Autonomous Intelligent Systems

ORIGINAL ARTICLE Open Access

Enhanced bearing RUL prediction based on dynamic temporal attention and mixed MLP

Zhongtian Jin[1], Chong Chen[2], Aris Syntetos[3] and Ying Liu[1*]

Abstract

Bearings are critical components in machinery, and accurately predicting their remaining useful life (RUL) is essential for effective predictive maintenance. Traditional RUL prediction methods often rely on manual feature extraction and expert knowledge, which face specific challenges such as handling non-stationary data and avoiding overfitting due to the inclusion of numerous irrelevant features. This paper presents an approach that leverages Continuous Wavelet Transform (CWT) for feature extraction, a Channel-Temporal Mixed MLP (CT-MLP) layer for capturing intricate dependencies, and a dynamic attention mechanism to adjust its focus based on the temporal importance of features within the time series. The dynamic attention mechanism integrates multi-head attention with innovative enhancements, making it particularly effective for datasets exhibiting non-stationary behaviour. An experimental study using the XJTU-SY rolling bearings dataset and the PRONOSTIA bearing dataset revealed that the proposed deep learning algorithm significantly outperforms other state-of-the-art algorithms in terms of RMSE and MAE, demonstrating its robustness and accuracy.

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

1 Introduction

The widespread use of rotating machinery is in the aerospace, automotive manufacturing, and textile manufacturing industries. These machines typically work in hostile settings and under variable loadings, exposing them to the possibility of faults that pose significant security threats [1, 2]. Bearing, as essential parts of machinery, has a major effect on the reliable operation. The accurate prediction of the remaining useful life (RUL) of bearings is important for predictive maintenance [3], preventing unexpected failures and increasing the lifespan of machinery [4]. The traditional RUL prediction methods are highly dependent upon feature extraction and expert knowledge, which takes much time and cost [5].

Many feature extraction techniques have been explored in the realm of RUL prediction to increase accuracy. In particular, due to their capability to utilize spatial hierarchies in data [6, 7], convolutional neural networks (CNNs) are particularly effective. Other techniques have been used to extract important features from complex signals such as Principal Component Analysis (PCA) [8] and Wavelet Transforms [9]. Although these techniques can consume much computational resources and may have difficulty with nonstationary data as is common in bearing degradation processes, there are other possible solutions. However, classical deep learning algorithms such as Long Short-Term Memory (LSTM) networks [10] and Recurrent Neural Networks (RNNs) [1] show promising results in practice, but they are not computationally or scalable efficient.

For the last few years, models with both feature extraction techniques, such as wavelet transforms and convolution and multi-head attention mechanisms have been shown to perform exceptionally well in predicting bearing life. While these models have achieved success, their robustness remains insufficient, especially when applied to nonstationary data characteristic of the bearing degrada-

*Correspondence: liuy81@cardiff.ac.uk
[1] Department of Mechanical Engineering, School of Engineering, Cardiff University, Cardiff CF24 3AA, UK
Full list of author information is available at the end of the article

tion process. Moreover, the high computational overhead, as well as the inconsistency of performance across different datasets demonstrate a clear demand for stronger and more generalizable models.

To address these challenges, this paper proposes a hybrid approach: CWT for feature extraction, a Channel-Temporal Mixed MLP layer for exploiting the intricate dependencies, and a dynamic attention mechanism. Using CWT and CT-MLP, we enhanced feature extraction and captured complex temporal dependencies in the deep learning model that we developed. This model makes the model robust to non-stationary data through a dynamic attention mechanism that focuses on relevant features over time. Our approach leads to improvement of prediction accuracy and generalization capability and is validated on comprehensive evaluations with XJTUSY and PRONOSTIA bearing datasets. These results highlight the model's potential to enhance the reliability and efficiency of predictive maintenance strategies, facilitating broader adoption in industrial applications. In summary, the main contributions of this study are:

- The integration of Continuous Wavelet Transform (CWT) for effective feature extraction from non-stationary signals and a Channel-Temporal Mixed MLP (CT-MLP) layer to capture complex dependencies in the data.
- The development of a dynamic attention mechanism that adapts to temporal importance, enhances the model's ability to focus on relevant features over time, thereby improving robustness and prediction accuracy.
- Comprehensive experimental validation on the XJTU-SY and PRONOSTIA bearing datasets, demonstrating superior performance in terms of RMSE and MAE compared to state-of-the-art algorithms.

The rest of this paper is organized as follows. This thesis begins by reviewing prior work in RUL prediction and on the use of wavelet transforms and attention mechanisms in either speech recognition or other sequence prediction tasks in Sect. 2. In Sect. 3, we present our proposed methodology including the model architecture and dynamic attention mechanism. In Sect. 4, we describe the experimental setup, datasets used, and our results for the ablation studies and parameter tuning. The results are presented in Sect. 5 and further discussed in detail. Section 6 is the conclusion of the paper and a sift of future directions for research.

2 Literature review

2.1 The studies of feature extraction in RUL prediction

Model-based (or physics-based) and data-driven models are two broad categories of RUL estimation methods [11]. However, physical models in [12] for RUL estimation are difficult to specify in practical applications in which fault propagation mechanisms are complex and not well understood. Since robust data-driven RUL prediction is reliant on efficient extraction of representative features through appropriate signal processing techniques, the focus of this work is on the selection of a signal representation method that is best suited to RUL prediction. Typically, data-driven RUL prediction encompasses three stages: The second step deals with data acquisition and processing, feature extraction and computation, and finally deep learning model training and RUL prediction [13]. Bearing RUL prediction generally involves three types of features: There are time-domain features, frequency-domain features, and time-frequency domain features. Features in the time domain directly extract statistical attributes from the raw time series data. It has been documented in the literature that up to 22 such time domain features exist for RUL prediction [14]. Nevertheless, using these features as input parameters to predict RUL can result in overfitting and requires optimization of the feature extraction technique for improved RUL prediction outcomes. One technique is the feature attention mechanism [15], in which input features are dynamically weighted to allow the model to place more emphasis on the most important attributes. However traditional prediction methods such as Recurrent Neural Networks (RNNs) suffer from problems such as gradient explosion [1] and inspired us to employ Long Short-Term Memory (LSTM) [16] networks for temporal prediction.

Two of the most frequently employed feature extraction methods either make direct use of deep learning models (e.g., Convolutional Neural Networks (CNNs)) for extracting features from time series data or use signal processing techniques like Short-Time Fourier Transform (STFT), Continuous Wavelet Transform (CWT), or Hilbert Huang Transform (HHT). Extensive Fourier-based methods have been used, such as CNNs for multi-scale feature extractions with the help of STFT to obtain the time-frequency information from original data [17]. MFCC and STFT spectra are also used to extract time-frequency features [18]. In particular, the wavelet functions are increasingly becoming popular for the ability to simultaneously capture frequency and location information, which is more appropriate for the analysis of nonstationary signals, hence analysis is increasingly performed with CWT, Discrete Wavelet Transform (DWT), Wavelet Packet Decomposition (WPD), and so on. For example, a multivariate time series is decomposed into wavelet domains using DWT [19], and graphical convolution is used to extract the feature and model inter-variable relationships. Graph convolutional networks (GCNs) capture frequency domain and time domain features, and modelling variable dependencies at different resolutions improve long sequence prediction accuracy.

Since CWT is primarily explained in several studies, however, DWT can be employed for signal decomposition into frequency components to individualize features associated with the motor degradation, or faults. This enables more complex patterns to be revealed that may not be captured easily by traditional methods increasing fault diagnosis and possible reliability for fault detection under varying load conditions. For instance, CWT produces scaleograms that capture energy variability concerning time-frequency scales, which increases leakage detection ability. A combination of the advantages of the CWT and the STFT allows a detailed interpretation of the acoustic emission signals of oil pipelines and thus the accuracy and reliability of oil leakage detection systems can be improved [20]. The flexibility in analyzing the signal frequency is one for which CWT has accomplished its name while the computational overhead is large. However, DWT enables efficient computation and multi-resolution analysis and hence is appropriate for signal denoising and compression. In engineering and computer science settings, DWT is usually used as a signal encoding tool whereas CWT is the preferred tool for signal analysis in scientific research [21]. The wide acceptance of wavelet transforms as a replacement for Fourier transforms in many domains is indisputable, and this has occurred even to the point where they have taken the place of Fourier transforms in many application fields. The primary distinction between DWT and CWT lies in their operational approach: CWT operates in all possible scales and shifts; DWT works with a particular subset of scale and shift values. Therefore, DWT is employed for the position encoding and learning from segmented data signals. For research purposes, we use CWT, but it doesn't cover the entire research basis. This is highly suitable given the fact that the fault signals degrade gradually.

Finally, I conclude that feature extraction techniques are of great importance to RUL prediction model accuracy and efficiency. Different methods like, time domain, frequency domain, etc. have been studied but wavelet transforms like CWT, and DWT have a better competence in processing a nonstationary signal. Furthermore, with the aid of these techniques, combined with state-of-the-art deep learning models, we can build much more robust and generalizable RUL prediction models, allowing for the adoption of more trustworthy preventive maintenance strategies in industrial settings.

