Scientific
Research
Publishing

Remaining Useful Life Prediction of Rail Based on Improved Pulse Separable Convolution Enhanced Transformer Encoder

Zhongmei Wang[1], Min Li[1*], Jing He[2], Jianhua Liu[1], Lin Jia[1]

[1]College of Railway Transportation, Hunan University of Technology, Zhuzhou, Hunan
[2]College of Electrical and Information Engineering, Hunan University of Technology, Zhuzhou, Hunan
Email: wangzhongmei@hut.edu.cn, *m21085400058@stu.hut.edu.cn, hejing@263.com, juliu0615@163.com, jialin@hut.edu.cn

Abstract

In order to prevent possible casualties and economic loss, it is critical to accurate prediction of the Remaining Useful Life (RUL) in rail prognostics health management. However, the traditional neural networks is difficult to capture the long-term dependency relationship of the time series in the modeling of the long time series of rail damage, due to the coupling relationship of multi-channel data from multiple sensors. Here, in this paper, a novel RUL prediction model with an enhanced pulse separable convolution is used to solve this issue. Firstly, a coding module based on the improved pulse separable convolutional network is established to effectively model the relationship between the data. To enhance the network, an alternate gradient back propagation method is implemented. And an efficient channel attention (ECA) mechanism is developed for better emphasizing the useful pulse characteristics. Secondly, an optimized Transformer encoder was designed to serve as the backbone of the model. It has the ability to efficiently understand relationship between the data itself and each other at each time step of long time series with a full life cycle. More importantly, the Transformer encoder is improved by integrating pulse maximum pooling to retain more pulse timing characteristics. Finally, based on the characteristics of the front layer, the final predicted RUL value was provided and served as the end-to-end solution. The empirical findings validate the efficacy of the suggested approach in forecasting the rail RUL, surpassing various existing data-driven prognostication techniques. Meanwhile, the proposed method also shows good generalization performance on PHM2012 bearing data set.

Keywords

1. Introduction

With increasing railway speed and the development of railway heavy haul transportation [1] [2], more and more failures relating to the rail have occurred. Accurately predicting the remaining life of a rail and formulating a rail flaw detection period for extending the remaining service life of the rail are significance in ensuring the safe operation of a railway line. Because rail maintenance intervals and service life have a direct impact on the operational status of the line.

Currently, the available methods for predicting Remaining Useful Life (RUL) can be classified into two distinct categories: one rely on model-based forecasting and the other utilize data-driven forecasting techniques [3]. The methodology with model employs failure mechanisms or damage laws to simulate the deterioration of the machine. Subsequently, statistical estimation techniques, including linear least squares, maximum likelihood estimation, and sequential Monte Carlo [4], are utilized to determine the model parameters and forecast the RUL in the methodology with model. However, the irregularity of damage to heavy rail tracks poses a challenge for developing a precise failure model in actuality. Consequently, the establishment of precise mathematical statistics and physical degradation models in practical applications is very difficult. For data-centric methodologies, it does not necessitate familiarity with overt machine malfunction mechanisms and have the ability to deduce concealed causal connections within the data.

In recent years, deep learning has gained a lot of interesting in data-driven RUL prediction [5] [6] [7] [8]. Compared with traditional machine learning technology, deep learning possesses a superior capacity for representation learning and is capable of autonomously acquiring multi-level representations from raw data [9] [10]. Hence, a prediction model is constructed directly from the raw sensor data via the utilization of deep learning technology, thereby eliminating the intricate procedure of manual feature extraction. And the predominant approach for time series modeling is applied by recurrent neural network (RNN) methodology. Moreover, the long short term memory (LSTM) network has obtained the significant attention due to its ability to address the issues of gradient explosion and vanish the gradient in RNN [11]. Zhao et al. introduced a method for predicting RUL, which is a combination of bidirectional long short term memory (BILSTM) and attention mechanism [12]. The effectiveness of this method was evaluated on two publicly available datasets. However, LSTM exhibits certain constraints, because LSTM is a serial structure in essence, which can not realize parallel computing. It has the low efficiency and hard capturing the long-term dependence in time series. Conversely, the transformer model [13] could not only improve the computational efficiency through parallel processing, but also effectively deal with the long-term dependence of time-series data over time. Remarkable outcomes have been attained across various domains, including but not limited to natural language processing and computer vision.

Recently, Transformer model was used to predict RUL. Mo et al. [14] adopted the Transformer encoder as the fundamental architecture of their model, and applied 1 gate convolution uni combined with the information of the local context with each time to achieve RUL prediction. Zhang et al. [15] achieved quite good prediction results by using a completely self-attention encoder decoder structure and establishing a Transformer structure with the characteristics of the sensor characteristics and time step length as the input. Chen [16] introduced a RUL prediction method for the electro-mechanical actuators via a multimodal Transformer. However, the amount of Transformer model parameters is large, and so is the calculation consumption. If the training sample is insufficient, overfitting can be easily caused in Transformer model. In addition, the multi-channel vibration data sets of the rail RUL prediction are composed of signals obtained by different sensors and different channels, which are of two difficulties: One is that the data from the many sensors include information has been degraded to differing degrees, and the other is the different sensors signals are of correlation.

