons for practicing ensemble learning. The primary evaluation factor in this study is performance enhancement. Soni et al. (2020) used LSTM and CNN deep learning methods to estimate the remaining useful life of gas turbines and showed that the CNN method exhibited more reliable performance results. Al-Dulaimi et al. (2019) practiced LSTM and CNN algorithms in order to improve their hybrid neural network RUL prediction model. In deep learning methods, the size of the data is much more than in machine learning methods. Ellefsen et al. (2019) developed a semi-supervised model and showed the effect of a small amount of labeled training data on the model while estimating the remaining useful life. In most of the studies investigated in the literature, it is recognized that the CNN method introduced improved results than the LSTM compared to it. In some studies, deep convolutional neural network was applied instead of CNN. Babu et al. (2016), Zheng et al. (2017), and Li et al. (2018) displayed in their studies that the use of deep convolutional neural networks provides better accuracy for estimating remaining useful life. A systematic literature review performed in this domain examined gaps in the field (Begüm et al. 2021).

It is clear from a review of current research that there has been an uptick in methods that exploit attention. Attention

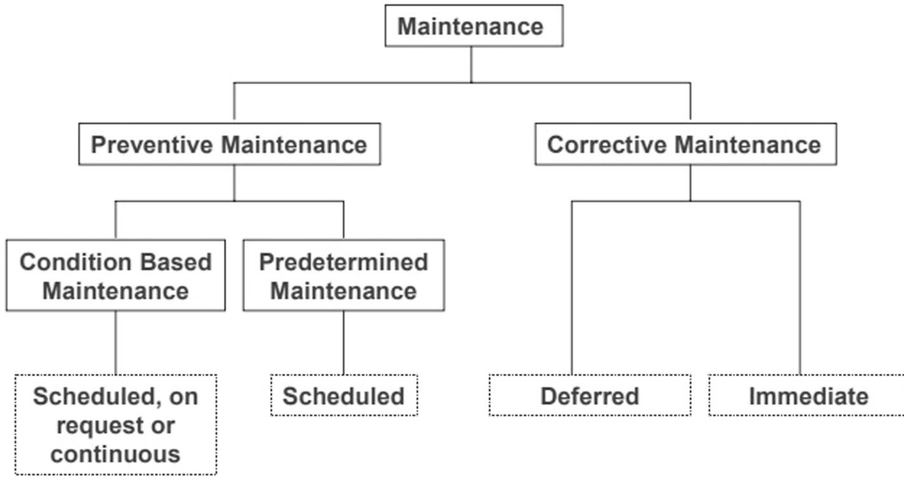

Fig. 1 Maintenance overview (EN-13306: 2010) [4]

is a mechanism used in deep learning models that allows the model to selectively focus on various parts of the input data rather than uniformly processing the entire input. This is especially beneficial when the input data is complex and of variable lengths, such as natural language text or time-series data. In predictive maintenance, for instance, the input data may consist of sensor readings from a piece of equipment over time, and attention can be used to determine which readings are most pertinent for predicting an impending failure.

Attention has been used as an alternative to conventional CNN architectures for processing time-series data in LSTM-based research on predictive maintenance (Li et al. 2023; Wang et al. 2023). CNNs can be effective at identifying local patterns in time-series data, but they may struggle with identifying long-term dependencies or determining which portions of the input sequence are most important for making a prediction. Attention, on the other hand, enables the model to dynamically weight various parts of the input sequence based on their importance to the prediction assignment. Several recent papers utilizing attention-based LSTM models for fault detection and failure prediction in industrial settings have demonstrated that this can contribute to improved performance on predictive maintenance tasks (Tian et al. 2023; Zhang et al. 2023).

Highly cited research papers and important contributions in this field are presented in Table 1. This study intends to develop ensemble learning and deep learning models to achieve highly accurate results for RUL estimation.

3 Methodology

3.1 Dataset

The dataset has been obtained from NASA's Prognostics Center of Excellence (PCoE) archive and includes simulated data of the failures (Run-to-Failure) of the turbofan jet engines during their operation. The engine degradation simulation was created with NASA's C-MAPSS (Commercial Modular Aero-Propulsion System Simulation) simulator. C-MAPSS is a tool for simulating realistic large turbofan engine data. The data for each simulation is generated up to the engine failure. The goal of this dataset is to predict the remaining useful lifetime (RUL) of each engine in the test dataset, i.e., the number of operating cycles after the last cycle the engine will continue to run. In this study, NASA's C-MAPSS (Commercial Modular Aero-Propulsion System Simulation) dataset has been used to validate the effectiveness of the proposed models because this dataset is widely used by researchers in this field and is publicly available. Since the dataset is publicly available, researchers can compare the performance of their models with the proposed model easily.

When the dataset is reviewed, four different datasets with different operating conditions and failure modes can be seen, namely FD001, FD002, FD003, and FD004. There are two failure modes in the dataset, and four different datasets were obtained using these failures. "One" and "two" are the failure modes' names. The word one indicates HPC corruption, whereas the word two indicates HPC corruption and fan breakdown. Each dataset has been divided into training and testing datasets. Also, there is a separate csv file that contains the test datasets' actual remaining useful lifetime (RUL) values. RUL is equivalent to the number of flights remaining for the engine after the last data point in the testing dataset.

