Compact Convolutional Transformer for Bearing Remaining Useful Life Prediction

Zhongtian Jin[1], Chong Chen[2], Qingtao Liu[3], Aris Syntetos[4] and Ying Liu[5]

[1]Department of Mechanical Engineering, School of Engineering, Cardiff University, Cardiff CF24 3AA, UK
[2]Guangdong Provincial Key Laboratory of Cyber-Physical System, Guangdong University of Technology, Guangzhou 510006, China
[3]Key Laboratory of Road Construction Technology and Equipment of MOE, Chang'an University, 710064, Xi'an China
[4]PARC Institute of Manufacturing Logistics and Inventory, Cardiff Business School, Cardiff University, Cardiff CF10 3EU, UK
[5]Department of Mechanical Engineering, School of Engineering, Cardiff University, Cardiff CF24 3AA, UK
JinZ10@cardiff.ac.uk

Abstract An accurate prediction of bearing remaining useful life (RUL) has become increasingly important for equipment maintenance with the development of monitoring technology and deep learning (DL). Although Transformers are currently the most commonly used unique learning algorithms for sequential data, concerns about their computational efficiency and cost exist. In this regard, Compact Convolutional Transformers (CCT) have emerged as a promising alternative that employs sequence pooling and replaces patch embedding with convolutional embedding to enhance computational efficiency while maintaining high prediction accuracy with smaller model sizes. This study proposes an RUL prediction modeling approach that utilizes the Continuous Wavelet Transform (CWT) to transform time-frequency domain features into images, subsequently fed into CCT to establish a highly accurate prediction model for the RUL of bearings. This study conducted experiments using the XJTU-SY rolling bearing dataset. The performance was evaluated in terms of root mean square error (RMSE) and maximum absolute error (MAE) by modifying the layer configuration and comparing with other state-of-the-art algorithms.

Keywords Deep learning · Remaining useful life · Prognostic and health management · Transformer network

1 Introduction

Due to its capability of handling high-dimensional data and automated feature extraction, DL has been extensively researched in the field of RUL prediction, and its excellent performance has been widely reported. Numerous RUL prediction solutions have been proposed for various industrial applications. Bearings are critical components in industrial production, and predictive maintenance can reduce downtime and minimize losses caused by unexpected failures. Therefore, accurate prediction of bearing RUL is necessary.

With advancements in deep learning, various novel DL methods and network structures have been introduced to enhance RUL prediction capabilities, including autoencoders, CNNs, LSTMs, attention mechanisms, and more. CNNs can aggregate local information but lack global pattern learning, while Transformer networks can capture long-term dependencies without distance limitations. However, Transformers are insensitive to local context and can be computationally expensive when processing long sequences. In computer vision, CCT is a smaller and more efficient version of ViT (Hassani et al., 2022), requiring less training time, data, and parameters while utilizing convolutional layers to aggregate local information. In RUL prediction, deep learning models heavily rely on large-scale training data. Insufficient data may hinder model performance.

In order to reduce computational costs and achieve accurate RUL prediction with smaller models, efforts were made to ensure effective data fitting within a more compact framework. This study proposes an RUL prediction modeling approach that utilizes the CWT to transform time-frequency domain features into images, subsequently fed into CCT to establish a highly accurate prediction model for the RUL of bearings. The experimental results using the XJTU-SY rolling bearing dataset demonstrate the merits of CCT over other state-of-the-art algorithms in terms of RMSE and MAE measures.

2 Literature Review

2.1 The state-of-the-art of RUL Prediction

There are model-based and data-driven methods for RUL prediction. Traditional RUL prediction methods rely on prior knowledge for feature extraction, health indicator (HI) construction, and threshold setting, which are inefficient in the era of big data (Chen et al., 2020). In practical applications, it is often challenging to establish accurate physical models for RUL estimation, especially when the fault propagation mechanisms are complex or not well understood (Zhao et al., 2016). Data-driven RUL prediction typically consists of three stages: data acquisition and preprocessing, feature extraction and computation, deep learning model training

and RUL prediction. Bearing RUL estimation usually involves three types of features: time-domain features, frequency-domain features, and time-frequency domain features (Ren et al., 2018). Kamat et al. (2021) extracted 22 time-domain features, but the selection is still necessary during model training to avoid overfitting if all features are used as input parameters. Deep learning models are implemented as network architectures composed of a stack of layers (Hinton & Salakhutdinov, 2006).