The spiking neural network (SNN), which is the third generation of artificial neural network, has been employed as a means to address the aforementioned issues. And it is closer to the working principle of biological neural systems [17]. With its working mechanism based on dynamic sparse pulse discharge, it is expected to overcome the existing shortcomings of the artificial neural network and realize the prediction function with strong generalization ability. At the same time, because the SNN itself processes the information through the dynamic pulse discharge process, it is easier to capture the spatio-temporal correlation information in the time series.

Some researchers have successfully improved the traditional SNN and have applied it to time series forecasting tasks. Hong et al. [18] introduced a multi-layer pulse neural network based on the improved pulse neuron model, and learned it through the pulse time error back-propagation algorithm to achieve short-term power load forecasting. However, the feature extraction ability of traditional SNN is limited by its structure, and its prediction accuracy of RUL time series is not high enough, which cannot be comparable with that of RNN and Transformer network. Therefore, how to improve the traditional SNN for effectively predict RUL is a problem to be solved. Inspired by reference [14], a model around a Transformer encoder is build to minimize model parameters and boost operational effectiveness. More importantly, the suggested network makes use of a brand-new, effective channel attention technique to further enhance network prediction performance.

This paper's principal contributions are outlined as follows:

1) An improved pulse-separable convolutional network is proposed, which leverages 3×3 channel-by-channel convolution and 1×1 point-by-point convolution to effectively capture the inter-channel dependencies in multi-channel data. Additionally, an alternative gradient back propagation algorithm is employed to optimize the network performance. Furthermore, an ECA attention

mechanism is incorporated to re-calibrate the pulse feature map, thereby enhancing the discriminability of the information in the prediction network.

2) A residual life prediction module with improved Transformer encoder as the core is proposed. Two encoder layers are used for stacking training, and the parallel computing power of multi-head self-attention mechanism is fully utilized to dig the dependency among input characteristics and rail degradation degree and themselves. Combined with the information processing mechanism of SNN, the two-layer encoder is improved by blending pulse maximum pooling, which preserves more degradation time sequence features and improves the accuracy of rail RUL prediction.

The other parts of this article are organized as follows. Section 2 describes the main issues. Then section 3 introduces the proposed method, and section 4 gives the experiment. Finally, conclusions are drawn in section 5.

2. Research Problem

The objective of this paper is to develop a prognostic model for estimating the RUL of a system using vibration monitoring data obtained from multi-channel sensors. Specifically, M sets of rail vibration data to train an offline prediction model, encompassing the entire degradation process from normal state to slight fault and eventually to severe fault and then to severe fault were used to train the prediction model offline, and then the trained model was used to forecast the RUL of a new rail.

In this paper, the multivariable time series prediction problem is defined as a sequence-to-sequence problem. The overall architecture for time series prediction is described before the network architecture is specified. An input sequence $(x_1, x_2, \cdots, x_T)$ was given to the time series signals, this is to predict the output $Y = (y_1, y_2, \cdots, y_T)$, corresponding to each respective time. The primary objective of serial modeling network z, as shown in Equation (1), is to build up a mathematical relationship between the monitoring data and the rail degradation process:

$$(y_1, y_2, \cdots, y_T) = f(x_1, x_2, \cdots, x_T) \tag{1}$$

3. The Proposed Method

Figure 1 depicts the comprehensive framework diagram of the network as presented in the paper, which is aimed at attaining a precise forecast of the rail RUL. The network is primarily comprised of two distinct components: improved pulse separable convolution coding module B and residual life prediction module C. The improved pulse separable convolution module converts time series signals into pulse signals, thereby capturing the inter-sensordata relationship and reduce the information loss in the coding process. The alternative gradient back propagation technique was utilized to improve the network, and ECA attention mechanism was added to re-calibrate the pulse feature map and highlight the useful pulse features. Then, based on the effective pulse characteristics,
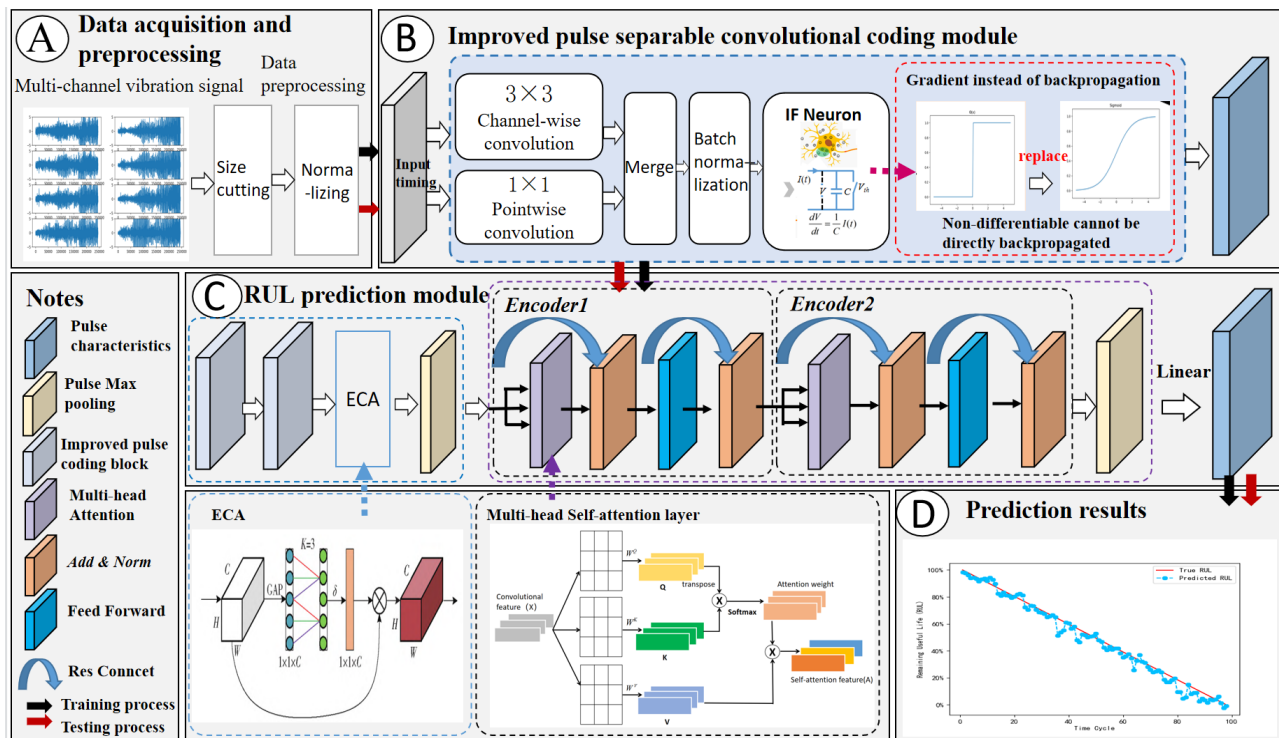
Figure 1. Overall framework of improved pulse separable convolution enhanced Transformer Encoder.

the residual life prediction module adopted multi-head self-attention mode to comprehensively control the dependence of each time step of the fault characteristics in the long timespan of the whole life cycle. Finally, the linear layer completed the RUL prediction.

3.1. Improved Pulse Separable Convolution Coding Module

The multi-channel data used in rail RUL prediction is collected by multi-sensors at different positions. On the one hand, the degradation information contained in vibration data captured by various sensors varies. On the other hand, these data are highly correlated (coupled), so it is difficult for traditional pulse coding to effectively model the pulse representation of the relationship between these data [19]. To solve the above problems, as shown in **Figure 2**, this paper proposes a multi-layer pulse separable convolutional coding module comprised of a depth separable convolutional layer and a pulse neuron coding layer. The data from the input multi-channel sensors was simply fed into the prediction network without any human feature extraction being performed beforehand. Secondly, depth-separable convolution was used so that the number of parameters maybe decreased while simultaneously realizing the separation of channels and regions. After that, Merge Channels and BatchNorm2d were processed so that the vibration data would not be too large before being sent into the improved pulse neuron layer, improving the stability of the network. Finally, the improved pulse neuron layer coding was constructed to get the pulse feature output.
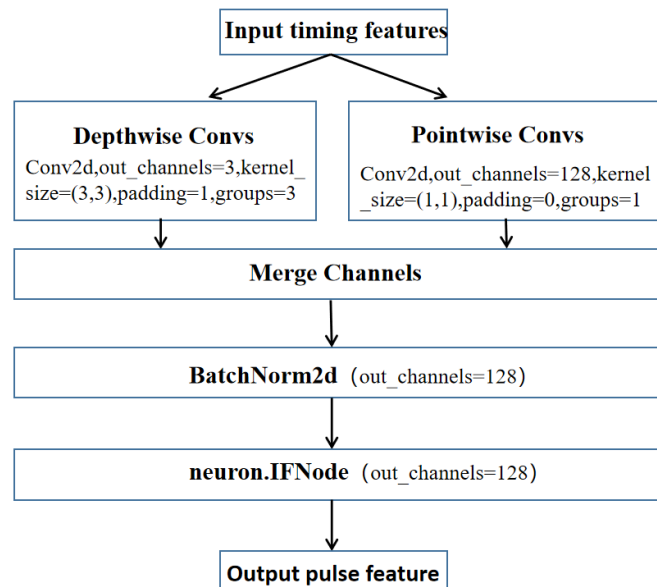
Figure 2. Depth-separable convolution structure.