2.2 The studies of attention mechanism in PHM

Time series prediction has been a hallmark task for long Short-Term Memory (LSTM) networks since they can learn long-range dependencies in sequential data. Further, try to improve the performance of the standard LSTM and a ton of variants have been proposed. One example is a method for estimating remaining useful life (RUL) via a sequence-to-sequence LSTM encoder-decoder structure [22]. A sliding window is being used to read a sequence of multi-dimensional time series sequences of our inputs and outputs, which has a tremendous effect on the sample efficiency of the training. In particular, the M-LSTM model employs two subnetworks of features extracted from layered Fully Connected Neural Networks (FCNN) and layered Long Short Term Memory (LSTM) networks. The extracted features are used to integrate with a cascading layer and ultimately used to deploy the FCNN for the final RUL prediction [23]. In [24], Niazi et al. showed that utilization of a TT-ConvLSTM model is effective in handling spatiotemporal dependencies for RUL prediction of bearings on time-series data. The second approach utilizes Deep Neural Networks (DNN) on top of the LSTM model, creating Health Indices (HI) that seamlessly combine multiple sensor signals into the degradation process for multiple engineering systems. This method also leverages domain knowledge (i.e., failure thresholds and the monotonicity of the degradation process) for improved interpretability [25]. The LSS model combines the advantages of LSTM networks and statistical process analysis on bearing vibration signal temporal features extraction and feeds the multilevel signals into LSS for prediction [26]. Further, bidirectional LSTM (Bi LSTM) networks have been used to capture dependency across forward and backward time directions enhancing the accuracy of time series prediction [27].

Transformers and attention mechanisms have spied the light on the world of time series prediction, especially for large datasets. These are exceedingly resource-hungry models yet they manage to deliver very good quality. For example, a new deep feature learning method is proposed for RUL estimation in [28] by using Time-Frequency Representation (TFR) and Multi-Scale Convolutional Neural Network (MSCNN). MTS Mixers, a framework for multivariate time series forecasting using factorized temporal and channel mixing to capture the dependency and be more efficient than the traditional transformer-based approach is proposed by Li et al. [29]. An alternative model, known as the Multi-Head Neural Network (MHNN), predicts the RUL of industrial equipment by utilizing an asymmetric constraint and an architecture consisting of bidirectional gated recurrent units (BGRU) and self-attention mechanisms to extract temporal features from condition monitoring data [30]. Additionally, the robustness and accuracy of RUL predictions have been improved using deep adversarial neural networks [31]. DCNNs in combination with multi-layer perceptrons (MLP) in dual network architecture have also been proposed for feature extraction and RUL estimation [32]. More specifically, a multi-domain adversarial network with stacked convolutional autoencoders is used to reduce discrepancies in extracted degradation features to improve the feature transfer process [33].

We also found success in the deployment of gated convolutional unit layers as the first hidden layer, followed by linear layers and position encoding operations to extract high-level features before the data is fed into transformer blocks [34]. The vanishing gradient problem is addressed using deep attention residual neural networks (DARNN) [35] and deep residual networks that use skip connections [36] to achieve better learning capabilities in RUL prediction models. Further prediction accuracy has been improved by integrating a dual network within a Bootstrap framework. In this method, three Hoyer indices are used to assess the important contribution of different frequency components to the bearing degradation from the frequency domain perspective [37]. In general, the RUL prediction models with advanced attention mechanisms and transformer models have performed better. These methods capture subtle temporal and frequency domain features which are especially suited for nonstationary and complex datasets.

Over the past few years, however, many modern time series prediction algorithms have started to incorporate attention mechanisms for better interpretability and performance. Examples of such global-local attention mechanisms in RNNs already exist for datasets with seasonal characteristics, [38] that learn to capture local and global dependencies in time series datasets. In this approach, a simple multi-scale framework is employed that uses downsampling convolution to obtain local features and isometric convolution for capturing global correlations and offers a balance between computational efficiency and the capability of extracting complex temporal correlations. Since data on bearing failure generally does not have seasonality, we turn to dynamic attention mechanisms. They are dynamically controlled concerning both context and content, permitting models to focus on pertinent elements to the task at hand and to neglect less important data, improving performance. There are variations of dynamic attention which are multi-head adaptive attention, shifted window dynamic attention, dynamic sparse attention, and hybrid adaptive attention.

A form of dynamic attention mechanism is the multihead Gaussian Adaptive Attention Mechanism (GAAM), which adapts attention by Gaussian distribution parameters (mean and variance) to account for changes in the data [39]. Although computationally expensive, this method enables greater accuracy and adaptiveness to nonstationary data provided valid knowledge of the data distribution is available. Another giant advancement is the development of dynamic sparse attention mechanisms in vision transformers. These mechanisms are intended to control computational and memory limits only where there is a need. Applying a two-stage routing method to implement this strategy, has been seen to improve performance in visual tasks like image classification and object detection by reducing unneeded computations and focusing on semantically meaningful regions [40]. In addition, we proposed a dynamic attention mechanism to enhance the robustness of the transformer-based model to adversarial attacks. The second method, which consists of dynamically adjusting attention weights to minimize the influence of the inputs that could mislead the output of the model [41], is applied. The Multi-Scale Fusion Transformer (MSFT) is another innovative approach that fuses dynamic attention mechanisms for time series prediction. The MSFT model encompasses local and global information in a time series dataset, and this works dynamically so that more emphasis is provided to a more significant event or anomaly in the dataset. At the cost of simplicity, this approach excels at integrating multi-scale data to achieve a holistic view that improves prediction accuracy [42]. Fu et al. [43] presented a dual-task learning framework to perform First Prediction Time (FPT) detection and Remaining Useful Life (RUL) prediction in a united model, using a multichannel attention mechanism to calculate the importance of input parameters and extracted features adaptively and eventually provide more accurate and more adaptable predictions. It also uses an improved Temporal Convolutional Network (TCN) to leverage long-term dependencies within multidimensional time series data.

These advanced methods focus on specific local anomalies and global overall trends improving sensitivity to changes in bearing conditions over time. Introducing layers to learn local features like particular wear patterns together with the layers corresponding to systemic trends in overall degradation improves the ability to predict RUL. To develop robust, generalizable models that can perform effective predictive maintenance in industrial applications, this integration is key.

3 Methodology

This section outlines the methodology employed for predicting the remaining useful life (RUL) of bearings using a combination of Continuous Wavelet Transform (CWT) for feature extraction and a dynamic attention mechanism within a multi-head attention framework to enhance prediction accuracy. Figure 1 shows the flow chart of the methodology. Additionally, we introduce a Channel-Temporal Mixed MLP (CT-MLP) layer to capture intricate dependencies within the time series data. The methodology consists of several key steps: Data acquisition and preprocessing, feature extraction using CWT, dynamic attention mechanism implementation, and the integration of the CT-MLP layer.

3.1 Data acquisition and preprocessing

Data from the XJTU-SY rolling element bearing dataset and the PRONOSTIA bearing dataset are utilized for model training and evaluation. The datasets consist of

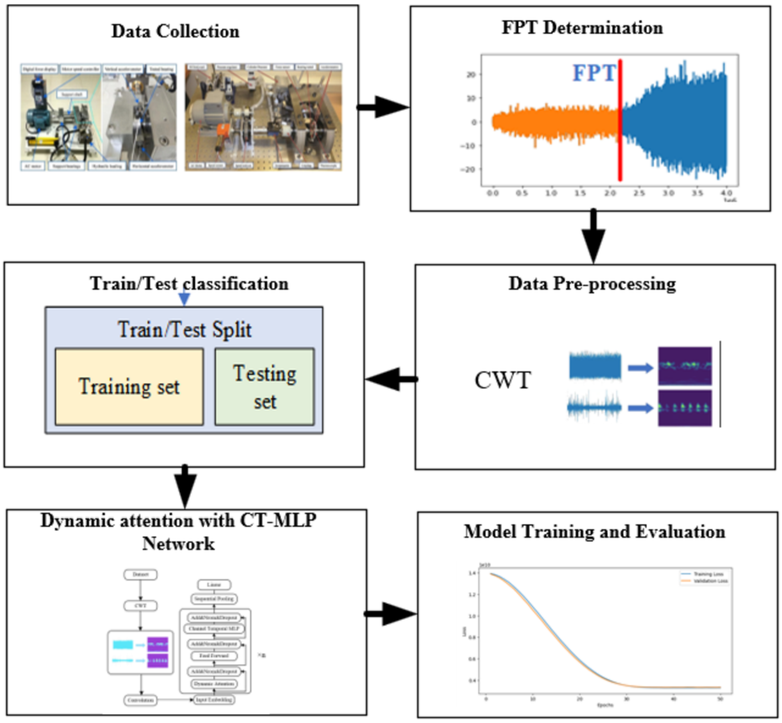

Figure 1 The flow chart of the methodology

time-series data capturing the degradation process of bearings under various operational conditions. Initially, the raw vibration signals are subjected to preprocessing steps, including normalization and denoising, to ensure the quality and consistency of the input data.

3.2 Feature extraction using Continuous Wavelet Transform (CWT)

The time-domain signals are then converted into time-frequency domain representations by using CWT, which can capture temporal and frequency characteristics of the vibration signals at the same time. The ability to analyze nonstationary signals, prevalent in bearing degradation, makes this transformation particularly desirable. The decomposition of time-series data into wavelet coefficients is achieved by convolving the signal with scaled and translated versions of a wavelet function. Being able to track oscillatory behaviours in the data, we decide to use the Morlet wavelet. CWT's formula can be expressed as follows:

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

The output of the CWT is a scaleogram, an image-like representation where the x-axis represents time, the y-axis represents a scale (related to frequency), and the pixel intensity corresponds to the magnitude of the wavelet coefficients. These scaleograms serve as input features for the subsequent deep learning mode.