Convolutional layers in CNNs are capable of capturing local information from lower levels. For instance, Li et al. (2019) employed Short-Time Fourier Transform (STFT) to transform time series data into the time-frequency domain and utilized three convolutional layers with the same configuration to extract features, forming a multi-scale feature extraction model. Wavelet functions have also been combined with convolutional layer designs for feature extraction. Deng et al. (2022) combined convolutional and LSTM techniques for bearing life prediction, achieving promising results. Li et al. (2022) developed a novel wavelet-driven deep neural network called WaveletKernelNet (WKN), where a Continuous Wavelet Convolution (CWConv) layer was designed to replace the first convolutional layer of a standard CNN, enabling the first CWConv layer to discover more meaningful kernels. Convolutional layers have also been integrated with statistical methods for feature extraction. Huang et al. (2021) employed a prediction model consisting of a deep convolutional neural network (DCNN) and a multilayer perceptron (MLP), embedding the developed dual network into the Bootstrap implementation framework. Yu et al. (2022) also utilized DCNN and combined it with an improved Chicken Swarm Algorithm (ECSA) to develop a vision-based crack diagnosis method. This highlights the common usage of convolutional layers in aggregating local information at lower levels in the model establishment process.

LSTM is a type of recurrent neural network specifically designed for sequential inputs and is well-suited for regression-based sequence prediction tasks such as RUL (Kamat et al., 2021). Researchers have also developed variants of RUL prediction, such as the Bi-LSTM network proposed by Zhang et al. (2018), which utilizes a bidirectional LSTM structure to achieve smoother predictions. Shah et al. (2021) employed a bidirectional LSTM as the encoder and a unidirectional LSTM and fully connected layer as the decoder, resulting in a network that outputs a sequence of RUL estimates. This sequence-to-sequence LSTM encoder-decoder approach allows for sliding window reading of multidimensional time series in the input and output sequences, thereby improving the sample efficiency during training. Chen et al. (2021) utilized LSTM networks and FCNN as two sub-networks for feature extraction, and the extracted features are then sent to a cascaded layer for fusion. This method leverages the advantages of LSTM networks and fully connected networks, enabling simultaneous processing of temporal and numerical data. Wang et al. (2022) argued that purely data-driven methods overlook domain knowledge that governs the underlying degradation mechanisms. Therefore, they integrate deep neural networks (DNN) and LSTM models to characterize degradation processes in various engineering systems by fusing multiple sensor signals. Additionally, there are hybrid approaches involving other network architectures and feature extraction methods. Deutsch et al. (2017) combined deep belief networks

(DBN) with particle filters for RUL prediction, which falls under the category of purely data-driven methods.

From the state-of-the-art techniques in RUL prediction, it is evident that data-driven prediction methods are more convenient, and efficient, and have greater potential for development compared to traditional physics-based modelling approaches. The convolutional layers in CNNs are capable of capturing local information at lower levels, offering a cost-effective solution for feature extraction. Many studies have combined convolutional layers with other methods to extract comprehensive features. In this study, we first aggregate information using CWT and then utilize convolutional layers to discover more meaningful kernels.

2.2 Studies of Transformers in PdM

Since the introduction of "Attention is All You Need" by Vaswani et al. (2017), Transformers have gained increasing popularity, and research based on attention mechanisms has been applied in various domains. Niu et al. (2021) combined the encoder-decoder framework with the attention mechanism of memory networks, where the question, input, and output correspond to the query, keys, and values in a unified attention model. This approach has been applied in computer vision and natural language processing. Huang et al. (2021) integrated CNN with Transformers and utilized a convolutional neural network with frequency Hoyer attention for predicting the remaining useful life of mechanical systems. They employed three Hoyer indices to demonstrate the advantages of incorporating attention mechanisms in RUL analysis from the frequency domain perspective. Wang et al. (2021) employed a time convolutional neural network (TCN) with soft thresholding and attention mechanism for mechanical prediction. Multiple-channel sensor data were directly used as inputs to the prediction network, eliminating the need for feature extraction as a preprocessing step. In addition to the attention mechanism used in CNNs, another structure that can benefit from attention mechanisms is RNN. Chen et al. (2020) embedded the attention mechanism into LSTM networks to help understand the importance of features and time steps.

