

# A Deep Learning Framework Based on Novel Hierarchical-LSTM Model for Enhanced Machinery Prognostics

Ahmed Ayman

School of Engineering and Computing  
University of Central Lancashire  
Preston, UK  
aaabdel-aal@uclan.ac.uk

Omneya Attallah

Department of Electronics and  
Communications Engineering  
Arab Academy for Science, Technology  
and Maritime Transport  
Alexandria, Egypt  
o.attallah@aast.edu

Ahmed Onsy

School of Engineering and Computing  
University of Central Lancashire  
Preston, UK  
aonsy@uclan.ac.uk

Hadley Brooks

School of Engineering and Computing  
University of Central Lancashire  
Preston, UK  
hlbrooks1@uclan.ac.uk

Iman Morsi

Department of Electronics and  
Communications Engineering  
Arab Academy for Science, Technology  
and Maritime Transport  
Alexandria, Egypt  
drimanmorsi@yahoo.com

**Abstract**—Machinery prognostics has garnered increasing research attention due to its critical role in industries such as manufacturing and renewable energy. Data-driven techniques, particularly recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have shown promise in accurately extracting features for estimating the remaining useful life (RUL) of machinery. However, the non-stationary and non-linear nature of machinery signals poses significant challenges to achieving accurate prognostics. This study introduces a novel hierarchical recurrent neural network method called hierarchical long short-term memory (H-LSTM) that is based on the long short-term memory (LSTM) model. H-LSTM is meant to address the problems with traditional RNNs that only use the previous time step for sequential data learning. It incorporates a hierarchical structure, enabling influence from multiple preceding time steps at each current step. Experimental evaluation on the FEMTO benchmark bearing dataset under varying operational conditions demonstrates that the proposed H-LSTM approach achieves up to fourfold improvements in performance compared to state-of-the-art methods, particularly for low signal-to-noise ratio (SNR) signals.

**Keywords**— Machinery prognostics, deep learning, recurrent neural networks, remaining useful life

## I. INTRODUCTION

The prediction of the remaining useful life (RUL) of rotating machinery is crucial for enhancing operational efficiency and preventing unexpected failures in various applications, including manufacturing, wind turbines, and space exploration [1], [2]. Predictive approaches for RUL estimation are broadly categorized into model-based and data-driven methods. Model-based techniques, such as particle filters, Kalman filters, and unscented filters, rely on mathematical models to describe equipment degradation but often require expert knowledge and struggle with non-linear and complex degradation patterns [3]. In contrast, data-driven approaches, leveraging machine learning and deep learning, have gained prominence due to their ability to automatically estimate equipment lifespan with higher reliability and reduced dependence on domain expertise [4], [5].

Convolutional neural networks (CNNs) have emerged as effective data-driven techniques for extracting spatial features in prognostics and health management (PHM) tasks. Ren et al. [6] demonstrated the capability of CNNs to analyze bearing

signal degradation and predict RUL, highlighting their efficiency in spatial feature learning. Enhancements to CNN-based approaches include incorporating statistical analyses, as proposed by Majali et al. [7] where time and spectral domain features are utilized during CNN training to improve lifespan prediction accuracy. Similarly, statistical feature extraction followed by autoencoder-based selection has been proposed by Ren et al. [3] to refine input features for CNNs. Despite these advancements, the reliance on preliminary statistical preprocessing steps may exclude non-stationary characteristics and increase computational demands [8]. Furthermore, conventional CNN approaches often neglect temporal correlations in degradation phases. Instead, they focus on Euclidean distance-based spatial feature extraction within grid structures, which makes it more useful for grid structure applications.