3.1.1. Depth Separable Convolution

Figure 2 depicts the structure of depth separable convolution [20], which is made up primarily of channel-by-channel convolution (Depthwise Convs) and point-by-point convolution (Pointwise Convs). The number of convolution nuclei in each channel of classical convolution is proportional to the number of input channels; however, a channel of per-channel convolution has only one convolution kernel, which means that per-channel convolution may significantly decrease the number of parameters and the amount of processing. The point-by-point convolution is a classic kind of convolution that utilizes a convolution kernel size of 1 by 1. Convolution on a point-by-point basis allows for the fusion of features that have been taken from matching points within each channel. This allows one to avoid critical characteristics that are only present in a single channel. In addition, both an increase in the number of features and a decrease in the dimensions may be accomplished by adjusting the number of channels appropriately.

3.1.2. Coding layer of pulse neuron

Pulsed neuron is the basic unit that makes up pulsed neural network (SNN) [17]. Different from convolutional neural network, pulsed neural network adopts a processing mechanism more in line with human brain, that is, pulse sequence as signal transmission. In current research, input information is usually encoded into pulse train by Poisson coding, delay coding and other methods. These coding methods [21] [22] have certain randomness, which may cause information loss and affect the effect of subsequent rail RUL prediction. In order to solve this problem, an improved pulse depth-separable convolution coding scheme is proposed in this section.

The improved pulse depth-separable convolution coding method combines

the depth-separable convolution with the improved pulse neurons, and the multi-channel sensor data was fed directly into the network model. The rail RUL time sequence data was convolved on the convolution kernel of the separable convolutional neural network, and the voltage was accumulated in the membrane potential of the postsynaptic neuron. If the voltage of the membrane potential exceeded the set threshold voltage, the pulse neuron would send a pulse and then return to the resting potential. If the threshold voltage was not reached, the membrane potential would accumulate voltage and wait for the input of the next time window.

In order to realize the coding process, the IF neuron model was adopted [23]. IF neuron model can be regarded as an ideal integrator. Its membrane potential voltage would not leak over time when it was not activated, as shown in Equation (2). The membrane potential voltage at time t was shown in Equation (3), where V (t) is the membrane potential voltage at time t, and I (t) is the input at the current time, Vth is the threshold voltage, and set to Vth = 1 in this paper. When the membrane potential voltage reached the threshold voltage, the pulse would be activated; otherwise, the pulse would be sent step by step, as shown in Equation (4).

As can be seen from Equation (4), pulsed neurons themselves are non-differentiable, that is, non-derivable. As a result, it is not possible to update model parameters through direct back propagation in the training process. Therefore, this section adopts alternative gradient descent algorithm to complete the updating process of back propagation parameters and improves the traditional IF neurons. Through the selection of a suitable function to substitute the impulse function during the back propagation process, the network still showed neuronal impulse characteristics in the forward propagation, while the continuous differentiable function was used to replace the impulse function in the back propagation.

In the proposed coding method, derivatives of Relu function, Sigmoid function and Piecewise LeakyRelu function were introduced to replace the derivative of θ during back propagation. The derivative expressions are described as equations (5) to (7):

$$-\frac{\mathrm{d}V(t)}{\mathrm{d}t} = \frac{1}{C}I(t) \tag{2}$$

$$V(t) = V(t-1) + I(t) \tag{3}$$

$$\theta(x) = \begin{cases} 1, & V(t) \geq V_{th} \\ 0, & V(t) < V_{th} \end{cases} \tag{4}$$

$$g_1^1(x) = \begin{cases} x, & x > 0 \\ 0, & x \leq 0 \end{cases} \tag{5}$$

$$g_2^1(x) = \frac{a\mathrm{e}^{-ax}}{\left(1+\mathrm{e}^{-ax}\right)^2} \tag{6}$$

$$g_3^1(x) = \begin{cases} x, & x > 0 \\ ax, & x \le 0 \end{cases} \tag{7}$$

where, coefficient a is the smoothness of the activation function. This part aims to evaluate and contrast the impact of various alternative activation functions on the predictive efficacy of the network's RUL in the experiment outlined in Section 4.3.3.

3.2. Remaining Life Prediction Module

3.2.1. Multilayer Pulse Separable Convolution Algorithm

Based on Section3.1, the multilayer pulse separable convolution algorithm includes a two-layer pulse separable convolution coding module, a high efficiency channel attention (ECA-Net) module and a pulse maximum pooling step. As shown in **Figure 3**, after the time sequence data of the input multi-sensor went through two-layer pulse separable convolutional coding module (Spilking Separable Convs_1, 2), the time series signals were converted into pulse signals to capture the connection between data from various sensors, and to minimize the information loss during the process of coding. In order to highlight the useful pulse features, ECA attention mechanism was added to re-calibrate the pulse feature map. Pulse maximum pooling was adopted to retain more pulse features and facilitate subsequent RUL prediction.