In this study, the presented experiments are implemented on the FD001 dataset. The dataset is represented as a text file containing 26 columns of numbers separated by spaces. Each row is a snapshot of data taken during a single operational cycle, and each column represents a different variable. The dataset includes engine unit number, time in the cycle, three operational settings, and the 21 sensor measurements such as

Table 1 Classification of related work

	Study	Selected approach	Applied method(s)	Performance metric(s)	Score
Machine learning (shallow)	Mosallam (2014)	Comparative approach	Gaussian process regression, K-nearest neighbors	MAPE	27.57
	Mathew et al. (2017)	Comparative approach	Linear regression, decision tree, random forest, K-nearest neighbors, AdaBoost, gradient boosting method, K-means algorithm, ANOVA	RMSE	29.73
	Polikar (2006) Polikar and Polikar (2006)	Ensemble-based approach	Bagging, boosting, AdaBoost, stacked generalization	N/A	N/A
	Hu et al. (2012)	Ensemble-based approach	Similarity-based interpolation, Bayesian linear regression, recurrent neural network, support vector machine	Validation error	0.6984
	Ahsan et al. (2017) Ahsan and Lemma (2017)	Autoregressive models	Autoregressive model	Fit percent, MSE	54.9 - 0.0
Deep learning	Soni et al. (2020)	Deep neural network	Long short-term memory, convolutional neural network	Accuracy	94.16
	Al-Dulaimi et al. (2019)	Deep neural network	Long short-term memory, convolutional neural network	RMSE	12.22
	Ellefsen et al. (2019)	Semi-supervised deep learning	Restricted Boltzmann machine, long short-term memory, feedforward neural networks, genetic algorithm for HP tuning	RMSE	12.56
	Babu et al. (2016)	Deep neural network	Convolutional neural network	RMSE	18.44
	Zheng et al. (2017)	Deep neural network	Convolutional neural network	RMSE	2.80
	Li et al. (2018)	Deep neural network	Convolutional neural network	RMSE	12.61
	Li et al. (2023)	Hybrid deep neural network	Self-attention mechanism, convolutional neural network, long short-term memory networks	RMSE	12.6
	Wang et al. (2023)	Hybrid deep neural network	Transformer encoder, attention	RMSE	10.35
	Tian et al. (2023)	Hybrid deep neural network	Spatial correlation and temporal attention mechanism, long short-term memory networks	RMSE	4.15
	Zhang et al. (2023)	Hybrid deep neural network	Embedded attention-based parallel network, long short-term memory networks, gated recurrent unit	RMSE	12.11

Table 2 Dataset description

Index	Data fields	Types	Descriptions
1	Engine Id	Integer	Aircraft engine number
2	Cycle	Integer	Time in cycles
3	Settings 1	Double	Operational setting 1
4	Settings 2	Double	Operational setting 2
5	Settings 3	Double	Operational setting 3
6	Settings 4	Double	Sensor measurement 1
7	–	–	–
8	Settings 21	Double	Sensor measurement 21

temperature, pressure, fan speed, coolant bleed, and fuel–air ratio. Table 2 presents an overview of the dataset.

3.2 Ensemble learning

Ensemble learning (a.k.a., multiple machine learning approac hes) is a technique that creates multiple predictive models and determines the final prediction result by combining the results retrieved from these models. Ensemble learning approaches maximize the benefit and advance the efficiency of individual machine learning models. Since there is more than one machine learning algorithm applied in ensemble learning, model training takes longer compared to individual learners. Although slow machine learning methods can also contribute to ensemble learning, faster machine learning algorithms are generally favored in ensemble learning methods. Several methods are available to design ensemble learning in regression models (Dietterich 2000; Akbulut et al. 2022).

This research aims to design ensemble learning-based models by applying multiple regression algorithms on the selected dataset. The most widely used advanced ensemble learning methods are bagging (bootstrap aggregating), boosting, and stacking. Breiman (1996) creates bootstrapped samples, and each machine learning algorithm (either classifier or regressor based on the underlying task) is applied to each sample. For a regression task, an average is computed based on the outputs of all regression algorithms. Boosting Duffy and Helmbold (2002) is another method in which each subsequent model aims to correct the errors of the previous model by creating subsets of the dataset. The final model is the average of all models generated. Finally, stacking is a technique that aims to provide high accuracy by using estimation results from multiple models (Breiman 1996). In this study, the stacking ensemble learning method has been investigated because of its capability in handling complex datasets.

3.3 Machine learning algorithms

In this study, the applied methods are supervised learning methods (Molnar 2020). The regression algorithms studied are linear regression, support vector regression (SVR), decision tree regression, random forest regression, and extreme gradient boosting regression. These five learning algorithms have been combined with the stacking ensemble learning method to increase performance. Figure 2 shows the steps taken to predict the RUL using machine learning algorithms.

3.4 Deep learning algorithms

Deep learning is a sub-field of machine learning and inspires by artificial neural networks. In addition to the fully connected (FC) layer type, deep learning models use additional layer types such as the convolutional layer, pooling layer, and dropout layer. In deep learning, features are learned at different levels of the hierarchy (Mathew et al. 2017; Yildirim et al. 2023). Features can be learned automatically in the deep learning models, and the model does not require hand-engineered features. In this study, because of the limited data, the classification problem was addressed. According to the literature search, long short-term memory (LSTM) and convolutional neural network (CNN) methods were found to be the most promising deep learning algorithms. As such, LSTM and CNN methods were used for the RUL estimation. Figure 3 shows the steps of deep learning-based RUL estimation.

3.4.1 Long short-term memory

The long short-term memory (LSTM) network is a more specific version of recurrent neural networks (RNNs) trained using back-propagation. LSTM is a method that can prevent problems experienced in typical RNNs. In the LSTM model, random periods are remembered and the model includes LSTM blocks rather than neurons. LSTMs are designed to prevent long-term dependency problems (Soni et al. 2020). In other words, the values in the memory can be kept for a long time, which is the expected behavior of the LSTM method. Thus, it forms the context for the data and can make inferences about data. LSTM networks can be applied for data classification problems using time-series data (Sadigov et al. 2023).