The gradual degradation of machinery during operation necessitates the analysis of temporal interdependencies between various conditions to predict the RUL. Recurrent neural networks (RNNs) are well-suited for capturing such temporal dependencies; however, issues like vanishing or exploding gradients during backpropagation hinder their performance [9]. Advanced RNN variants, such as long short-term memory (LSTM) and gated recurrent units (GRU), mitigate these challenges by utilizing gating mechanisms to retain relevant information. Studies such as Mao et al. [10] have demonstrated the efficacy of LSTM in temporal feature extraction for RUL prediction, while Chen et al. [11] incorporated empirical mode decomposition (EMD) as a preprocessing step to enhance LSTM's ability to learn from noisy signals. Other advancements include adversarial learning to filter noise and bidirectional LSTM (Bi-LSTM) to capture both forward and backward dependencies, as well as GRU-based architectures with health indicator frameworks [12]. Despite these innovations, limitations persist: EMD may introduce mode aliasing, adversarial approaches face challenges with imbalanced normal and fault state samples, and the sequential nature of LSTM and GRU models risks losing critical features due to reliance on stepwise data dependencies. These constraints underscore the need for further improvements in temporal modeling for machinery prognostics.

To address these challenges, this paper proposes a variant of the traditional LSTM named hierarchical LSTM (H-

LSTM), designed to learn various hidden patterns in machinery signals. The algorithm gets updated from an input vector of time data samples and the hidden states of an arbitrary number of subordinate units instead of relying only on one preceding unit in the conventional structure, enabling the formation of more complex network topologies. Each LSTM unit can thus incorporate information from multiple subordinate units, facilitating more robust feature learning. The proposed method was evaluated using the PRONOSTIA dataset, which comprises seventeen distinct bearings operating under various conditions. In this dataset, the bearings are subjected to different operational scenarios and run until failure, providing real-world performance data for evaluation.

This paper is structured as follows: Section 2 includes the methods and materials adopted; Section 3 delves into the methodology. Section 4 presents the results of the proposed study. Finally, Section 5 concludes the proposed approach.

## II. METHODS AND MATERIALS

### A. Data Description

The PRONOSTIA dataset is adopted for the evaluation purpose of this study [13]. This dataset consists of 17 run-to-failure experiments involving bearings under three distinct operating conditions, as detailed in Table I. To accelerate the degradation of the bearings, a radial force exceeding their maximum dynamic load was applied, leading to their failure within a few hours. During the tests, the bearing speed was kept constant. Data was collected using two accelerometers, which ensured precise measurement of bearing vibrations. A critical state, indicating bearing failure, is defined by a vibration signal exceeding 20 g, beyond which the bearing is considered non-operational.

The vibration data was sampled at a rate of 25,600 samples per second, providing high-resolution observations. Each sample, covering a duration of 0.1 seconds, consists of 2,560 data points. Data was recorded at regular 10-second intervals to capture consistent and accurate information throughout the experiments.

TABLE I. CHARACTERISTICS OF THE PRONOSTIA DATASET [13]

Characteristics	1	2	3
Load (N)	4000	4200	5000
Speed (rpm)	1800	1650	1500
Training	B1_1 B1_2 B1_3	B2_1 B2_2 B2_3	B3_1 B3_2
Validation/ Testing	B1_4 B1_5 B1_6 B1_7	B2_4 B2_5 B2_6 B2_7	B3_3

### B. Recurrent Neural Networks

Recurrent Neural Networks (RNNs) can handle input sequences of varying lengths by recursively applying a transition function to a hidden state vector  $h_t$ . At each time step  $t$ , the hidden state  $h_t$  is updated based on the input vector  $x_t$  at the current step and the hidden state from the previous step  $h_{t-1}$ . Typically, the transition function in an RNN consists of an affine transformation followed by a pointwise nonlinearity, such as the hyperbolic tangent function, as expressed in Equation 1:

$$h_t = \tanh(Wx_t + Uh_{t-1} + b) \quad (1)$$

However, a significant issue with RNNs using this transition function is that, during training, certain components of the gradient vector may either explode or vanish over long sequences [14].