The majority attention mechanisms strive to enhance their performance by developing intricate attention modules, which unavoidably leads to an escalation in the model's complexity. The ECA-Net [24] mainly improves the SE-Net module [25] by introducing a local cross-channel interaction strategy (ECA module) that does not involve any dimension reduction. Additionally, it incorporates a method for the adaptive selection of one-dimensional convolution kernel size.
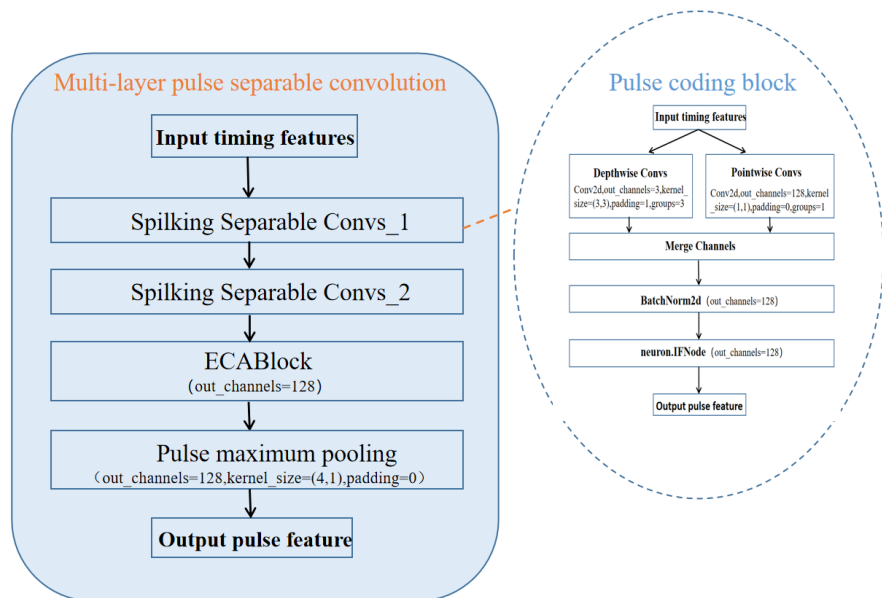

Figure 3. Flow of multilayer pulse separable convolution algorithm.

This particular module has the ability to achieve noteworthy improvements in performance despite the addition of only a limited number of parameters. **Figure 4** illustrates the schematic diagram of ECA-Net.

3.2.2. Improving Transformer Encoder Algorithm

When simulating lengthy time periods of rail deterioration, standard neural networks like RNN and SNN have trouble capturing the long-term dependency, which are easy to appear gradient disappearance or explosion, and sensitive to the length of input sequence. An enhanced Transformer encoder method is suggested to take advantage of the original multi-head self-attention mechanism in order to address this issue. The relationship between input timing sequence characteristics and rail degradation degree itself and among each other is better excavated. The detailed formulation of the multi-head self-attention mechanism in the paper is shown in **Figure 7** and Equation (11) section. The existing Transformer encoder is improved by integrating pulse maximum pooling operation to retain more effective pulse timing sequence features and to enhance the precision of rail RUL forecasting by combining the advantages of pulse separable convolution network in Section 3.1.

Figure 5 illustrates the RUL prediction mechanism that is used by the enhanced Transformer encoder. On the basis of the pulse features provided by multi-layer pulse separable convolution, after two layers of Transformer encoder structure, pulse maximum pooling and linear full connection layer successively,
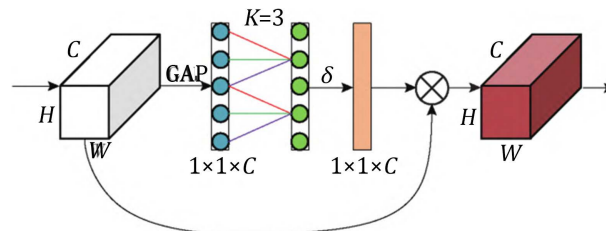

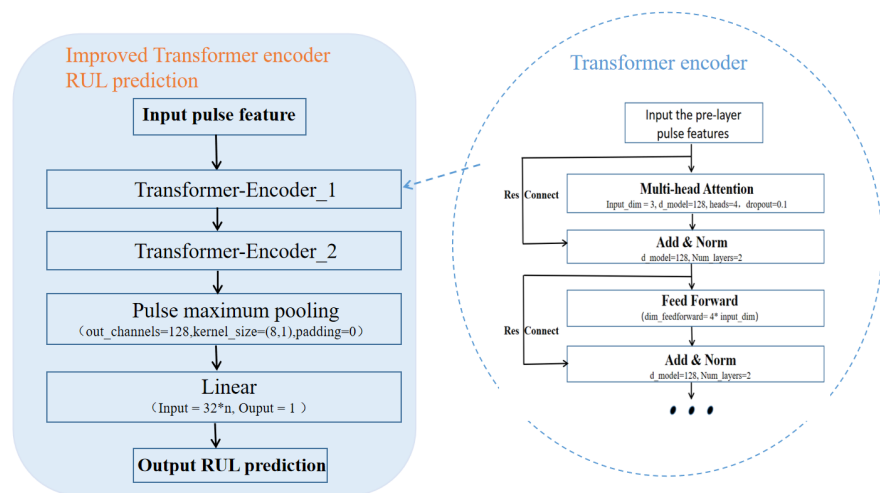**Figure 4.** Architecture schematic diagram of ECA-Net.