3.4.2 Convolutional neural network

Convolutional neural network (CNN) is a specialized version of artificial neural networks that uses many layers and different types of layer types. CNN-based models present outstanding results in image processing, voice and face recognition, object identification in autonomous vehicles, and traffic sign identification. The data to be transmitted to

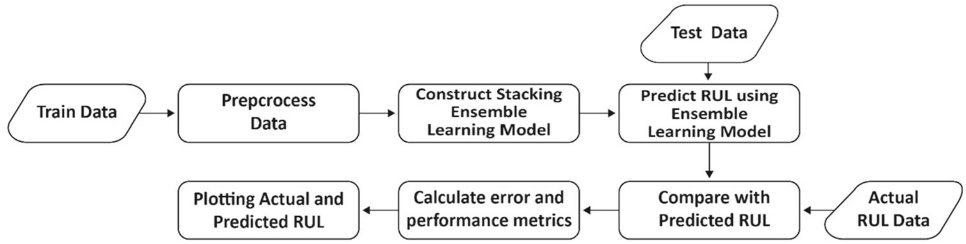

Fig. 2 The prediction flowchart for the RUL model

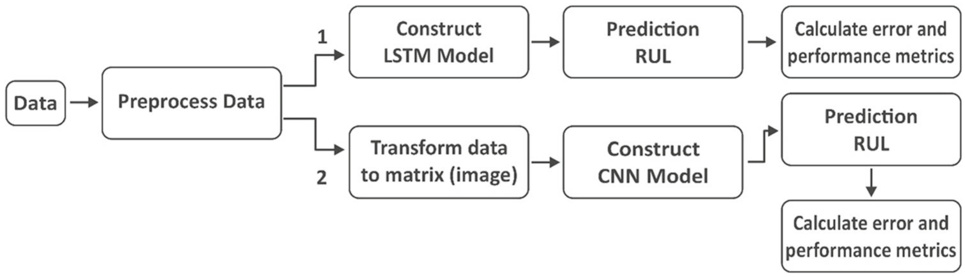

Fig. 3 The prediction flowchart of the RUL model using deep learning algorithms

the CNN model must have the type of image (i.e., matrix) or video (i.e., frames). For this reason, in this study, after the pre-processing stage, the sensor data specific to the CNN model was transformed into a matrix using the recurrence plot method. It learns how to automatically extract the features from the raw data by separating the given data according to specific filters and can make inferences about the given data using these features.

3.5 Experimental setup

The Python programming language was used to implement and test the RUL estimation models. The data has been read from.csv files. The machine learning and deep learning methods were applied to the data using Python on Google Colaboratory (Colab). The best models have been benchmarked in terms of performance and up-time using Jupiter Notebook on a machine with 64 GB RAM and 1080 GTX SLI requirements.

3.6 Preparation of data

The data should be clean and prepared for the model before applying the predictive maintenance model. If the data is imbalanced, contains noise, or has too many missing values, the data must be processed with pre-processing techniques. Ranjan et al. (2018) explained that traditional machine learning approaches do not work correctly if a given dataset is highly imbalanced. Data has been organized and analyzed in accordance with pre-processing methods. F0001 is the example data used for these actions.

The longest-running engine breaks after 362 cycles, while the earliest failed engine after 128 cycles when the maximum time cycles are examined. It is known that an engine degrades on average every 199 to 206 cycles. The required target column values should be added to the data before the pre-processing step. The RUL values corresponding to the records in the testing dataset are provided as a separate.csv file. The maximum number of cycles per motor is presented in the RUL column. This number represents the number of cycles in which the unit number of each engine ends. For each motor record, Formula 1 is used to calculate the RUL value. This calculation is executed for each engine data and the RUL column derivations for the training data.

$$\text{RUL} = \text{ maximum cycle of motor - current cycle of motor.}$$

After adding the target value column, we checked whether there is a NaN or a null value in the dataset. If we had missing data in the dataset, we would update these values using different approaches such as mean, median, most frequent, or constant. Since there are no missing values in the data, this method is not necessary. Descriptive statistics can be used to identify minimum and maximum values that are distant from the mean. The minimum and maximum values for some columns are the same, and the standard deviation is 0. Sensors with a standard deviation of 0 have been removed because of the outlier values.

Normalization is used to scale the data of a feature so that it falls in a smaller range, such as −1 to 1 or 0 to 1. When data columns are examined, the features of the data are in different intervals. The ranges of the data attributes must be close for the methods to be applied. Scaling is used to normalize the data as a result. We chose the normalization algorithm, Min–Max, to scale the dataset and normalized our dataset to the 0–1 range. Before scaling, the distribution of the original data values reveals that the data are extremely distinct from one another. By scaling the data, all of the values are placed inside a range of 0–1. The scaled versions of our data with the previous training and testing structure are shown in Fig. 4.

The features obtained through data pre-processing steps are used to train predictive maintenance models using regression-type machine learning techniques. The data is not extra pre-processed or given a new feature. However, the data should be translated to matrix format and the class labels added before constructing predictive maintenance models using deep learning methods in the classification type. Since these processes are added extra after the general pre-processing steps, they are explained in the deep learning section.