### C. Proposed Hierarchical LSTM

One limitation of the RNN and its variants, such as LSTM and GRU architectures, is their reliance on strictly sequential information flow. To address this, we introduce the H-LSTM structure, which allows for more intricate network topologies. In this model, each LSTM unit can incorporate information from multiple subordinate units, enabling the system to capture richer hierarchical dependencies. Like standard LSTM units, each H-LSTM unit (indexed by  $j$ ) comprises input and output gates  $i_j$  and  $o_j$ , along with a memory cell  $c_j$  and hidden state  $h_j$ . The key distinction between conventional LSTM and H-LSTM units lies in how the gating vectors and memory cell updates depend on the states of potentially multiple child units. Unlike standard LSTMs, which have a single forget gate, each H-LSTM unit employs a unique forget gate  $f_{jk}$  for each child  $k$ , allowing selective integration of information from different child units. Figure 1 illustrates the internal structure of the proposed H-LSTM.

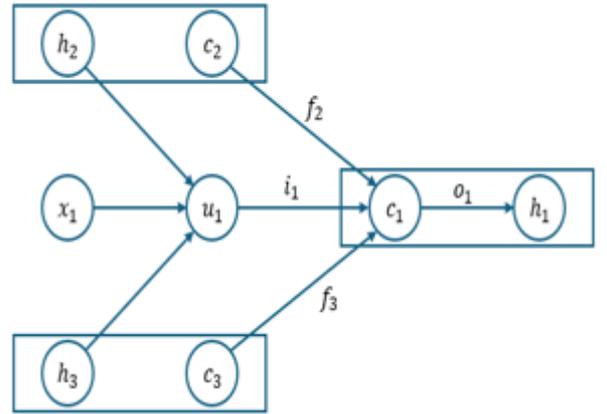


Fig. 1. Internal structure of H-LSTM

As with conventional LSTMs, each H-LSTM unit receives an input vector  $x_j$ , which in this context represents a specific point in the life cycle of the machinery being modelled. The time input  $x_j$  varies depending on the topology of the tree. Let  $(C_j)$  denote the collection of children for node  $j$  in a given tree. The H-LSTM transition equations are formulated as follows:

$$\tilde{h}_j = \sum_{k \in C(j)} h_k \quad (2)$$

$$i_j = \sigma(W^{(i)}x_j + U^{(i)}\tilde{h}_j + b^{(i)}) \quad (3)$$

$$f_{jk} = \sigma(W^{(f)}x_j + U^{(f)}h_k + b^{(f)}) \quad (4)$$

$$o_j = \sigma(W^{(o)}x_j + U^{(o)}\tilde{h}_j + b^{(o)}) \quad (5)$$

$$u_j = \tanh(W^{(u)}x_j + U^{(u)}\tilde{h}_j + b^{(u)}) \quad (6)$$

$$c_j = i_j \odot u_j + \sum_{k \in C(j)} f_{jk} \odot c_k \quad (7)$$

$$h_j = o_j \odot \tanh(c_j) \quad (8)$$

In these equations,  $k \in C(j)$  denotes a child of node  $j$ . Each parameter matrix encodes relationships between the input vector  $x_j$ , the hidden states of the child units  $h_k$ , and the components of the H-LSTM unit. In the case of a dependency tree application, the model can learn parameters such as  $W^{(i)}$  that allow the input gate  $i_j$  to open (i.e., approach values close to 1) when an important feature is presented, or close (i.e., approach values close to 0) when the input is relatively insignificant or noisy.

This flexible gating mechanism enables the model to integrate information in a more refinement manner, allowing the H-LSTM to better handle complex, hierarchical data structures.

### III. PROPOSED FRAMEWORK BASED ON H-LSTM

This section discusses the proposed framework based on the H-LSTM, which comprises four interconnected stages: data acquisition, data processing, feature learning, and RUL estimation.