Figure 5. Improved Transformer encoder RUL prediction flow.

the final residual life prediction value RUL will be output. The multiple encoder layers used have a similar structure, with each encoder layer stacked with multiple sub encoder layers of the same structure. As shown in **Figure 5**, each sub-coder layer includes a Multi-head Attention layer and a Feed Forward layer, and both of them apply Res Connect and Add & Norm operation to mitigate the issue of vanishing gradient and expedite the convergence of the model.

Pooling processing can reduce the size of feature mapping after convolutional output and further reduce the overfitting degree of the model and the amount of training parameters of the network. Pooling processing methods include maximum pooling (selecting the largest neuron output in the pooling window) and mean pooling (carrying out two-dimensional mean pooling operation on every value in the pooling window). In the practical application, the maximum pooling operation is not only more in line with the information processing mechanism of SNN, but also retains more characteristic information in the timing data [26]. In order to retain more effective pulse timing features, this study adopted the maximum pooling operation (pooling kernel size of 8 × 1) for training.

An essential component of the Transformer network is the self-attention mechanism [13]. In the process of computing the self-attention mechanism, each feature will be given a relative weight in comparison to the other characteristics. This will make it possible to recognize the connection that exists between the various features, and it will also enable the extracted features to be determined in accordance with the degree of correlation that exists between the features. **Figure 6** illustrates the calculating process that is involved in the self-attention mechanism.

The convolution feature output matrix *X* of the previous layer was multiplied with the corresponding three weight matrices W^Q, W^K and W^V, to obtain the corresponding three vectors，query vector *Q*, key vector *K* and content vector *V*, respectively. The specific formula is Equation (9).

$$\begin{cases} Q = XW^q \\ K = XW^k \\ V = XW^v \end{cases} \tag{9}$$

The correlation matrix was obtained by calculating the dot product of *Q* and *K*, and the weight corresponding to each position was obtained after activated by Softmax function. Finally, the weight was superimposed to *V* to get the self-attention output *A*. The specific formula is Equation (10).

$$A(Q,K,V) = \text{Softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{D_K}}\right)V \tag{10}$$

where $\sqrt{D_K}$ is the square root of the key vector dimension, which is used as a scaling factor to alleviate the gradient disappearance problem.

Multi-head Attention layer adopts multi-head attention mechanism. This mechanism involves the calculation of multiple groups of *Q*, *K*, and *V*, which are subsequently combined to form the final output. In order to balance the possible

deviation of the same attention mechanism, and thus improve the model effect. **Figure 7** presents a diagrammatic representation of its structural components.

The corresponding calculation formula is

$$\begin{cases} A_{\text{MuliHead}}(Q,K,V) = [h_1,\cdots,h_H]W, \\ h_i = A(Q_i,K_i,V_i) \end{cases} \tag{11}$$

In Equation (11), W represents the multi-head attention weight matrix; h_i refers to the i-th self-attention output; H is for attention heads.

4. Experiment and Discussion

The proposed method was used to predict the whole life cycle vibration data set of railway rail and the effectiveness of network generalization is verified using IEEE PHM2012 bearing data set. The experiment was implemented in Python, the environment was built under PyTorch framework, the number of random
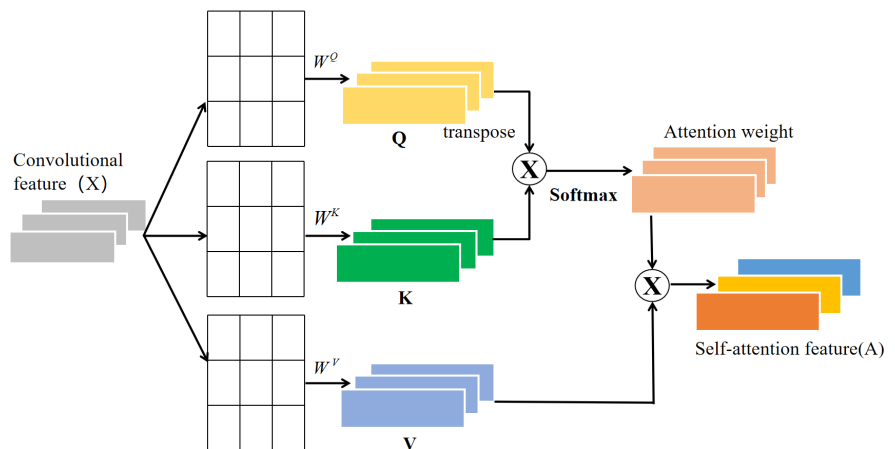