3.3 Feature extraction with convolutional layers

High-level features are automatically extracted by supplying CWT scaleograms to convolutional layers. The CNN layers make use of multiple consecutive convolutional and pooling layers that aim to encode spatial hierarchies of the scaleogram at multiple scales. These layers aid in the discovery of informative patterns indicative of bearing health. More specifically, convolutional layers scan the scaleograms against filters to identify specific features, and the pooling layers reduce the spatial dimensionality of the feature maps to make the maps more manageable — and less prone to overfitting — by performing maximum, minimum, or average operations. In the CNN layers, the output is a set of feature maps, which are flattened and further processed by the transformer-based architecture. We can write the convolutional operation as:

$$x_j^l=\sigma(\sum_{i\in M_j}x_i^{l-1}*k_{ij}^l+b_j^l) \tag{2}$$

3.4 Dynamic attention mechanism for time series prediction

The model is endowed with a multi-head attention mechanism to dynamically focus on the most relevant features

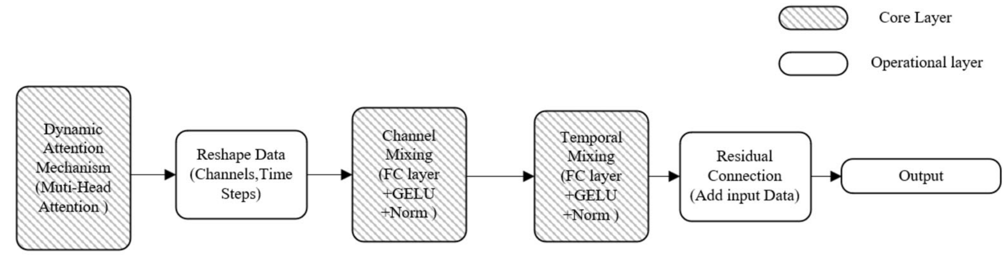
Core Layer
Operational layer
Dynamic Attention Mechanism (Muti-Head Attention)
Reshape Data (Channels,Time Steps)
Channel Mixing (FC layer +GELU +Norm)
Temporal Mixing (FC layer +GELU +Norm)
Residual Connection (Add input Data)
Output

Figure 2 Data Flow from Dynamic Attention Mechanism to CT-MLP

to better predict RUL. The multi-head attention mechanism allows us to project the input into several sub-spaces so that the model can attend to different parts of the input sequence at once. The improved convergence of the model to capture the complex spatiotemporal dependencies within the data. The dynamic attention mechanism focuses on how important the features in the time series are with a dynamic attention rectification that attenuates less important features by weakening their attention or masking through zeroing out their attention. Moreover, this process is essential for dealing with non-stationary data and cases of varying operational conditions. Also, the dynamic attention mechanism dynamically creates the set of the candidate features and adjusts their attended scores in every transformer layer, thus the model focuses on various features at different times and highlights the most relevant information at all times.

3.5 Channel-Temporal Mixer MLP (CT-MLP) layer

Following the dynamic attention mechanism, the CT-MLP layer is designed to capture intricate dependencies between different channels (features) and time steps in the multivariate time series data. Initially, the output from the dynamic attention mechanism is reshaped into a suitable format for the CT-MLP layer.

The CT-MLP layer operates in two stages: Channel mixing and temporal mixing. Assume the input feature tensor is $X \in R^{T \times C}$, where T is the sequence length and C is the number of channels.

Channel Mixing: A fully connected layer is applied along the channel dimension to model interactions between features within each time step:

$$X_C = \text{LayerNorm}(X) + \sigma(XW_C + b_C), \tag{3}$$

where:

$W_C \in R^{C \times d}$ and $b_C \in R^d$ are the weights and biases of the channel mixing layer, σ is a non-linear activation function (GELU in this case).

Temporal Mixing: A fully connected layer is applied along the temporal dimension to capture dependencies across time steps:

$$X_t = \text{LayerNorm}(X_c) + \sigma(X_c W_t + b_t), \tag{4}$$

where:

$W_t \in R^{T \times d}$ and $b_t \in R^d$ are the weights and biases for temporal mixing.

Residual Connections: Residual connections are included to ensure effective gradient flow and prevent vanishing gradients in deeper architectures.

The CT-MLP layer then applies a fully connected (dense) layer across the channel dimension to learn interactions between different features at each time step, followed by another fully connected layer across the temporal dimension to capture dependencies and patterns over time for each feature. Figure 2 shows data flow from dynamic attention mechanism to CT-MLP. Each fully connected layer is followed by a non-linear activation function, GELU, and a normalization layer, Layer Norm, to ensure stable training and improve model performance. Residual connections are incorporated to facilitate gradient flow and prevent vanishing gradients, ensuring that the model can learn effectively even with deep architectures. The specific hyperparameters for this layer include a learning rate of 0.001, a batch size of 64, two layers for both channel and temporal mixing, and a dropout rate of 0.1.

3.6 Model architecture

The key components of the overall architecture are, amongst other things. We first input the CWT scaleograms to CNN layers, which extract spatial features from the scaleograms and yield feature maps. The transformer encoder maximizes the feature representation by processing these feature maps through multi-head and dynamic attention mechanisms. Finally, we obtain the attention-enhanced feature maps, which are further processed by the Channel temporal Mixed MLP (CT MLP) layer. Figure 3 shows the network structure. The CT MLP layer fuses channel and temporal information in the time series data, leveraging their inherent dependencies to help the model better learn complex patterns and relationships

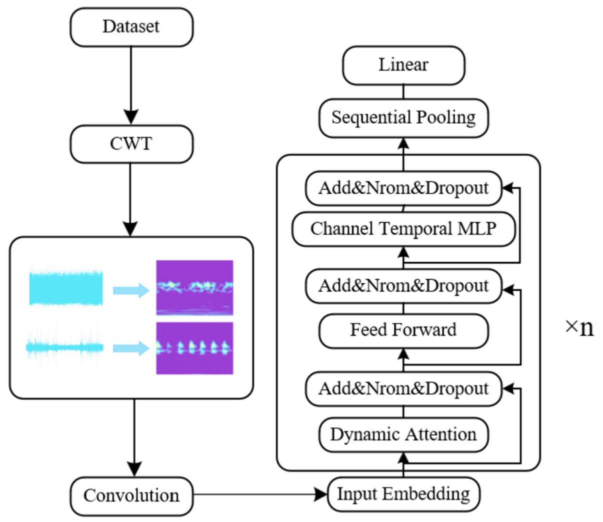

Figure 3 Network structure

in the input data. The hybrid of channel mixing and temporal mixing, this layer provides a much more complete picture. In this architecture, the tentative phase combines the strengths of CWT, which captures fine-grained details in the data, as well as the dynamic attention mechanism, which captures broad trends in the data. It is further fed into a linear layer to predict the RUL of the bearings.

3.7 Training and evaluation

The XJTU-SY and PRONOSTIA datasets are used to train the model. In the training process, lots of Hyperparameters have to be tuned to make the model perform better. We systematically performed hyperparameter tuning with Bayesian optimization, a probabilistic model-based approach. The reason for this is that this method is very efficient in exploring big and complex hyperparameter spaces by repeating the search and refining it by using previous evaluations. In the convolutional neural network (CNN), the number of convolutional layers (3 to 5), kernel size (3 × 3 or 5 × 5), number of filters (32, 64, 128), the pooling size (2 × 2), activation function (ReLU), dropout (0.2 to 0.5) are chosen as parameters. Parameters for the transformer model were tuned, including the number of transformer layers (2 to 4), number of attention heads (4, 8, 12), model dimension (128, 256), feedforward network dimension (512, 1024), and the dropout rate (0.1 to 0.3). We also carefully select training parameters such as batch size (32, 64), learning rate (0.001 to 0.0001 with learning rate decay), optimizer (Adam), and number of epochs (50 to 100) so that the performance is optimized. Configurations that balanced predictive performance with computational cost were identified through the use of Bayesian optimization and ultimately selected a model with eight attention heads, a model dimension of 256, a feedforward dimension of 1024, four MLP layers with GELU activation, and a learning rate of 0.0005. We evaluate the performance of the model according to metrics such as root mean square error (RMSE) and maximum absolute error (MAE). The proposed approach is shown through a large number of experiments to provide not only better prediction accuracy but also better generalization skills on various datasets. In particular, the dynamic attention mechanism significantly improves the model's capacity to concentrate on key features, eliminating the model's sensitivity to operational conditions and signal characteristics' changes.

4 Experiment study

4.1 Data description

Choosing suitable datasets to evaluate our proposed model is very important. For our experiments, there was the XJTU-SY rolling element bearing dataset and the PRONOSTIA bearing dataset owing to their wide acceptance and use in the field of bearing RUL prediction. These are common datasets used to benchmark and validate the performance of several bearing prediction algorithms, and therefore, are good candidates to assess the robustness and accuracy of our model. The XJTU and PRONOSTIA datasets both are well-known public datasets due to their comprehensive and detailed recordings of bearing degradation in various operation conditions. We demonstrate the generalization ability of our model across various scenarios and data characteristics by using these datasets. Our results are therefore comparable to existing studies given that these well-established datasets are included in the validation framework.