Initially, the raw vibration signal from the monitored bearings was sampled at a frequency of 25.6 kHz. The raw signal was then divided into segments of equal length, with each segment containing 2,560 instantaneous accelerometer readings. Subsequently, the standard deviation (Std) was calculated for each segment, representing each window with a single value corresponding to its standard deviation. This process not only mitigates the low SNR commonly associated with complex machinery signals but also facilitates faster model convergence. After segmentation and the extraction of Std, the data is normalized using the min-max scaling technique. This normalization step is essential for adjusting the data range, which enhances the convergence speed of the model during training. For this experiment, the first three sets from the first and second operational conditions listed in Table 1 were used for training, while the remaining sets were reserved for testing. This approach maintains an 80-20 split between training and testing data, ensuring robust evaluation of the proposed method.

The data is subsequently input into the proposed H-LSTM model. However, the selection of hyperparameters is a critical aspect of the training process. Poorly chosen hyperparameters can lead to suboptimal model performance, underscoring the importance of their careful tuning during training. In this study each H-LSTM cell contains 32 subordinate cells, each has its independent memory cell, update gate, and set of weights. During backpropagation, the model's weights are adjusted after quantifying the error of each training epoch of the 30 epochs, ensuring that the model parameters are refined iteratively. Stochastic Gradient Descent has been selected as the optimizer, with a learning rate of 0.01, as this configuration demonstrated optimal performance for the model.

On the other hand, since the hardware specifications are crucial to ensure efficient handling of large datasets and minimize the training time for iterative experiments. This

experiment is conducted on a workstation with an Intel Core i7 and 16GB of RAM.

Finally, to evaluate the performance of the model, the mean squared error (MSE) is applied.

### IV. RESULTS AND DISCUSSION

This section presents the results of the proposed study. Initially, the procedure was repeated nine times, each targeting a different bearing set, namely B1\_4, B1\_5, B1\_6, B1\_7, B2\_4, B2\_5, B2\_6, B2\_7, and B3\_3. This approach adheres to the 80-20 percent training testing procedures discussed in the previous section.

Initially, the H-LSTM is trained, and the error is quantified. Table II presents the MSE evaluation results of the H-LSTM approach.

Furthermore, to comprehensively evaluate the proposed approach, the experiment was replicated using traditional LSTM and GRU models under identical configurations. The hyperparameters of the LSTM and GRU networks include 64 LSTM cells, with a dropout rate of 0.3 applied to mitigate overfitting. The training process remained consistent with that of the prior experiment. As displayed in Table III, the proposed H-LSTM model achieved significantly superior results compared to the conventional LSTM and GRU models, underscoring its advantages in this application.

Comparative analysis shows that the proposed model does better than traditional LSTM and GRU methods on all test bearing sets. However, it can be noticed that the performance of the proposed H-LSTM method is more significant on bearing sets with very low-SNR bearing sets like B2\_4, B2\_5, B2\_6, and B2\_7. This highlights the model's capability to effectively eliminate the irrelevant temporal characteristics depicted in the vibration signal. At the same time, the proposed method can be less significant in performance when the observed trend has high SNR and obvious failure threshold such as the case in test set B3\_3.

TABLE II. EVALUATION RESULTS OF THE PROPOSED METHOD FOR THE FIRST OPERATIONAL CONDITION SETTINGS

Set	B1_4	B1_5	B1_6	B1_7
MSE	0.025	0.023	0.02	0.019

TABLE III. EVALUATION RESULTS OF THE PROPOSED METHOD FOR THE SECOND AND THIRD OPERATIONAL CONDITION SETTINGS

Set	B2_4	B2_5	B2_6	B2_7	B3_3
MSE	$8.9 * 10^{-3}$	$8.7 * 10^{-3}$	$6 * 10^{-3}$	$7 * 10^{-2}$	0.024

TABLE IV. RESULTS OF THE PROPOSED METHOD COMPARED TO OTHER STUDIES EVALUATED ON THE PRONOSTIA DATASET

Bearing set	LSTM	GRU	H-LSTM
B1_4	0.083	0.053	0.025
B1_5	0.354	0.185	0.023
B1_6	0.0599	0.076	0.02
B1_7	0.1	0.08	0.019
B2_4	0.164	0.119	$8.9 * 10^{-3}$
B2_5	0.047	0.0255	$8.7 * 10^{-3}$
B2_6	0.719	0.6461	$6 * 10^{-3}$
B2_7	0.259	0.26	$7 * 10^{-2}$
B3_3	0.222	0.187	0.024