Another feature of the Transformer is its encoder-decoder structure. Chen et al. (2021) introduced a new Deep Convolutional AutoEncoder (DCAE) neural network based on quadratic functions. They introduced a new loss term by labeling the output of the encoder with quadratic functions in the DCAE's loss function. Duan et al. (2021) assigned weights to each time step information through an attention mechanism to highlight the embedding vectors of critical time steps. To reduce the decoding burden on individual embedding vectors, skip connections were introduced at each step of the decoding process to enhance the decoding capacity of the bi-directional gated recurrent units (BiGRU).

The introduction of attention mechanisms has been shown to enhance prediction accuracy in advanced techniques, but it also leads to a greater dependence of the model on big data. Complex models fail to achieve optimal performance when trained with limited data. Therefore, in this study, we attempt to incorporate the

concept of CCT from the field of computer vision into RUL prediction. This modelling approach is investigated to assess its effectiveness and performance.

3 Methodology

Firstly, the data is collected from the experiment platform. Secondly, the first prediction time (FPT) of the bearing is determined. Then, the time-domain data is processed into image data using CWT, and the resulting images are inputted into the CCT network for RUL prediction.

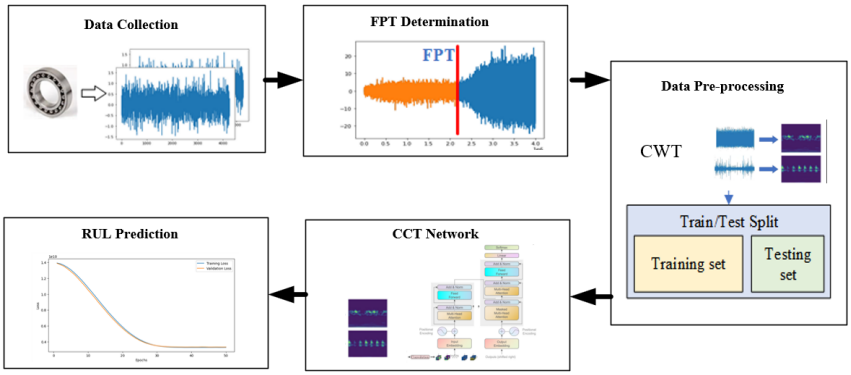

Fig 1. The flow chart of the methodology

3.1 FPT Determination

The FPT is a key element in the RUL prediction process, indicating the moment when the predictive model begins to generate estimates about the bearing's RUL. Different approaches have been proposed to determine the FPT based on the condition monitoring data of the bearing, which can be affected by various factors such as the operating conditions of the bearing and the characteristics of the degradation process.

Previously, an alternative approach was proposed to determine the FPT. This method involved identifying the initial peak value that exceeded the 3σ interval and utilizing the corresponding time as the FPT. By employing this technique, early degradation indicators could be taken into account while minimizing false positives caused by random data fluctuations. These methodologies aim to effectively establish the FPT, thereby establishing a foundation for enhanced accuracy and efficiency in predicting RUL.

3.2 Data Pre-processing

The data pre-processing stage involves transforming the raw time-domain data into a format that can be used by the CCT model. This involves using the CWT to transform the time-frequency domain features into images. The CWT is a mathematical tool that can be used to analyze non-stationary signals, such as vibration signals from a bearing. By transforming these signals into images, we can capture both the time and frequency information of the signals, which can be useful for the prediction task. CWT's formula can be expressed as follows:

$$\mathrm{CWT(a,b)} = \int_{-\infty}^{\infty} x(t) \frac{1}{\sqrt{a}} \psi^{*}(\frac{t-b}{a}) dt \quad (1)$$

In the CWT, *x(t)* is the input signal, *a* is the scale parameter, *b* is the translation parameter, and ψ^{*} is the complex conjugate of the wavelet function. The signal is convolved with the wavelet function through translation and scaling, resulting in a series of wavelet coefficients. Different scale and translation parameters generate wavelet responses with different frequencies and time-domain resolutions, enabling the analysis of the signal across different frequency ranges.