The XJTU-SY bearing dataset contains complete run-to-failure data for 15 rolling element bearings. Figure 4 shows the XJTU-SY test bench [44]. Each subset of the XJTU-SY dataset (e.g., 1-1, 1-2) corresponds to a specific bearing tested under distinct operational conditions. The dataset includes three experimental conditions defined by combinations of rotational speed and radial force: 2100 RPM and 12 kN (Condition 1), 2250 RPM and 11 kN (Condition 2), and 2400 RPM and 10 kN (Condition 3). For each condition, five bearings were tested (e.g., Bearings 1_1 to 1_5 for Condition 1), with run-to-failure vibration signals sampled at 25.6 kHz every minute. Each subset also records the bearing's failure mode and actual lifespan, providing diverse scenarios for RUL prediction.

The PRONOSTIA dataset is part of the IEEE PHM 2012 Data Challenge and contains run-to-failure data for bearings under different operating conditions. Figure 5 shows the PRO-NOSTIA test bench [45]. The platform consists of a rotating part (asynchronous motor with gearbox), a degradation generation part (pneumatic jack applying ra-

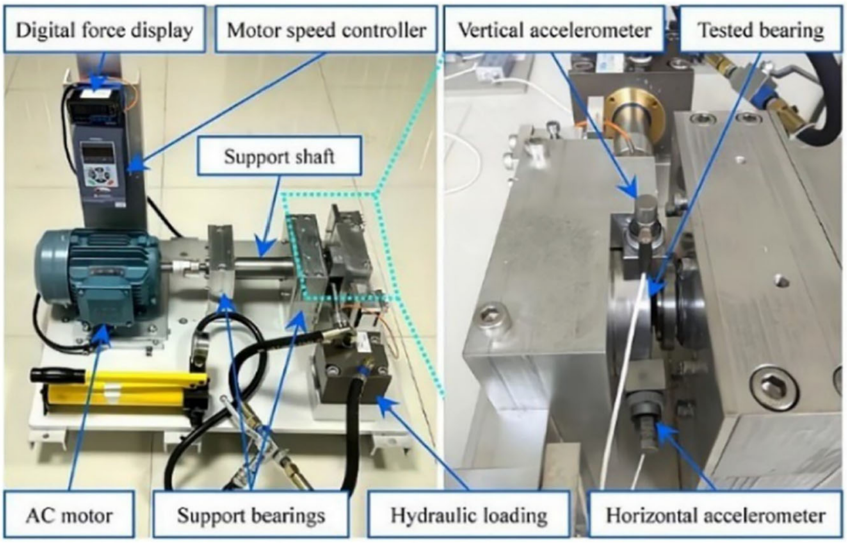

Figure 4 The XJTU-SY test bench [44]

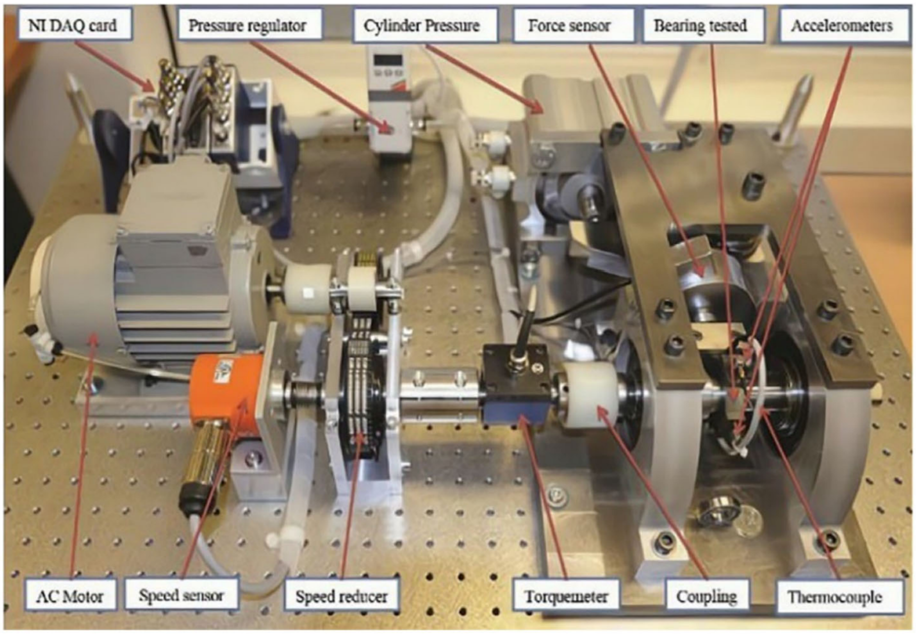

Figure 5 PRO-NOSTIA test bench [45]

dial force), and a measurement part (vibration and temperature sensors). In the accelerated degradation test, there are three operational conditions which are 1800 rpm and 4 kN, 1650 rpm and 4.2 kN, and 1500 rpm and 5 kN. Vibration signals were sampled at 25.6 kHz, with 2560 samples recorded every 10 seconds. Temperature signals were sampled at 10 Hz, with 600 samples recorded every minute.

4.2 Modal setup

The Modal steps involve several key processes, including data preprocessing, feature extraction, and model training. The raw vibration signals are first normalized and denoised to ensure consistency and quality. The normalization process adjusts the amplitude of the signals to a common scale, while denoising removes unwanted noise that could affect the accuracy of the feature extraction process.

After preprocessing, the time-domain signals are converted into time-frequency domain representations using Continuous Wavelet Transform (CWT). The CWT decomposes the time-series data into wavelet coefficients by convolving the signal with scaled and translated versions of a wavelet function. The Morlet wavelet is chosen due to its suitability for capturing oscillatory behaviours in the data. The output of the CWT is a scaleogram, an image-like representation where the x-axis represents time, the y-axis represents a scale (related to frequency), and the pixel intensity corresponds to the magnitude of the wavelet coefficients. These scaleograms are then fed into convolutional layers to automatically extract high-level features. The CNN layers consist of multiple convolutional and pooling layers designed to capture spatial hierarchies within the scaleogram. These layers help in extracting relevant patterns that are indicative of bearing health. The output of the CNN layers is a set of feature maps, which are then flattened and passed on to the transformer-based architecture for further processing.

During model training, we tune many hyperparameters to find the best-performing one. The CNN parameters like the number of conv layers, kernel size, number of filters, pool size, activation function, and dropout rate are changed. Parameters such as the number of transformer layers, the number of attention heads, model dimension, feedforward network dimension, and dropout rate, are tuned for the transformer model. Careful selection of their training parameters – batch size, learning rate with decay, optimizer, and number of epochs – is performed for the best performance. The temporal importance of features in the time series is used to adjust the mechanism of the dynamic attention diagnostic, focusing on where it matters most. Attention rectification corrects attention scores that have been assigned to different features, i.e., masks or weakens the attention that has been given to less important features. This allows for handling nonstationary data and varying operating conditions. Moreover, the dynamic attention mechanism dynamically constructs and updates the attention scores of the set of candidate features in each transformer layer. With this, the model will emphasize different features at different times to always highlight the most important information.

The performance of the model is evaluated based on metrics such as root mean square error (RMSE) and maximum absolute error (MAE). Extensive experiments demonstrate that the proposed approach improves prediction accuracy and generalization capability across different datasets. The dynamic attention mechanism enhances the model's ability to focus on relevant features, particularly making it robust to variations in operational conditions and signal characteristics.

4.3 Ablation experiments

We do ablation experiments in this section to measure the importance of the MLP layer and dynamic attention on model performance. Specifically, we compare the original model with three variants: I compare two variants of our overlay, one without the MLP layer (NoMLP), one without the dynamic attention mechanism (NoAttention), and one without both MLP layer and dynamic attention mechanism (NoMLP+NoAttention). The XJTU-SY and PRONOSTIA bearing datasets are then used in these experiments.

- Original Model: Incorporates both the MLP layer and dynamic attention mechanism.
- NoMLP Model: Removes the MLP layer while retaining the dynamic attention mechanism.
- NoAttention Model: Removes the dynamic attention mechanism while retaining the MLP layer.
- NoMLP+NoAttention Model: Removes both the MLP layer and dynamic attention mechanism for comparison.

To identify the optimal number of attention heads and MLP layers, a systematic series of experiments was conducted on both the XJTU-SY and PRONOSTIA datasets. The number of attention heads (H) was varied across (4, 8, 12), and the number of MLP layers (L) was tested with configurations of 2 and 4 layers. Each combination of H and L was evaluated using root mean square error (RMSE) and mean absolute error (MAE) as performance metrics.

4.4 Compare with different models

We validated the superiority of the proposed approach through comparison with several state-of-the-art (SOTA) models, which are well used in the area of Remaining Useful Life (RUL) prediction. Baseline models include traditional as well as recent deep learning-based models, providing a complete evaluation. For benchmark datasets in industrial applications on RUL prediction, the comparison was conducted on the XJTU-SY and PRONOSTIA-bearing datasets. Included in this comparison are models:

- CNN: Convolutional Neural Network for feature extraction from time-series data.
- ConvLSTM: A hybrid Convolutional LSTM network that combines spatial and temporal feature learning.
- MLP-MSCNN: A Multi-Layer Perceptron integrated with Multi-Scale CNN for capturing multi-resolution features.
- CNN-ResNet: Convolutional ResNet architecture leveraging residual connections to enhance feature learning.
- TT-ConvLSTM: The TT-ConvLSTM model effectively combines tensor-train (TT) decomposition and Convolutional Long Short-Term Memory (ConvLSTM) networks to handle spatiotemporal dependencies in time-series data.