To further demonstrate the effectiveness of the proposed method, recent state-of-the-art approaches were reviewed to

provide a basis for comparison with our results. In [15], a CNN-multi-layer perceptron (MLP) approach was introduced, combining influential features extracted from both 1D time-domain and 2D grid structures to enhance RUL prediction. In contrast, [16], proposed a multi-scale CNN model specifically for bearing RUL prediction, focusing on the impact of spatial characteristics in machinery prognostics tasks. Additionally, [17] investigated a bearing prognostics method based solely on LSTM, alongside a comparative approach utilizing Bi-LSTM, to illustrate the potential benefits of Bi-LSTM in this context, denoted as 27.a and 27.b respectively. Table 4 lists the MSE results of the proposed H-LSTM and the results of the discussed studies.

The results clearly indicate that the proposed study outperforms existing approaches, underscoring the effectiveness of the H-LSTM model in handling machinery prognostics tasks. This superior performance suggests that the H-LSTM framework offers enhanced capabilities in capturing critical features and addressing the complexities associated with time-series data in comparison to conventional models. Consequently, these findings highlight the potential of the H-LSTM approach as a robust solution for improving accuracy and reliability in machinery condition monitoring and RUL prediction applications.

MSE	[15]	[16]	[17].a	[17].b	H-LSTM
<b>Bearing1_4</b>	0.017	0.157	0.18	0.142	0.025
<b>Bearing1_5</b>	0.035	0.1	0.452	0.44	0.023
<b>Bearing1_6</b>	0.1	0.1156	0.522	0.42	0.02
<b>Bearing1_7</b>	0.042	0.127	0.739	0.368	0.1
<b>Bearing2_4</b>	-	-	0.06	0.04	<b>8.9</b> <b>* 10<sup>-3</sup></b>
<b>Bearing2_5</b>	-	-	0.36	0.426	<b>8.7</b> <b>* 10<sup>-3</sup></b>
<b>Bearing2_6</b>	-	-	0.0529	0.050	<b>6 * 10<sup>-3</sup></b>
<b>Bearing2_7</b>	-	-	0.1216	0.002	<b>7 * 10<sup>-2</sup></b>

## V. CONCLUSION

In recent years, the critical role of machinery prognostics across various applications has garnered the attention of research. Nonetheless, existing data-driven methods face challenges in capturing influential temporal dynamics over extended periods. Moreover, prevalent temporal models such as LSTM and GRU often struggle under complex operating conditions, especially in environments with low SNR. This paper introduced a novel temporal learning approach that incorporates gate information from a set number of preceding cells into the learning process at each time step  $t$ . This approach effectively mitigated the noise learning while significantly improving lifespan estimation accuracy. Experiments on a benchmark-bearing dataset under diverse operational settings validate the proposed method, demonstrating up to four times better results on a signal of low SNR. Future research will explore the inclusion of spatial features to further advance prognostic capabilities. That's in addition to demonstrating its applicability in other rotating machinery applications.