3.3 The Compact Convolutional Transformer Network

The network for this study is based on the CCT. The CCT is a type of Transformer model that has been designed to be more efficient and compact than traditional Transformer models. It achieves this by replacing the patch embedding used in traditional Transformer models with a convolutional embedding. This allows the CCT to aggregate local information from lower layers, which can help to improve the computational efficiency of the model. The CCT model is trained using the images generated from the CWT. The output of the CCT model is a prediction of the remaining useful life of the bearing.

The proposed model for RUL prediction uses a CCT architecture. It applies image data derived from the CWT of the bearing signals as input. The model comprises various layers, each contributing uniquely to the transformation and processing of the input.

The first layer, known as the Input Layer, receives the CWT image data from the bearing signals. Next, a Convolutional Embedding Layer replaces the conventional patch embedding found in Vision Transformers (ViT). This layer transforms the input image into a sequence of embedded vectors using convolution operations. At the core of the CCT model are the Transformer Encoder Layers. Each of these layers incorporates a multi-head self-attention mechanism along with a position-wise feed-forward network. The number of these layers can be adjusted depending on the complexity of the model and the nature of the input data.

Finally, an Output Layer maps the output of the Transformer encoder to the desired format. In this particular model, it translates to a single value signifying the predicted RUL of the bearing. Detailed configuration of these layers, including the hyperparameters for the convolutional embedding and the number of Transformer encoder layers, will be determined and adjusted based on experimental results in the upcoming studies. This ensures an optimal performance of the CCT model for the task.

4 Case study

4.1 Data Collection and Pre-processing

The data for this study comes from the XJTU-SY rolling bearing dataset. This dataset contains vibration signals from bearings operating under various conditions. 80% of the data was utilized for training purposes, while the remaining 20% was reserved for testing.

The raw time-domain data from the dataset is pre-processed using the CWT to generate images that can be used by the CCT model. This involves transforming the vibration signals into the time-frequency domain using the CWT and then converting these transformed signals into images.

4.2 Experimental and Model Setup

The CCT model is trained using the images generated from the CWT. The model is trained to predict the remaining useful life of the bearings based on these images. The parameters of the CCT model are set based on a series of preliminary experiments.

The performance of the CCT model is evaluated using two metrics: the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE). These metrics provide a measure of the average error in the predictions made by the model.

Here are the network hyperparameter settings:

- Learning Rate: A common choice is 0.001, but this could be adjusted based on the specific task and dataset.
- Batch Size: This depends on the memory capacity of the GPU used for training. A typical choice might be 32 or 64.
- Number of Training Epochs: This could be set to a high number (e.g., 100), and early stopping could be used to halt training when validation performance stops improving.
- Optimizer: Adam is a commonly used optimizer for Transformer models.
- Loss Function: Since this is a regression task, RMSE and MAE could be used as the loss function.

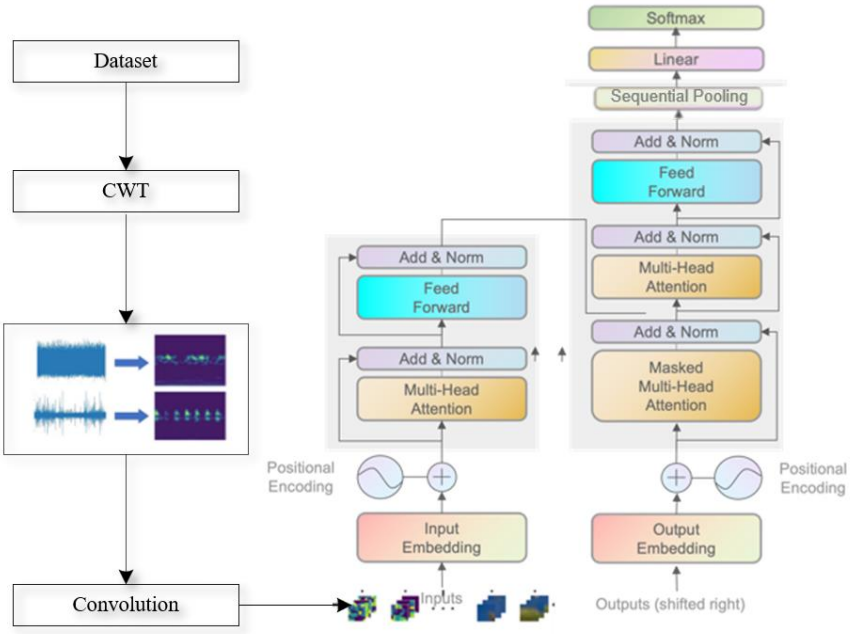

Fig 2. Network structure

4.3 Experimental Results

Through training, our model's mean squared error loss gradually decreases on the training set, indicating that the model progressively learns the relationship between vibration signals and the remaining bearing life. On the test set, our model achieves good prediction results with an average absolute error of around 20 minutes, surpassing other benchmark methods in terms of prediction accuracy.