These comparisons highlight the diversity of baseline models considered, ranging from traditional CNNs to advanced hybrid architectures. This comprehensive evaluation ensures a fair and robust assessment of the proposed approach's performance relative to existing methods.

5 Results and discussion

5.1 Ablation experiment

The results of the ablation study on the XJTU-SY dataset can be found in Table 1. The data shows that the model as originally designed featuring the MLP layer and dynamic attention is superior. We found that all the zero-shot variants in this model consistently outperformed the NoMLP and NoAttention variants, highlighting the key importance of both components to improved model performance. In particular, the RMSE was 0.195 and MAE was 0.179 for the original model in Test 1-1, compared to RMSE 0.210 and MAE 0.204 for NoMLP and RMSE 0.221 and 0.216 for NoAttention. Our results demonstrate the ability of the MLP layer, the dynamic attention mechanism, and the reported combination to improve both the model's prediction accuracy and robustness.

Results on the PRONOSTIA dataset (Table 2) show the original model's superior accuracy, with an RMSE of 0.168 and MAE of 0.154 in Test 1-1, outperforming NoMLP (RMSE 0.183, MAE 0.166) and NoAttention (RMSE 0.199, MAE 0.187).

The XJTU-SY bearing dataset and the PRONOSTIA bearing dataset were used for conducting the ablation experiments whose results are listed. The experiments aimed to evaluate the performance impact of two key components in the proposed model: the dynamic attention mechanism and the MLP layer. Four different model configurations were tested. From these ablation experiments, we show that the MLP layer and dynamic attention mechanism are key components to boosting the performance of the model. With these components added in, the Original Model can capture intricate dependencies, apply dynamic focus on important features through their temporal importance, and increase prediction accuracy and robustness.

Figure 6 shows the comparison of the number of layers in the model and its performance. In the first configuration with 4 attention heads and 2 MLP layers, the model reached RMSE equal to 0.195 and MAE 0.179. Increasing the MLP layers to 4 (and keeping the number of the attention heads constant) showed some improvement — RMSE and MAE both improved to 0.190 and 0.174. The results suggest that the more layers of MLP introduced into the model can help derive more complex dependencies, and hence increase prediction accuracy. For the case where the number of attention heads was increased to 8, with 2 MLP layers, the RMSE and MAE further decreased to 0.185 and 0.169 respectively. This showed that a larger number of attention heads results in the model implicitly focusing on various parts of the input features, thereby increasing the overall robustness and accuracy. With 8 attention heads and 4 MLP layers, the model can improve its performance with an RMSE of 0.180 and an MAE of 0.164. At first, for configurations with 12 attention heads, the model did better with 2 MLP layers (RMSE = 0.175, MAE = 0.159, accuracy = 42.5). Yet, as

Table 1 Ablation experiments in the XJTU-SY dataset

Metric		1-1	1-2	1-3	1-5	2-1	2-2	2-3	2-4	2-5	3-3	3-4	3-5
RMSE	Original	**0.195**	**0.220**	**0.162**	**0.173**	**0.122**	**0.143**	**0.159**	**0.149**	**0.126**	**0.179**	**0.153**	**0.209**
	Non-MLP	0.210	0.235	0.181	0.195	0.147	0.154	0.191	0.173	0.153	0.201	0.169	0.243
	Non-Attention	0.221	0.249	0.194	0.213	0.183	0.176	0.189	0.193	0.149	0.215	0.187	0.274
	Non-MLP+Non-Atten	0.314	0.251	0.218	0.357	0.221	0.240	0.255	0.247	0.199	0.248	0.239	0.313
MAE	Original	**0.179**	**0.199**	**0.127**	**0.156**	**0.099**	**0.131**	**0.140**	**0.142**	**0.113**	**0.154**	**0.137**	**0.194**
	Non-MLP	0.204	0.213	0.168	0.187	0.132	0.143	0.170	0.163	0.139	0.192	0.151	0.215
	Non-Attention	0.216	0.224	0.184	0.193	0.173	0.159	0.166	0.170	0.127	0.197	0.173	0.266
	Non-MLP+Non-Atten	0.279	0.233	0.190	0.315	0.211	0.219	0.240	0.232	0.176	0.224	0.199	0.304

Table 2 Ablation experiments in the PRONOSTIA dataset

Metric		1-1	1-2	1-3	1-4	1-5	1-7
RMSE	Original	**0.168**	**0.213**	**0.226**	**0.190**	**0.135**	**0.194**
	Non-MLP	0.183	0.251	0.246	0.214	0.174	0.210
	Non-Attention	0.199	0.276	0.257	0.229	0.195	0.233
	Non-MLP+Non-Atten	0.211	0.317	0.304	0.267	0.231	0.274
MAE	Original	**0.154**	**0.194**	**0.187**	**0.139**	**0.107**	**0.177**
	Non-MLP	0.166	0.233	0.201	0.207	0.155	0.189
	Non-Attention	0.187	0.250	0.231	0.210	0.167	0.211
	Non-MLP+Non-Atten	0.202	0.293	0.284	0.243	0.217	0.238

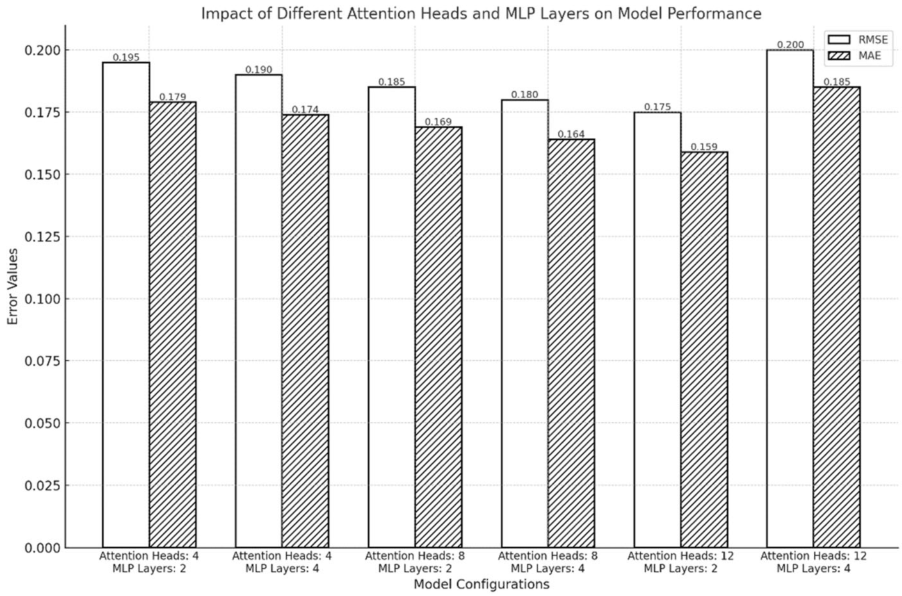

Figure 6 The comparison of the number of layers in the model and its performance

Table 3 Performance comparisons of different models for XJTU-SY bearing dataset

Test	CNN		ConvLSTM		MLP-MSCNN		CNN-ResNet		TT-ConvLSTM		Proposed(ours)	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
1-1	0.214	0.189	0.242	0.213	0.206	**0.176**	0.212	0.187	0.210	0.190	**0.195**	0.179
1-2	0.233	0.204	0.262	0.229	0.240	0.207	0.225	**0.195**	0.231	0.196	**0.220**	0.199
1-3	0.252	0.162	0.184	0.155	0.178	0.151	**0.153**	0.132	0.174	0.133	0.162	**0.127**
1-5	0.221	0.175	0.215	0.181	0.184	0.155	**0.160**	**0.148**	0.177	0.164	0.173	0.156
2-1	0.217	0.209	0.148	0.126	0.117	0.099	**0.115**	**0.097**	0.124	0.101	0.122	0.099
2-2	0.203	0.179	0.232	0.194	**0.122**	**0.102**	0.123	0.112	0.137	0.128	0.143	0.131
2-3	0.196	0.176	0.199	0.164	0.158	**0.126**	0.168	0.135	0.163	0.151	**0.159**	0.140
2-4	0.243	0.201	0.231	0.195	0.177	0.141	0.152	0.134	0.152	**0.131**	**0.149**	0.142
2-5	0.143	0.121	0.108	0.090	**0.091**	**0.075**	0.164	0.148	0.133	0.120	0.126	0.113
3-1	0.257	0.223	0.247	0.214	0.244	0.204	0.254	0.217	0,213	0.191	**0.206**	**0.181**
3-3	0.221	0.196	0.191	0.156	**0.158**	**0.129**	0.161	0.240	0.183	0.163	0.179	0.154
3-4	0.198	0.167	0.165	0.139	**0.132**	**0.107**	0.157	0.121	0.155	0.141	0.153	0.137
3-5	0.223	0.214	0.267	0.225	0.266	0.219	**0.207**	0.188	0.211	**0.187**	0.209	0.194

we move from 2 MLP layers to 4 MLP layers the RMSE and MAE jumped to 0.200 and 0.185. Therefore the results also suggest that going beyond a certain point, more MLP layers and attention heads do not improve the model accuracy but incur overfitting and higher computational cost.