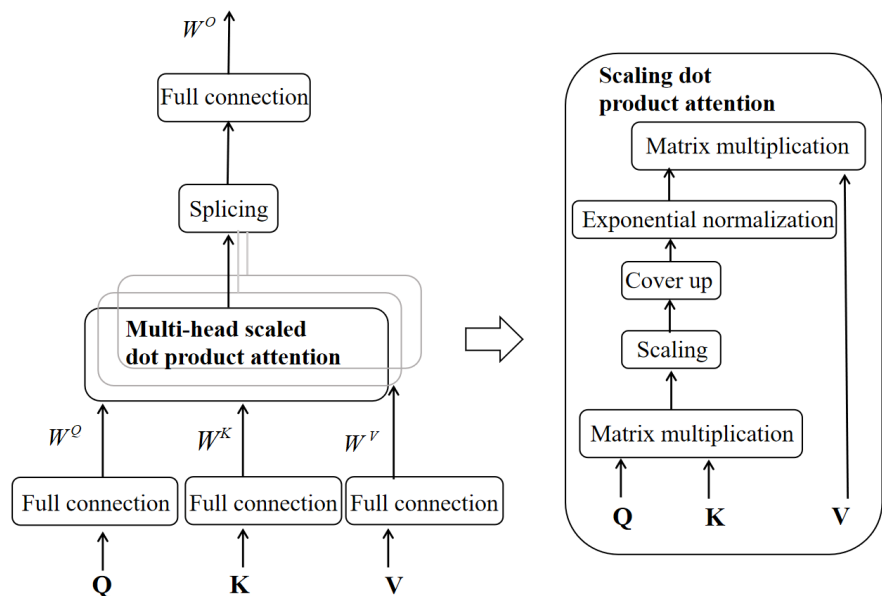**Figure 6.** Schematic diagram of self-attention mechanism.

Figure 7. Schematic diagram of multi-head self-attention mechanism.

seeds was fixed, and the weight of the network was randomly initialized. The configuration of the computer used is: 1) Processor (12th Gen Intel Core i9-12900H 2.90GHz); 2) Running memory (32G); 3) Graphics card (NVIDIA GeForce GTX 3070 Ti).

4.1. Performance Indicators

In order to conduct a thorough evaluation of the method's efficacy, three distinct performance metrics have been employed, namely the mean absolute error (MAE), the root mean square error (RMSE), and an enhanced scoring function (Score) which builds upon the original scoring function (iScore).

Where, Equations (12) and (13) define the MAE and RMSE, respectively. The actual RUL is denoted as RUL_t^{act}, the predicted RUL is denoted as RUL_t^{act}, and n represents the total number of samples. It can be seen that, MAE and RMSE deploy equal weight for each prediction. However, in actual applications, later predictions (ert < 0) are punished more seriously than early predictions (ert > 0). Thus, the scoring function (Score) is introduced as a supplement.

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}\left|RUL_t^{act} - RUL_t^{pre}\right| \tag{12}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(RUL_t^{act} - RUL_t^{pre}\right)^2} \tag{13}$$

In addition, it is worth noticing that the possibility of encountering failures in the early phases of the life cycle is comparatively minimal over the course of its operational lifespan. In another word, the precision of RUL estimation during the later phase assumes greater significance compared to the early stage. It means that it is better to give higher weight to the prediction of later stages. The prediction results will be better. So improve the scoring function in the 2012 Forecast and Health Management Data Challenge, and propose an improved scoring function (i_{Score}) to make a comprehensive assessment of the performance of the predicted model. Therefore, an improvement action was taken on the scoring function for the 2012 Prognostics and Health Management Conference Data Challenge, and the optimized scoring function (i_{Score}) was proposed, for the purpose of a comprehensive performance assessment of the prediction model.

$$iA_t = \begin{cases} \exp\left(-\ln(0.6)\times\left(RUL_t^{act} - RUL_t^{pre}/10\right)\right), & RUL_t^{act} - RUL_t^{pre} \leq 0 \\ \exp\left(\ln(0.6)\times\left(RUL_t^{act} - RUL_t^{pre}/40\right)\right), & RUL_t^{act} - RUL_t^{pre} \geq 0 \end{cases} \tag{14}$$

$$i_{Score} = \omega_1\sum_{t=1}^{m}iA_t + \omega_2\frac{1}{n-m}\sum_{m+1}^{n}iA_t \tag{15}$$

According to Equation (14), iA_t represents the weighted error that exists between the observed and anticipated values of PrediRUL at the time step *t*. The final improved scoring function is as shown in Equation (15), where *n* is the total number of sample, *m* is the percentage of early stages. ω_1 and ω_2 are the weights of early and late stages, respectively. In this paper, it is set that

$\omega_1 = 0.35$, $\omega_2 = 0.65$, $m = n/2$, which makes the prediction of late stages more reliable and meaningful than that of the early stages [27]. Values of ω_1 and ω_2 are determined according to the experiment entered by the user. The Score variable is constrained to a range between 0 and 1. A positive correlation exists between the value and the predicted performance, indicating that a higher value corresponds to superior performance.