4 Experimental results

In this section, we present our experimental results. Eight learning algorithms including different machine learning and deep learning methods were adopted. For machine learning, 14 different tests were conducted by testing the learning algorithms selected with each data separation method. Moreover, to determine the ensemble learning model's meta-regressor algorithm, ten different tests were completed using two different cross-validation methods and five different machine learning algorithms. For deep learning methods, two specific tests were performed and the results are discussed in this section.

4.1 Results regarding the ensemble learning model

The training and testing stages were implemented on the FD001 dataset. The dataset contains data from one hundred motors, and each motor has 21 different sensor values. The dataset description is shown in Table 2. Five regression algorithms were applied to the FD001 dataset. Before the training step, different approaches were utilized to separate the dataset as training and testing datasets. The first approach is the K-fold cross-validation, and the other one is the leave-one-out (LOO) method that separates the dataset based on the engine numbers. The dataset was divided into tenfold when using K-fold cross-validation (i.e., k=10). The other method in LOO, the training process uses the data points in the training set. One engine was adopted for testing purposes, and all remaining engines were utilized for training. Thus, each engine took part in both training and testing, and each data was applied in the process. The results were collected in each step, and finally, the final result was produced.

Table 3 shows the performance results of the regression algorithms when the tenfold cross-validation is used. As shown in this table, the linear regression algorithm provides the lowest accuracy, and the random forest regression and stacking ensemble learning algorithms provide the best accuracy. Researchers attempt to reduce the RMSE values more than the standard deviation value. When the RMSE and standard deviation rates are evaluated, the stacking ensemble learning method returns better results.

Figure 5 depicts the relationship between the real RUL and the estimated RUL. When the data is separated by the leave-one-out (LOO) method, the performance results are presented in Table 4. As shown in Table 4, stacking ensemble learning, a combined version of five regression algorithms, provides the best performance. When RMSE and standard deviation areas are evaluated, we observe that the RMSE ratio of support vector regression is the lowest. However, the accuracy and standard deviation are not satisfactory compared to stacking ensemble learning. Since the RMSE ratios of the stacking ensemble learning and support vector regression algorithms are very close, the stacking ensemble learning algorithm performs better based on the standard deviation and accuracy metrics.

Figure 6 depicts the relationship between the real RUL and the estimated RUL when the stacking ensemble learning technique is used. The high number of overlapping points on the graph shows that the prediction accuracy is high.

While applying stacking ensemble learning, different combinations of five algorithms were investigated. Each algorithm is employed as the meta-regressor, and the results are listed in Table 5. While the linear regression algorithm provides better results as a meta-regressor in tenfold cross-validation, the support vector regression algorithm is used as a meta-regressor when data is separated with LOO.

Figure 7 shows the sample flowchart of the stacking ensemble learning model. This study illustrates that the stacking ensemble learning method enhances the model's performance. The RUL estimation results have been compared with their actual values, and the corresponding figures were interpreted.

4.2 Deep learning result

As used in machine learning experiments, the data used in deep learning models is FD001. Convolutional neural network (CNN) and Long short-term memory (LSTM) methods were preferred for deep learning models to evaluate the remaining useful life. The pre-processing of data is implemented in the same way as machine learning algorithms, and steps are taken as described in the Preparation of Data section. Deep learning algorithms perform better with vast amounts of data. Due to the limited data, we can approach this task as a classification problem. Since this problem has been treated as a classification problem, a class label is created for each data point. The added class label data was designed via the RUL column. The dataset is labeled as 2 for RUL values between 0 and 15, 1 for 16-30, and 0 for those greater than 30. The pre-processing of the data for the LSTM model has been

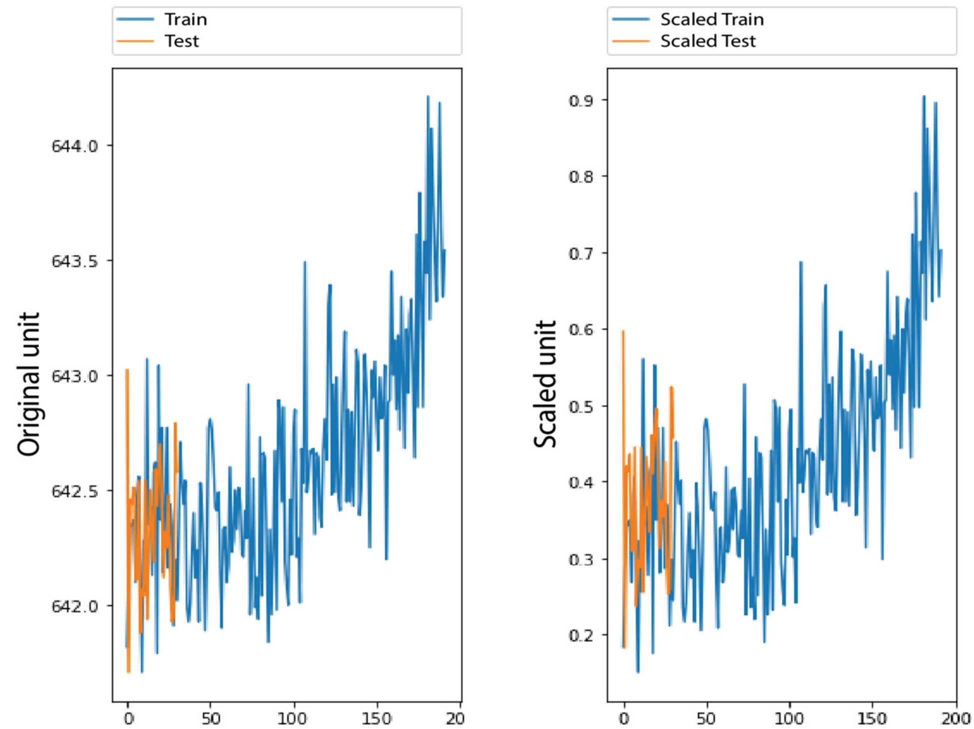

Fig. 4 Flowchart prediction RUL model with ML

Table 3 Results with tenfold cross-validation

K-fold cross-validation	Average accuracy score	RMSE	Std.Dev.
Linear regression	87.35 %	34.02	42.44
Support vector machine	93.90 %	32.80	44.91
Decision tree	93.90 %	32.80	44.19
Random forest	95.72 %	33.33	43.96
Extreme gradient boosting	95.66 %	40.07	35.52
Stacking ensemble learning	95.72 %	33.25	40.93