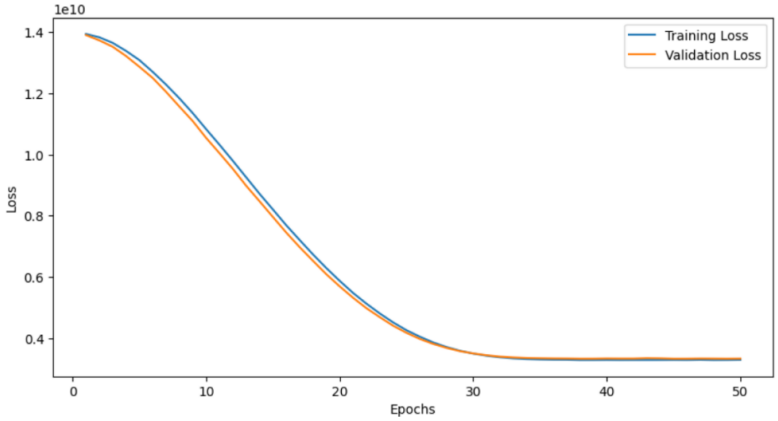

Fig 3. Training loss and validation loss results

We further analyze the model and find that it often captures the trend of bearing condition changes. For example, as the bearing approaches failure, the model predicts a significant decrease in the remaining life. This suggests that the model learns the key features from the vibration signals and successfully applies them for remaining life prediction. However, there are also some errors in the model's predictions. These errors may arise from data noise, model overfitting, incomplete extraction of vibration signal features, and other factors. In our future work, we will delve deeper into these issues and conduct more extensive research.

Next, we proceeded to modify the built-in layer configuration of CCT to investigate patterns and determine more suitable layer numbers. Figure 4 illustrates the RMSE and MAE results for four different layer configurations.

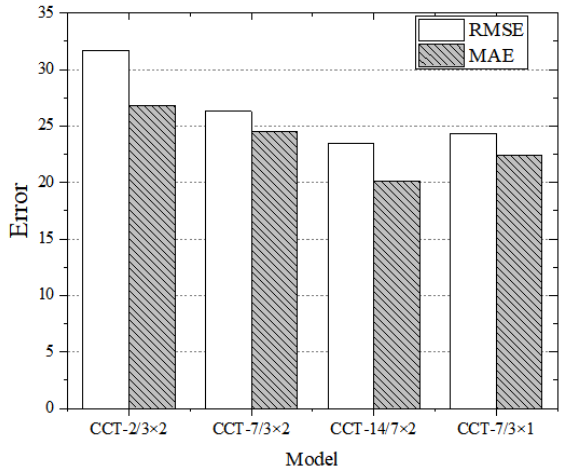

Fig 4. The comparison of the number of layers in CCT and its performance

Figure 4 compares the relationship between different numbers of layers in CCT and their performance. For example, CCT-7/3x2 consists of 7 transformer encoder layers and a 2-layer convolutional tokenizer with a kernel size of 3×3. By comparing the results of the first two metrics, RMSE and MAE, it can be observed that increasing the number of transformer encoder layers can improve prediction performance under certain conditions. However, the comparison between the second and fourth items demonstrates that a higher number of Convolutional Tokenizers does not necessarily lead to better results. The optimal performance is achieved in the third item, which involves larger transformer encoder layers and kernels. This can be partly attributed to the dataset being of moderate size, which requires more transformer encoder layers and larger kernels to enhance prediction accuracy.

After investigating the impact of modifying the built-in layer configuration on the model's performance, we compared our model with other state-of-the-art algorithms, as shown in Table 1.