## REFERENCES

- [1] L. Ren, J. Cui, Y. Sun, and X. Cheng, "Multi-bearing remaining useful life collaborative prediction: A deep learning approach," *J Manuf Syst*, vol. 43, pp. 248–256, 2017.
- [2] M. A. Djeziri, S. Benmoussa, and E. Zio, "Review on health indices extraction and trend modeling for remaining useful life estimation," *Artificial Intelligence Techniques for a Scalable Energy Transition: Advanced Methods, Digital Technologies, Decision Support Tools, and Applications*, pp. 183–223, 2020.
- [3] H. Liu, W. Song, Y. Niu, and E. Zio, "A generalized cauchy method for remaining useful life prediction of wind turbine gearboxes," *Mech Syst Signal Process*, vol. 153, p. 107471, 2021.
- [4] J. Deutsch and D. He, "Using deep learning-based approach to predict remaining useful life of rotating components," *IEEE Trans Syst Man Cybern Syst*, vol. 48, no. 1, pp. 11–20, 2017.
- [5] L. Ren, Y. Sun, J. Cui, and L. Zhang, "Bearing remaining useful life prediction based on deep autoencoder and deep neural networks," *J Manuf Syst*, vol. 48, pp. 71–77, 2018.
- [6] F. Cheng, L. Qu, and W. Qiao, "Fault prognosis and remaining useful life prediction of wind turbine gearboxes using current signal analysis," *IEEE Trans Sustain Energy*, vol. 9, no. 1, pp. 157–167, 2017.
- [7] N. Gebraeel, Y. Lei, N. Li, X. Si, and E. Zio, "Prognostics and remaining useful life prediction of machinery: advances, opportunities and challenges," *Journal of Dynamics, Monitoring and Diagnostics*, pp. 1–12, 2023.
- [8] L. Ren, Y. Sun, H. Wang, and L. Zhang, "Prediction of bearing remaining useful life with deep convolution neural network," *IEEE access*, vol. 6, pp. 13041–13049, 2018.
- [9] A. Majali, A. Mulay, V. Iyengar, A. Nayak, and P. Singru, "Fault identification and remaining useful life prediction of bearings using Poincare maps, fast Fourier transform and convolutional neural networks," *Mathematical Models in Engineering*, vol. 8, no. 1, pp. 1–14, 2022.
- [10] Z. Zhao, J. Wu, T. Li, C. Sun, R. Yan, and X. Chen, "Challenges and opportunities of AI-enabled monitoring, diagnosis & prognosis: A review," *Chinese Journal of Mechanical Engineering*, vol. 34, no. 1, p. 56, 2021.
- [11] Y. Chen, G. Peng, Z. Zhu, and S. Li, "A novel deep learning method based on attention mechanism for bearing remaining useful life prediction," *Appl Soft Comput*, vol. 86, p. 105919, 2020.
- [12] W. Mao, J. He, J. Tang, and Y. Li, "Predicting remaining useful life of rolling bearings based on deep feature representation and long short-term memory neural network," *Advances in Mechanical Engineering*, vol. 10, no. 12, p. 1687814018817184, 2018.
- [13] Z. Chen, Y. Liu, and S. Liu, "Mechanical state prediction based on LSTM neural network," in *2017 36th Chinese control conference (CCC)*, IEEE, 2017, pp. 3876–3881.
- [14] M. S. Rathore and S. P. Harsha, "Prognostics analysis of rolling bearing based on bi-directional LSTM and attention mechanism," *Journal of Failure Analysis and Prevention*, vol. 22, no. 2, pp. 704–723, 2022.
- [15] P. Nectoux *et al.*, "PRONOSTIA: An experimental platform for bearings accelerated degradation tests," in *IEEE International Conference on Prognostics and Health Management, PHM'12.*, IEEE Catalog Number: CPF12PHM-CDR, 2012, pp. 1–8.
- [16] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.
- [17] C.-G. Huang, H.-Z. Huang, Y.-F. Li, and W. Peng, "A novel deep convolutional neural network-bootstrap integrated method for RUL prediction of rolling bearing," *J Manuf Syst*, vol. 61, pp. 757–772, 2021.
- [18] J. Zhu, N. Chen, and W. Peng, "Estimation of bearing remaining useful life based on multiscale convolutional neural network," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 4, pp. 3208–3216, 2018.
- [19] M. S. Rathore and S. P. Harsha, "Prognostics analysis of rolling bearing based on bi-directional LSTM and attention mechanism," *Journal of Failure Analysis and Prevention*, vol. 22, no. 2, pp. 704–723, 2022.