The experiments show how, although adding more attention heads and MLP layers in general improves model performance, they point out an optimal combination after which performance plateaus or even degrades. Therefore, this balance must be managed carefully to avoid unnecessary computational overhead and to produce the most efficient and accurate predictions of the remaining useful life of bearings.

5.2 Comparison with different models

We also compare the performance of our proposed model with several state-of-the-art algorithms, including CNN, ConvLSTM, MLP-MSCNN, CNN-ResNet, and TT-ConvLSTM, on the XJTU-SY and PRONOSTIA bearing datasets. In Table 3, we observe that our proposed model has competitive performance across a variety of test cases and does especially well in cases with nonstationary and complicated data patterns. The proposed model demonstrates clear advantages in many cases but is not always the winning baseline under all conditions. It demonstrates the possible impact of dataset-specific characteristics and operational conditions on performance. However, the outcome highlights the efficacy of the proposed model for dealing with a wide range of challenging RUL prediction tasks when both training data are sufficient and the degradation pattern is consistent.

Figures 7 and 8 show the RMSE and MAE comparison for the XJTU-SY bearing dataset. It can be observed that the proposed method achieves the best results in most tests on the XJTU–SY bearing dataset, especially for RMSE and MAE. In Test 1-1, for instance, the proposed model yields an RMSE of 0.195 and an MAE of 0.179 while CNN's RMSE was 0.214 and MAE was 0.189. A similar trend can be seen across different test cases, and this shows the potential of the proposed model in dealing with complicated and nonstationary data.

This is mainly because CWT is used for feature extraction and the dynamic attention mechanism is applied. CWT provides effective time domain to time-frequency domain transformation of the raw time series data that is both temporal and spectral. Additionally, this is further en-

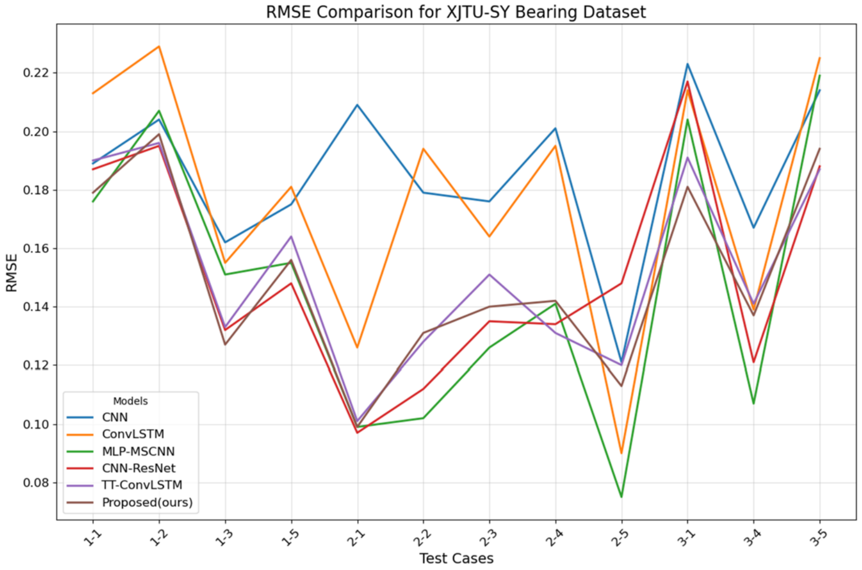

Figure 7 RMSE Comparison for XJTU-SY Bearing Dataset

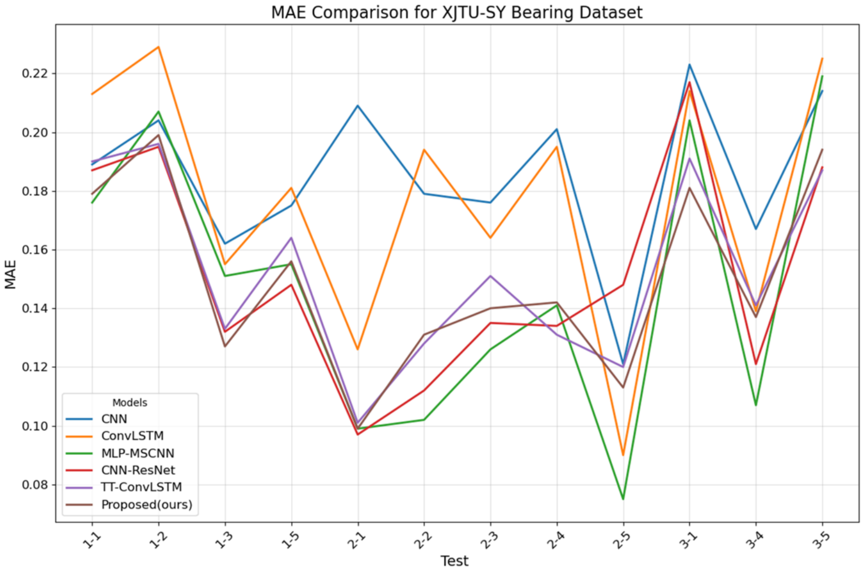

Figure 8 MAE Comparison for XJTU-SY Bearing Dataset

Table 4 The detailed result for the PRONOSTIA rolling element dataset

Test	CNN		ConvLSTM		MLP-MSCNN		CNN-ResNet		TT-ConvLSTM		Proposed(ours)	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
1-1	0.301	0.265	0.268	0.245	0.194	0.161	**0.157**	**0.135**	0.174	0.161	0.168	0.154
1-2	0.314	0.268	0.281	0.242	0.254	0.219	0.271	0.226	0.237	0.211	**0.213**	**0.194**
1-3	0.337	0.289	0.331	0.270	0.199	0.164	**0.180**	**0.154**	0.240	0.199	0.226	0.187
1-4	0.416	0.379	0.513	0.443	0.132	0.107	**0.148**	**0.125**	0.193	0.144	0.190	0.139
1-5	0.231	0.199	0.208	0.174	0.187	0.158	0.147	0.123	0.141	0.122	**0.135**	**0.107**
1-7	0.198	0.169	**0.165**	**0.141**	0.205	0.172	0.208	0.180	0.205	0.189	0.194	0.177

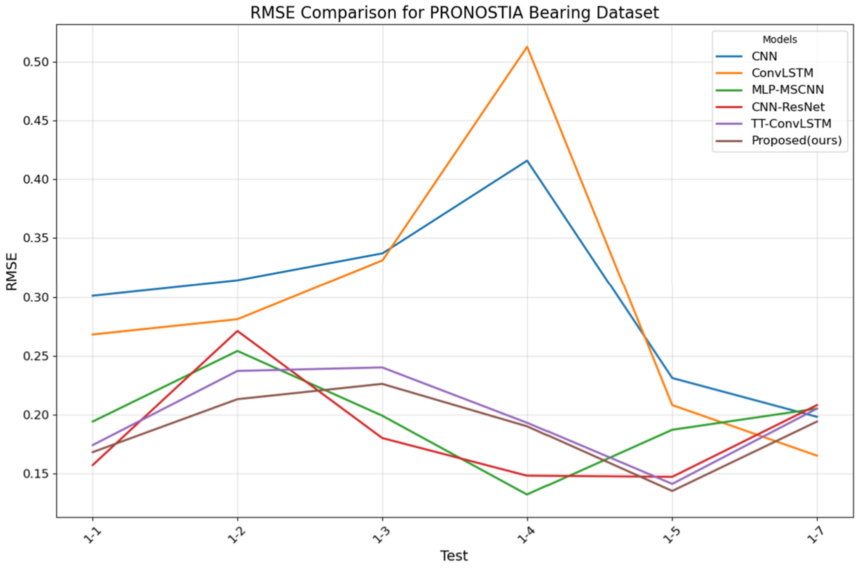

Figure 9 RMSE Comparison for PRONOSTIA Bearing Dataset

hanced with the use of the dynamic attention mechanism, which lets the model focus on the most relevant features in different settings to boost RUL prediction accuracy and robustness.

Similarly, we present the performance comparison of various models for RUL estimation using the PRONOSTIA bearing dataset. Table 4 summarizes the results of different models on the PRONOSTIA dataset.

Figures 9 and 10 show the results on the PRONOSTIA bearing tested dataset demonstrating that our proposed model consistently produces lower values of RMSE and MAE than other models. For example, the proposed model exhibits the RMSE of 0.168, and the MAE of 0.154 in Test 1-1, outperforming the RMSE of 0.301 and the MAE of 0.265 obtained by CNN. It shows that the model generalizes well across datasets and operational conditions. Test 1-4 in that tests of increased variability and noise also demonstrate how effective the proposed model is in the PRONOSTIA dataset. More importantly, we show that the proposed model significantly outperforms ConvLSTM and CNN-ResNet in RMSE (0.190) and MAE (0.139) while being robust to challenging data.