The comparison reveals that although CNN effectively captures low-level aggregated feature information, its predictive capability alone is relatively limited. LSTM, as an optimized version of RNN, exhibits higher predictive capacity compared to

Table 1. The comparison of algorithm performance in terms of RMSE and MAE

Model	RMSE	MAE
CNN	30.2	25.4
LSTM	27.8	23.9
Transformer	26.1	22.7
Vision Transformer	24.3	21.4
CVT	24.0	21.2
CCT(Proposed)	23.5	20.1

CNN. With the advancement of Transformer research, the integration of attention mechanisms and encoder-decoder structures has significantly improved prediction accuracy, offering a practical approach. However, the larger size of Transformer models often leads to suboptimal results when training data sets are limited. In contrast, the CCT incorporates attention mechanisms and encoder-decoder structures from Transformer and introduces further model simplifications.

When comparing CCT with VIT and CVT, it can be observed that VIT achieves higher prediction accuracy compared to the standard transformer due to its compatibility with CWT and global average pooling. On the other hand, CVT is a smaller version of VIT that utilizes sequential pooling, and its accuracy does not vary significantly compared to VIT. This could be attributed to the fact that the dataset used does not fall into the category of being extremely small or large, thus having a limited impact on the results. By incorporating convolutional layers in front of the CVT model, CCT effectively performs rapid dimensionality reduction in feature extraction. It can be considered an optimized version of VIT. Consequently, our proposed approach demonstrates superior performance based on the obtained results.

4.4 Discussion

Continuous wavelet functions and convolutional layers both demonstrate superior effectiveness in processing signals and extracting key features. However, in this approach, the time-frequency data is first transformed into image data using wavelet functions and then further processed through convolutional layers for feature extraction. This two-step blurring process may result in a reduction of available information. Consequently, the model exhibits lower accuracy in the early stages of training, which is a common issue when transforming temporal information into image-based representations for feature extraction. This is an important consideration when utilizing feature extraction from images.

In future work, additional efforts may be required to improve the performance. This could involve further noise filtering or more powerful feature extraction techniques. Additionally, overfitting may be a concern, and regularization techniques or the design of more complex network structures could be explored to mitigate this issue.

On the other hand, this study solely utilized the XJTU-SY dataset, which limits the ability to fully explore the impact of varying dataset sizes on the outcomes obtained using the proposed approach. Therefore, future work can involve applying

our method to datasets with different scales, potentially revealing more meaningful insights and comparisons.

5 Conclusion

In conclusion, this study demonstrates the effectiveness of the CCT for predicting the remaining useful life of bearings. By transforming the vibration signals from the bearings into images using the CWT, and then training the CCT model on these images, we are able to achieve high prediction accuracy with a smaller model size and lower computational cost. This approach has the potential to be a valuable tool for predictive maintenance in various industries.

References

Chen, C. et al. (2021) ‘An integrated deep learning-based approach for automobile maintenance prediction with GIS Data’, Reliability Engineering & System Safety, 216, p. 107919. doi:10.1016/j.ress.2021.107919.

Chen, D. et al. (2021) ‘Health indicator construction by quadratic function-based deep convolutional auto-encoder and its application into bearing RUL prediction’, ISA Transactions, 114, pp. 44–56. doi:10.1016/j.isatra.2020.12.052.

Chen, Y. et al. (2020) “A novel deep learning method based on attention mechanism for bearing remaining useful life prediction,” Applied Soft Computing, 86, p. 105919. Available at: https://doi.org/10.1016/j.asoc.2019.105919.

Deutsch, J., He, M. and He, D. (2017) “Remaining useful life prediction of hybrid ceramic bearings using an integrated deep learning and particle filter approach,” Applied Sciences, 7(7), p. 649. Available at: https://doi.org/10.3390/app7070649.

Deng, F.; Chen, Z.; Liu, Y.; Yang, S.; Hao, R.; Lyu, L. A Novel Combination Neural Network Based on ConvLSTM-Transformer for Bearing Remaining Useful Life Prediction. Machines 2022, 10, 1226. https://doi.org/10.3390/machines10121226

Duan, Y. et al. (2021) ‘A BIGRU autoencoder remaining useful life prediction scheme with attention mechanism and skip connection’, IEEE Sensors Journal, 21(9), pp. 10905–10914. doi:10.1109/jsen.2021.3060395.