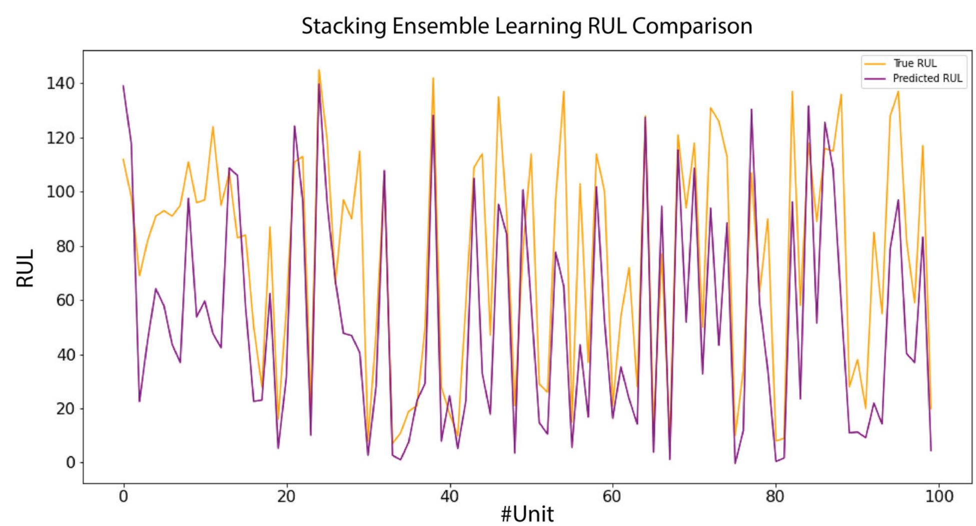

Fig. 5 RUL comparison of stacking ensemble learning with tenfold cross-validation

Table 4 Results with leave-one-out cross-validation

Leave-one-out cross-validation	Average accuracy score	RMSE	Std.Dev.
Linear regression	87.00%	33.80	42.38
Support vector machine	92.67 %	29.23	37.80
Decision tree	90.61 %	40.97	51.93
Random forest	95.30%	33.15	43.80
Extreme gradient boosting	95.31%	39.07	35.06
Stacking ensemble learning	95.69 %	31.30	35.15

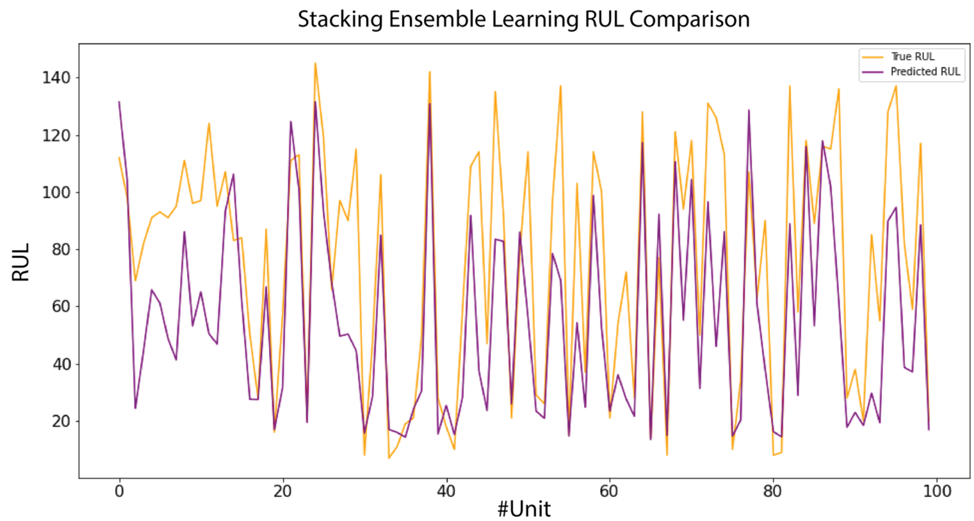

Fig. 6 RUL comparison of stacking ensemble learning with LOO

Table 5 Stacking ensemble learning

K-fold cross-validation		Accuracy	RMSE	Std.Dev.
Stacking	Linear regression	95.72 %	33.25	40.93
Ensemble	Support vector machine	91.01 %	31.34	35.20
Learning	Decision tree	95.22 %	35.71	46.19
Main	Random forest	90.64%	41.67	54.55
Learner	Extreme gradient boosting	95.68 %	33.19	44.03
Leave-one-out cross-validation		**Accuracy**	**RMSE**	**Std.Dev.**
Stacking	Linear regression	95.38 %	33.38	41.15
Ensemble	Support vector machine	95.69 %	31.30	35.15
Learning	Random forest	94.89 %	35.67	44.88
Main	Decision tree	90.29 %	44.54	53.46
Learner	Extreme gradient boosting	95.36 %	32.87	43.83