5.3 Discussion

The model was trained on the XJTU-SY dataset in nearly 2 hours with NVIDIA GTX 1660Ti GPU, and on the PRONOSTIA dataset in nearly 1.5 hours, with an average inference time of 20 ms /sequence. The proposed model, under its optimal configuration, consists of approximately 5.9 M parameters. Smaller kernel and pooling sizes, while increasing parameter count significantly (up to 17.9 M), did not yield proportional performance gains, highlighting the trade-offs in model design. The use of CWT and a dynamic attention mechanism adds computational overhead over more simple CNN models. However, this overhead is acceptable as the overhead is mitigated with efficient GPU parallelization and batch processing, and allows scaling to larger datasets. The model yields a large reduction in RMSE as compared to the baseline models despite a modest addition of the computational cost of up to 15% less compared to the baseline models. These results

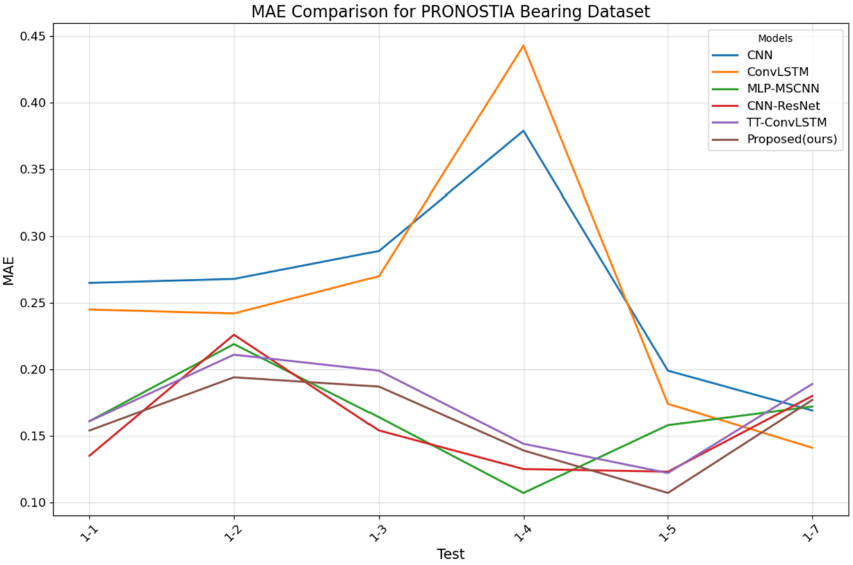

Figure 10 MAE Comparison for PRONOSTIA Bearing Dataset

show a good compromise between accuracy and efficiency; the model is shown to be applicable for predictive maintenance and to have the potential for further optimization in constrained operating environments.

The experimental results indicate that the proposed model, which integrates Continuous Wavelet Transform (CWT) for feature extraction and a dynamic attention mechanism within a Channel-Temporal Mixed MLP (CT-MLP) framework, provides significant improvements in Remaining Useful Life (RUL) prediction accuracy and generalization capability. Specifically, the dynamic attention mechanism enhances the model's ability to focus on the most relevant features, effectively handling the nonstationary nature of bearing degradation data. The incorporation of a multi-head attention mechanism further strengthens the model by capturing complex temporal dependencies, thereby improving the robustness and reliability of the predictions. The CT-MLP layer also contributes to performance gains by comprehensively understanding the intricate dependencies within the time series data, which is reflected in the lower RMSE and MAE values across different test scenarios on both the XJTU-SY and PRONOSTIA datasets.

In most scenarios, the proposed algorithm shows merits, especially when the dataset has a constant degradation pattern and sufficient training data (Tests 1-1 and 2-1 of the XJTU-SY dataset). Two primary challenges might account for the proposed algorithm's lack of superiority in many cases, as observed in Table 3 and Fig. 7: 1) High Dataset Variability: Higher complexity of the model can result in overfitting such subset when degradation patterns present themselves irregular (e.g., Test 3-3) or noise is highly present (e.g., Test 4-1). 2) Limited Training Samples: In contrast, the dynamic attention mechanism can fail to capture dependent features in subsets with fewer training samples, hence causing a performance reduction. Nevertheless, the overall performance of the model remains stable and can reliably process nonstationary and highly complex datasets. These results demonstrate the effectiveness of the proposed approach as well as its limitations in the aspects that can be further optimized: Adaptability to highly variable or sparse datasets.

Although this study mainly addresses data imbalance and improves prediction performance by the use of dynamic attention mechanisms and CWT-based feature extraction, it is important to test the robustness of the model concerning varied noise cases. By the richness of the feature extraction and attention mechanisms, the proposed methodology can inherently offer some resilience to nonstationary and noisy signals. In future work, we will carry out targeted experiments on different noise settings to evaluate and strengthen the robustness of the model. We will introduce a controlled noise in benchmark datasets and assess the model's capacity to maintain its accuracy and reliability. Additionally, we hope that by addressing this aspect, we can further demonstrate the practical usefulness of the proposed approach in industrial practice.

For all that was accomplished, there are still problems to be fixed. The dynamic attention mechanism and CT-MLP layers are computationally expensive, imposing stringent requirements on computational resources, and may not suit the real-time needs of such resource-constrained environments. The model can demonstrate robustness on each of these datasets, and it would be interesting to run experiments to visualize performance on more diverse datasets to ensure broad generalization. The second limitation is that adding more attention heads and more MLP layers beyond an optimal point will not make the accuracy dramatically increase, rather it would induce a waste of the computational costs.

From these challenges, future work could seek to optimize the computational efficiency of the model. For example, a lightweight dynamic attention mechanism and CT-MLP layers could be developed which reduce computational overheads without losing on predictive performance. Alternatively, transfer learning techniques for model generalization across a range of datasets and operational conditions is an area for additional improvement. Further, more sophisticated regularization techniques can be incorporated to help counter the problem of overfitting. Additionally, future studies can investigate the integration of additional types of sensor data into the predictive maintenance model construction to provide a more robust and accurate predictive maintenance model.

6 Conclusions

In this work, we explore an approach to bearing remaining useful life (RUL) prediction based on Continuous Wavelet Transform (CWT) for feature extraction integrated with a dynamic attention mechanism in a multi-head attention formation. The CT-MLP layer part is also proposed in the model to exploit the temporal and channel dependencies within time series data. The model was evaluated against two public datasets, XJTU-SY and PRONOSTIA, and was shown to reduce root mean square error (RMSE) and maximum absolute error (MAE) performance compared to other state-of-the-art models. CWT's integration in the model gives the model the ability to manage nonstationary signals and convert them to expressive time-frequency domain representations. The CT-MLP layer provides a comprehensive understanding of the data, whereas the dynamic attention mechanism improves its performance, by paying attention to the most important features. However, the model's complexity increases the computational costs, which may hinder its use in resource-constrained environments. Finally, its performance concerning some of the operational conditions did not outperform all baseline methods in all situations. The overall result of the proposed approach is to improve the reliability and efficiency of RUL predictions, facilitating more effective predictive maintenance strategies in industrial applications. Finally, future work would look into integrating more data sources and additional refinement of the attention mechanisms in an attempt to improve the performance of the model even more.

Acknowledgements
Not applicable.

Author contributions
Zhongtian Jin contributed to the conceptualization, methodology design, experimental study, and writing of the manuscript. Chong Chen and Aris Syntetos provided validation and critical revisions. Ying Liu supervised the project, contributed to methodology design and writing, and provided project resources. All authors read and approved the final manuscript.

Funding
Not applicable.

Data availability
The datasets used in this study are publicly available and can be accessed from the XJTU-SY rolling element bearing dataset and the PRONOSTIA bearing dataset.

Code availability
Not applicable.

Declarations

Competing interests
Prof. Ying Liu is an editorial board member for Autonomous Intelligent Systems and was not involved in the editorial review, or the decision to publish, this article. All authors declare that there are no other competing interests.

Author details
[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]PARC Institute of Manufacturing Logistics and Inventory, Cardiff Business School, Cardiff University, Cardiff CF10 3EU, UK.

Received: 30 July 2024 Revised: 1 December 2024
Accepted: 19 December 2024 Published online: 10 January 2025

References

1. Y. Chen, et al., A novel deep learning method based on attention mechanism for bearing remaining useful life prediction. Appl. Soft Comput. **86**, 105919 (2020). https://doi.org/10.1016/j.asoc.2019.105919. Available at
2. P. Ji, et al., Deep learning prediction of Amplitude Death. Auton. Intell. Syst. **2**(1), 26–36 (2022). https://doi.org/10.1007/s43684-022-00044-0
3. C. Chen, D. Wu, Y. Liu, Recent advances of AI for Engineering Service and maintenance. Auton. Intell. Syst. **2**(1), 19–21 (2022). https://doi.org/10.1007/s43684-022-00038-y
4. H. Fu, Y. Liu, A deep learning-based approach for electrical equipment remaining useful life prediction. Auton. Intell. Syst. **2**(1), 16–27 (2022). https://doi.org/10.1007/s43684-022-00034-2
5. M. Zhao, B. Tang, Q. Tan, Bearing remaining useful life estimation based on time–frequency representation and supervised dimensionality reduction. Measurement **86**, 41–55 (2016). https://doi.org/10.1016/j.measurement.2015.11.047
6. B. Zhao, Q. Yuan, A novel deep learning scheme for multi-condition remaining useful life prediction of rolling element bearings. J. Manuf. Syst. **61**, 450–460 (2021). https://doi.org/10.1016/j.jmsy.2021.10.004
7. H. Wang, et al., Remaining useful life prediction of bearings based on convolution attention mechanism and temporal convolution network. IEEE Access **11**, 24407–24419 (2023). https://doi.org/10.1109/access.2023.3255891
8. A. Tayade, et al., Remaining useful life (RUL) prediction of bearing by using regression model and principal component analysis (PCA) technique. Vibroeng. Proc. **23**, 30–36 (2019). https://doi.org/10.21595/vp.2019.20617
9. T. Li, et al., WaveletKernelNet: an interpretable deep neural network for industrial intelligent diagnosis. IEEE Trans. Syst. Man Cybern. Syst. **52**(4), 2302–2312 (2022). https://doi.org/10.1109/tsmc.2020.3048950
10. S.R. Shah, et al., A sequence-to-sequence approach for remaining useful lifetime estimation using attention-augmented bidirectional LSTM. Intell. Syst. Appl. **10–11**, 200049 (2021). https://doi.org/10.1016/j.iswa.2021.200049
11. Y. Liu, et al., Degradation-trend-aware deep neural network with attention mechanism for bearing remaining useful life prediction. IEEE Trans. Artif. Intell. **5**(6), 2997–3011 (2024). https://doi.org/10.1109/tai.2023.3333767