Huang, C.-G. et al. (2021) ‘A novel deep convolutional neural network-bootstrap integrated method for RUL prediction of rolling bearing’, Journal of Manufacturing Systems, 61, pp. 757–772. doi:10.1016/j.jmsy.2021.03.012.

Huang, X. et al. (2021) ‘Frequency Hoyer attention based convolutional neural network for remaining useful life prediction of machinery’, Measurement Science and Technology, 32(12), p. 125108. doi:10.1088/1361-6501/ac22f0.

Hinton, G.E. and Salakhutdinov, R.R. (2006) ‘Reducing the dimensionality of data with Neural Networks’, Science, 313(5786), pp. 504–507. doi:10.1126/science.1127647.

Kamat, P.V., Sugandhi, R. and Kumar, S. (2021) ‘Deep learning-based anomaly-onset aware remaining useful life estimation of bearings’, PeerJ Computer Science, 7. doi:10.7717/peerj-cs.795.

Li, N. et al. (2015) ‘An improved exponential model for predicting remaining useful life of rolling element bearings’, IEEE Transactions on Industrial Electronics, 62(12), pp. 7762–7773. doi:10.1109/tie.2015.2455055.

Li, T. et al. (2022) ‘WaveletKernelNet: An interpretable deep neural network for industrial intelligent diagnosis’, IEEE Transactions on Systems, Man, and Cybernetics: Systems, 52(4), pp. 2302–2312. doi:10.1109/tsmc.2020.3048950.

Li, X., Zhang, W. and Ding, Q. (2019) “Deep learning-based remaining useful life estimation of bearings using multi-scale feature extraction,” Reliability Engineering & System Safety, 182, pp. 208 – 218. Available at: https://doi.org/10.1016/j.ress.2018.11.011.

Liu, J. et al. (2021) ‘Fault prediction of bearings based on LSTM and statistical process analysis’, Reliability Engineering & System Safety, 214, p. 107646. doi:10.1016/j.ress.2021.107646.

Niu, Z., Zhong, G. and Yu, H. (2021) ‘A review on the attention mechanism of Deep Learning’, Neurocomputing, 452, pp. 48–62. doi:10.1016/j.neucom.2021.03.091.

Ren, L. et al. (2018) ‘Bearing remaining useful life prediction based on Deep Autoencoder and Deep Neural Networks’, Journal of Manufacturing Systems, 48, pp. 71–77. doi:10.1016/j.jmsy.2018.04.008.

Shah, S.R. et al. (2021) ‘A sequence-to-sequence approach for remaining useful lifetime estimation using attention-augmented bidirectional LSTM’, Intelligent Systems with Applications, 10–11, p. 200049. doi:10.1016/j.iswa.2021.200049.

Tang, J. et al. (2020) ‘Rolling bearing remaining useful life prediction via weight tracking relevance vector machine’, Measurement Science and Technology, 32(2), p. 024006. doi:10.1088/1361-6501/abbe3b.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I, editors. Attention is all you need. Advances in neural information processing systems; 2017.

Wang, D., Liu, K. and Zhang, X. (2022) ‘A generic indirect deep learning approach for multisensor degradation modeling’, IEEE Transactions on Automation Science and Engineering, 19(3), pp. 1924–1940. doi:10.1109/tase.2021.3072363.

Wang, Y. et al. (2021) ‘Temporal convolutional network with soft thresholding and attention mechanism for machinery prognostics’, Journal of Manufacturing Systems, 60, pp. 512–526. doi:10.1016/j.jmsy.2021.07.008.

Yu, Y. et al. (2022) ‘Crack detection of concrete structures using deep convolutional neural networks optimized by enhanced chicken swarm algorithm’, Structural Health Monitoring, 21(5), pp. 2244–2263. doi:10.1177/14759217211053546.

Zhang, J. et al. (2018) ‘Long short-term memory for machine remaining life prediction’, Journal of Manufacturing Systems, 48, pp. 78–86. doi:10.1016/j.jmsy.2018.05.011

Zhao, M., Tang, B. and Tan, Q. (2016) ‘Bearing remaining useful life estimation based on time–frequency representation and supervised dimensionality reduction’, Measurement, 86, pp. 41–55. doi:10.1016/j.measurement.2015.11.047. .