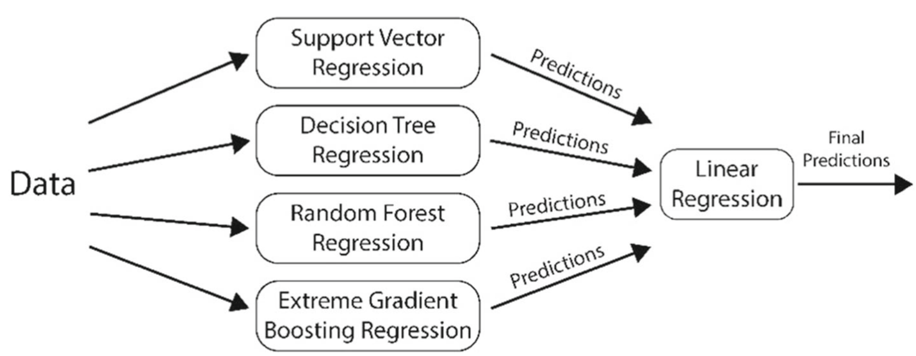

Fig. 7 The stacking ensemble learning flowchart

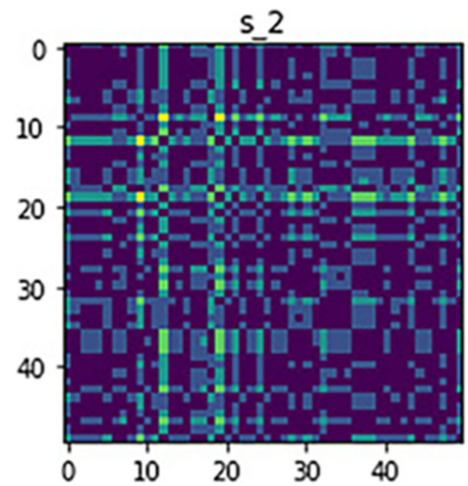

Fig. 8 Transform time series to image

completed. For the CNN model, the data has been converted to a matrix type because the CNN algorithm accepts such data. Therefore, the time-series data has been pre-processed and transformed into a matrix type via the recurrence plot function. In Fig. 8, as an example, sensor data number 2 is transformed into an image. The drawing size is 50x50x17. 50x50 is the size of the recurrence plot function to convert time-series data to image type. The number 17 also symbolizes our features. Since the columns with a standard deviation of 0 are removed from our data after pre-processing, we have 17 feature columns left.

Therefore, the pre-processing stages for LSTM and CNN models and preparation for training were completed. After this step, the CNN and LSTM models were designed. First, we investigate the structure of our LSTM model and the parameters used in its design. As shown in Fig. 9, the layer structure of our LSTM model consists of three LSTM layers, three dropout layers, and one dense layer as the output layer.

Softmax was preferred as the activation function in the output layer. This activation function performs well for classifiers, which can be used in problems requiring more than two class labels. The early stopping function was employed for the training step. At the same time, cross-validation has been adopted to avoid the overfitting of the model. As practiced in machine learning methods, with the leave-one-out algorithm, the data is separated according to their engine number in every step. After the training phase, the model's accuracy was evaluated with testing data. The accuracy of the LSTM model was 83.99%. After the LSTM model, the CNN algorithm was used to develop a new remaining useful life estimation model. The layer structure of the CNN-based model is shown in Fig. 10.

The data transformed to image (i.e., matrix) format was trained using this model. The early stop function was applied to attain the most appropriate result during the training. Additionally, the leave-one-out algorithm is applied to prevent overfitting. The model has been verified with the testing dataset, and the CNN model's accuracy was found to be 93.93

4.3 Comparison of the best scores of ensemble learning and deep learning models

Classical machine learning methods and deep learning methods were adopted for the remaining useful life estimation process. All algorithms used were described and contrasted with other methods. According to the comparison results, it is comprehended that the stacking ensemble learning algorithm, which was developed using classical machine learning methods, provides the most beneficial results. On the other hand, it is observed that the convolutional neural network algorithm has the highest accuracy for deep learning-based models. The developed models have been investigated in terms of training and testing times.

Since deep learning methods involve many hidden layers and neurons in their structures, they are more complex than traditional machine learning algorithms (i.e., shallow learning techniques). In addition, deep learning methods require a large amount of data and thereby more computing power is needed. The complexity of these models also affects the training time, which is typically much longer than the required time of traditional machine learning models. As such, the