12. M. Zhao, B. Tang, Q. Tan, Bearing remaining useful life estimation based on time–frequency representation and supervised dimensionality reduction. Measurement **86**, 41–55 (2016). https://doi.org/10.1016/j.measurement.2015.11.047
13. L. Ren, et al., Bearing remaining useful life prediction based on Deep Autoencoder and Deep Neural Networks. J. Manuf. Syst. **48**, 71–77 (2018). https://doi.org/10.1016/j.jmsy.2018.04.008
14. P.V. Kamat, R. Sugandhi, S. Kumar, Deep learning-based anomaly-onset aware remaining useful life estimation of bearings. PeerJ Comput. Sci. **7**, e795 (2021). https://doi.org/10.7717/peerj-cs.795
15. H. Liu, et al., Remaining useful life prediction using a novel feature-attention-based end-to-end approach. IEEE Trans. Ind. Inform. **17**(2), 1197–1207 (2021). https://doi.org/10.1109/tii.2020.2983760
16. Y. Wu, et al., Remaining useful life estimation of Engineered Systems using Vanilla LSTM neural networks. Neurocomputing **275**, 167–179 (2018). https://doi.org/10.1016/j.neucom.2017.05.063
17. X. Li, W. Zhang, Q. Ding, Deep learning-based remaining useful life estimation of bearings using multi-scale feature extraction. Reliab. Eng. Syst. Saf. **182**, 208–218 (2019). https://doi.org/10.1016/j.ress.2018.11.011. Available at
18. C. Jin, X. Chen, An end-to-end framework combining time–frequency expert knowledge and modified transformer networks for vibration signal classification. Expert Syst. Appl. **171**, 114570 (2021). https://doi.org/10.1016/j.eswa.2021.114570
19. F. Yang, et al., Waveform: graph enhanced wavelet learning for long sequence forecasting of multivariate time series. Proc. AAAI Conf. Artif. Intell. **37**(9), 10754–10761 (2023). https://doi.org/10.1609/aaai.v37i9.26276
20. S.Z. Hejazi, M. Packianather, Y. Liu, A novel customised load adaptive framework for induction motor fault classification utilising MFPT bearing dataset. Machines **12**(1), 44 (2024). https://doi.org/10.3390/machines12010044
21. M.F. Siddique, et al., A hybrid deep learning approach: integrating short-time Fourier transform and continuous wavelet transform for improved pipeline leak detection. Sensors **23**(19), 8079 (2023). https://doi.org/10.3390/s23198079
22. S.R. Shah, et al., A sequence-to-sequence approach for remaining useful lifetime estimation using attention-augmented bidirectional LSTM. Intell. Syst. Appl. **10–11**, 200049 (2021). https://doi.org/10.1016/j.iswa.2021.200049
23. C. Chen, et al., An integrated deep learning-based approach for automobile maintenance prediction with GIS Data. Reliab. Eng. Syst. Saf. **216**, 107919 (2021). https://doi.org/10.1016/j.ress.2021.107919
24. S.G. Niazi, et al., Multi-scale time series analysis using TT-CONVLSTM technique for bearing remaining useful life prediction. Mech. Syst. Signal Process. **206**, 110888 (2024). https://doi.org/10.1016/j.ymssp.2023.110888
25. D. Wang, K. Liu, X. Zhang, A generic indirect deep learning approach for multisensor degradation modeling. IEEE Trans. Autom. Sci. Eng. **19**(3), 1924–1940 (2022). https://doi.org/10.1109/tase.2021.3072363
26. J. Liu, et al., Fault prediction of bearings based on LSTM and statistical process analysis. Reliab. Eng. Syst. Saf. **214**, 107646 (2021). https://doi.org/10.1016/j.ress.2021.107646
27. J. Zhang, et al., Long short-term memory for machine remaining life prediction. J. Manuf. Syst. **48**, 78–86 (2018). https://doi.org/10.1016/j.jmsy.2018.05.011
28. J. Zhu, N. Chen, W. Peng, Estimation of bearing remaining useful life based on multiscale convolutional Neural Network. IEEE Trans. Ind. Electron. **66**(4), 3208–3216 (2019). https://doi.org/10.1109/tie.2018.2844856
29. Z. Li, et al., MTS-mixers: Multivariate Time Series forecasting via factorized temporal and channel mixing (2023). https://arxiv.org/abs/2302.04501 (Accessed: 28 July 2024)
30. Z. Liu, et al., A multi-head neural network with unsymmetrical constraints for remaining useful life prediction. Adv. Eng. Inform. **50**, 101396 (2021). https://doi.org/10.1016/j.aei.2021.101396
31. X. Li, et al., Data alignments in machinery remaining useful life prediction using deep adversarial neural networks. Knowl.-Based Syst. **197**, 105843 (2020). https://doi.org/10.1016/j.knosys.2020.105843
32. C.-G. Huang, et al., A novel deep convolutional neural network-bootstrap integrated method for RUL prediction of rolling bearing. J. Manuf. Syst. **61**, 757–772 (2021). https://doi.org/10.1016/j.jmsy.2021.03.012
33. Y. Zou, et al., A method for predicting the remaining useful life of rolling bearings under different working conditions based on multi-domain adversarial networks. Measurement **188**, 110393 (2022). https://doi.org/10.1016/j.measurement.2021.110393
34. Y. Mo, et al., Remaining useful life estimation via transformer encoder enhanced by a gated convolutional unit. J. Intell. Manuf. **32**(7), 1997–2006 (2021). https://doi.org/10.1007/s10845-021-01750-x
35. F. Zeng, et al., A deep attention residual neural network-based remaining useful life prediction of machinery. Measurement **181**, 109642 (2021). https://doi.org/10.1016/j.measurement.2021.109642
36. X. Yu, Z. Yu, S. Ramalingam, Learning strict identity mappings in deep residual networks, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018). https://doi.org/10.1109/cvpr.2018.00466. [Preprint]
37. X. Huang, et al., Frequency Hoyer attention based convolutional neural network for remaining useful life prediction of machinery. Meas. Sci. Technol. **32**(12), 125108 (2021). https://doi.org/10.1088/1361-6501/ac22f0
38. M. Jiang, et al., MLGN: multi-scale local-global feature learning network for long-term series forecasting. Mach. Learn.: Sci. Technol. **4**(4), 045059 (2023). https://doi.org/10.1088/2632-2153/ad1436
39. G. Ioannides, A. Chadha, A. Elkins, Gaussian adaptive attention is all you need: Robust contextual representations across multiple modalities (2024). https://arxiv.org/abs/2401.11143 (Accessed: 13 June 2024)
40. L. Zhu, et al., Biformer: vision transformer with bi-level routing attention, in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023). https://doi.org/10.1109/cvpr52729.2023.00995. [Preprint]
41. L. Shen, et al., Improving the robustness of transformer-based large language models with dynamic attention, in *Proceedings 2024 Network and Distributed System Security Symposium* (2024). https://doi.org/10.14722/ndss.2024.24115. [Preprint]
42. Y. Zhang, et al., A free lunch from VIT: adaptive attention multi-scale fusion transformer for fine-grained visual recognition, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2022). https://doi.org/10.1109/icassp43922.2022.9747591. [Preprint]
43. S. Fu, et al., MCA-DTCN: a novel dual-task temporal convolutional network with multi-channel attention for first prediction time detection and remaining useful life prediction. Reliab. Eng. Syst. Saf. **241**, 109696 (2024). https://doi.org/10.1016/j.ress.2023.109696
44. B. Wang, Y. Lel, N. Li, et al., A hybrid proanostics approach for estimating remaining useful life of rolling element bearinas. IEEE Trans. Reliab. **69**(1), 401–412 (2020). https://doi.org/10.1109/TR.2018.2882682
45. P. Nectoux, R. Gouriveau, K. Medjaher, E. Ramasso, B. Morello, N. Zerhouni, C. Varnier, PRONOSTIA: an experimental platform for bearings accelerated life test, in *IEEE International Conference on Prognostics and Health Management*, Denver, CO, USA (2012)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.