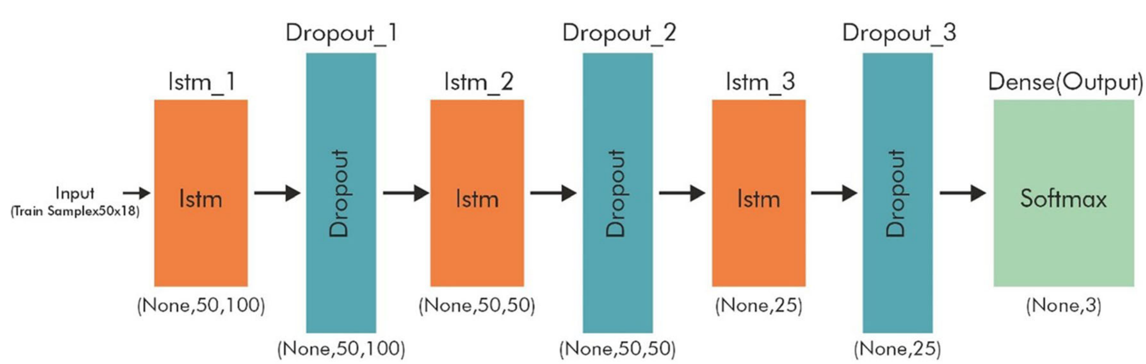

Fig. 9 The LSTM model architecture

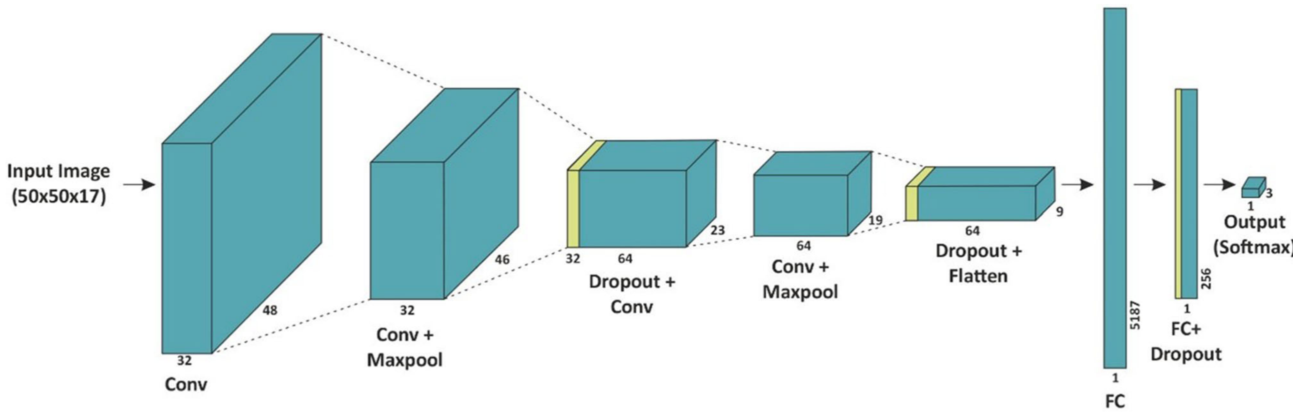

Fig. 10 The CNN model architecture

Table 6 The comparison of best results

Model	Accuracy	Training time (second)	Testing time (second)
Stacking ensemble learning	95.72	1023.36	1.86
CNN	93.93	3386.38	12.16

time spent on training for deep learning models is longer than the time of the proposed stacking ensemble learning model. Model accuracy and training and testing times are presented in Table 6. As a result, the accuracy and speed of the stacking ensemble learning algorithm are better than others.

5 Conclusion

Maintenance prediction is critical to avoid unnecessary maintenance activities, prevent economic losses, reduce unnecessary workload, and increase working performance. The main objective of predictive maintenance is to estimate equipment failure to determine the RUL of the equipment. Maintenance planning could be accomplished more effectively if equipment failures were predicted in advance. In this study, models that estimate the RUL of turbofan jet engines have been developed using both classical machine learning methods and deep learning methods. Both regression and classification models were formed, and model requirements were described. When the experimental results are investigated, it is seen that the stacking ensemble learning method, which was generated using five different machine learning methods, returns better results than the other models (i.e., 95.72% accuracy).

This research has significant scientific value when considering different industries such as automotive, manufacturing, healthcare, and aerospace because such a highly accurate prediction model can minimize the risk of system failures and can dramatically reduce downtime. In the current application domain (i.e., aerospace), the presented model can reduce maintenance costs, improves system safety, and maximizes availability. Also, in other industries, devices can be easily replaced before they fail and improve safety. To fully understand the scientific value of this research in different industries such as healthcare and automotive, follow-up research must be performed.

Recently, new feature selection techniques such as hybrid J-SLnO, which is a hybrid of the Jaya algorithm and Sea Lion Optimization (SLnO) and an LSTM-based redundant attribute removal technique (Jiang et al. 2022) have been proposed in the literature (Abidi et al. 2022). Also, new regression models such as the dendritic neural regression (DNR) model, the heap-based optimizer (HBO)-LSTM model, and the modified ANFIS model using the Aquila optimizer (AO) with the opposition-based learning (OBL) called AOOBL-ANFIS have been developed recently (Al-qaness et al. 2022; Ewees et al. 2022; Al-qaness et al. 2022). As future work, this type of feature selection techniques and regression models can be investigated to improve the overall performance.

In future work, the performance of the RUL prediction model can be improved by using a considerable amount of real-time data. Also, new deep learning algorithms such as transformers can be investigated to improve the performance of the prediction models.

Author Contributions All authors contributed to the study and design. All authors contributed to the manuscript writing. All authors read and approved the final manuscript.

Funding Open access funding provided by Qatar National Library (QNL).

Data Availability The datasets analyzed during the current study are available in the www.nasa.gov/content/prognostics-center-of-excellence-data-set-repository

Declarations

Ethical Approval Ethical approval is not applicable because this article does not contain any studies with human or animal subjects

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Abidi MH, Mohammed MK, Alkhalefah H (2022) Predictive maintenance planning for industry 4.0 using machine learning for sustainable manufacturing. Sustainability 14(6):3387

Ahsan S, Lemma TA (2017) Remaining useful life prediction of gas turbine engine using autoregressive model, In: MATEC Web of Conferences, Vol. 131, EDP Sciences, p 04014

Akbulut A, Gungor F, Tarakci E, Aydin MA, Zaim AH, Catal C (2022) Identification of phantom movements with an ensemble learning approach. Comput Biol Med 150:106132

Al-Dulaimi A, Zabihi S, Asif A, Mohammadi A (2019) A multimodal and hybrid deep neural network model for remaining useful life estimation. Comput Ind 108:186–196

Al-qaness MA, Ewees AA, Elaziz MA, Samak AH (2022) Wind power forecasting using optimized dendritic neural model based on seagull optimization algorithm and aquila optimizer. Energies 15(24):9261

Al-qaness MA, Ewees AA, Fan H, AlRassas AM, Abd Elaziz M (2022) Modified aquila optimizer for forecasting oil production. Geo-Spatial Inf Sci 25(4):519–535

Begüm A, Akbulut A, Zaim AH (2021) Techniques for apply predictive maintenance and remaining useful life: a systematic mapping study. Bilecik Şeyh Edebali Üniversitesi Fen Bilimleri Dergisi 8(1):497–511

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Breiman L (1996) Stacked regressions. Mach Learn 24(1):49–64

CEN, 13306: Maintenance terminology, European Committee for Standardization: Brussels, Belgium

Chen Z, Cao S, Mao Z (2017) Remaining useful life estimation of aircraft engines using a modified similarity and supporting vector machine (svm) approach. Energies 11(1):28

Dietterich TG (2000) Ensemble methods in machine learning, In: International workshop on multiple classifier systems, Springer, pp 1–15

Duffy N, Helmbold D (2002) Boosting methods for regression. Mach Learn 47(2):153–200

Ellefsen AL, Bjørlykhaug E, Æsøy V, Ushakov S, Zhang H (2019) Remaining useful life predictions for turbofan engine degradation using semi-supervised deep architecture. Reliab Eng Syst Saf 183:240–251

Ewees AA, Al-qaness MA, Abualigah L, Abd Elaziz M (2022) Hbo-lstm: optimized long short term memory with heap-based optimizer for wind power forecasting. Energy Convers Manage 268:116022

Hu C, Youn B, Wang P, yoon (2012) Ensemble of datadriven prognostic algorithms for robust prediction of remaining useful life. Reliab Eng Syst Saf 103:120–135

Isermann R (1984) Process fault detection based on modeling and estimation methods-a survey. Automatica 20(4):387–404

Jiang Y, Dai P, Fang P, Zhong RY, Zhao X, Cao X (2022) A2-lstm for predictive maintenance of industrial equipment based on machine learning. Comput Ind Eng 172:108560

Lee J, Wu F, Zhao W, Ghaffari M, Liao L, Siegel D (2014) Prognostics and health management design for rotary machinery systems-reviews, methodology and applications. Mech Syst Signal Process 42(1–2):314–334

Lei Y, Li N, Gontarz S, Lin J, Radkowski S, Dybala J (2016) A model-based method for remaining useful life prediction of machinery. IEEE Trans Reliab 65(3):1314–1326

Li Z, He Q (2015) Prediction of railcar remaining useful life by multiple data source fusion. IEEE Trans Intell Transp Syst 16(4):2226–2235

Li B, Li X, Zhao Z (2006) Novel algorithm for constructing support vector machine regression ensemble. J Syst Eng Electron 17(3):541–545

Li X, Ding Q, Sun J-Q (2018) Remaining useful life estimation in prognostics using deep convolution neural networks. Reliab Eng Syst Saf 172:1–11

Li J, Jia Y, Niu M, Zhu W, Meng F. Remaining useful life prediction of turbofan engines using cnn-lstm-sam approach, IEEE Sensors J

Mathew V, Toby T, Singh V, Rao BM, Kumar MG (2017) Prediction of remaining useful lifetime (rul) of turbofan engine using machine learning, In: 2017 IEEE international conference on circuits and systems (ICCS), IEEE, pp 306–311

Molnar C (2020) Interpretable machine learning, Lulu. com

Mosallam A (2014) Remaining useful life estimation of critical components based on bayesian approaches., Ph.D. thesis, Université de Franche-Comté

Polikar R, Polikar R (2006) Ensemble based systems in decision making. IEEE Circuit Syst Mag 6:21–45

Ranjan C, Reddy M, Mustonen M, Paynabar K, Pourak K. Dataset: rare event classification in multivariate time series, arXiv preprint arXiv:1809.10717

Sadigov R, Yıldırım E, Kocaçınar B, Akbulut F Patlar, Catal C (2023) Deep learning-based user experience evaluation in distance learning, Cluster Computing 1–13

Sateesh Babu G, Zhao P, Li X-L (2016) Deep convolutional neural network based regression approach for estimation of remaining useful life, In: International conference on database systems for advanced applications, Springer, pp 214–228

Si X-S, Wang W, Hu C-H, Zhou D-H (2011) Remaining useful life estimation-a review on the statistical data driven approaches. Eur J Oper Res 213(1):1–14

Soni H, Kansara A, Joshi T. Predictive maintenance of gas turbine using prognosis approach

Tian H, Yang L, Ju B (2023) Spatial correlation and temporal attention-based lstm for remaining useful life prediction of turbofan engine, Measurement 112816

Wang X, Li Y, Xu Y, Liu X, Zheng T, Zheng B (2023) Remaining useful life prediction for aero-engines using a time-enhanced multi-head self-attention model. Aerospace 10(1):80

Yildirim E, Akbulut FP, Catal C (2023) Analysis of facial emotion expression in eating occasions using deep learning, Multimedia Tools and Applications 1–13

Zhang X, Guo Y, Shangguan H, Li R, Wu X, Wang A (2023) Predicting remaining useful life of a machine based on embedded attention parallel networks. Mech Syst Signal Process 192:110221

Zheng S, Ristovski K, Farahat A, Gupta C (2017) Long short-term memory network for remaining useful life estimation, In: IEEE international conference on prognostics and health management (ICPHM). IEEE 2017:88–95