4.2 Data Description and Application Details

4.2.1 Description of Data Set

Data on vibrations reflecting the full life cycle of railway rail from normal to damage and to final failure was collected. The vibration signals were obtained by sensors distributed at different positions in the train. Each sensor has three channels that can capture the vibration signal in three different directions: horizontally, longitudinally and vertically. Depending on the train speed and loading status, 4 different operating conditions were considered and the corresponding data was collected. At each operating condition, there were three different types of rail damage types: corrugation, corner fine crack, and shelling defect, as shown in Table 1. The vibration signal and actual appearance of different defects are shown in Figure 8 and Figure 9.

4.2.2. Design of Experimental Tag and Parameter Setting

The life span of the rail is used as the output tag. As shown in the Equation (16), the actual RUL of the rail is normalized to the range of 0% - 100%, where S is the total time step, and S^t is the normalized value of the actual RUL at the time step t.

$$S_t^n = \frac{S^t}{S} \times 100 \tag{16}$$

The Model structure parameter settings are shown in Table 2.

Table 1. Railway Rail data set.

Operating condition	Speed of operation	No load/load	Type of injury		
condition1	Low speed	No load	corrugation(A_1)	Fish scaly injury(B_1)	Stripping off blocks(C_1)
condition2	Low speed	load	corrugation(A_2)	Fish scaly injury(B_2)	Stripping off blocks(C_2)
condition3	High speed	No load	corrugation(A_3)	Fish scaly injury(B_3)	Stripping off blocks(C_3)
condition4	High speed	load	corrugation(A_4)	Fish scaly injury(B_4)	Stripping off blocks(C_4)

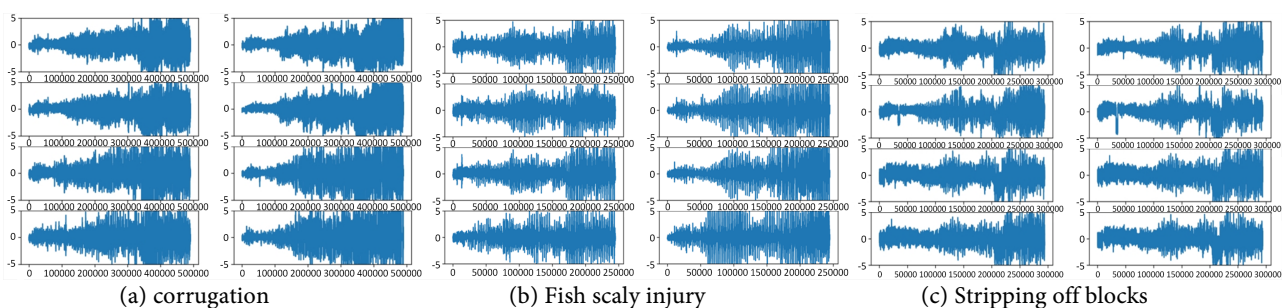(a) corrugation   (b) Fish scaly injury   (c) Stripping off blocks

Figure 8. Data set of rail lifecycle degradation.
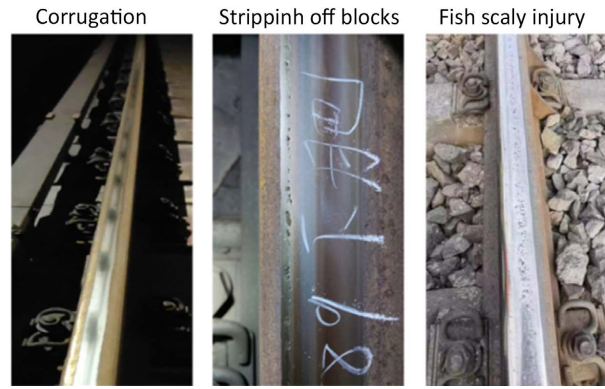
Figure 9. Three kinds of damage scene.

Table 2. Model structure parameter settings.

Types	Operation	parameter settings
Improved pulse codingmodule	Spilking Separable Convs_1 Spilking Separable Convs_2	
Spilking Separable Convs_1	Depthwise Convs	Conv2d, out_channels = 3, kernel_size = (3,3), padding = 1, groups = 3
	Pointwise Convs	Conv2d, out_channels = 128, kernel_size = (1,1), padding = 0, groups = 1
	BatchNorm2d	out_channels = 128
	neuron.IFNode	out_channels = 128
Spilking Separable Convs_2	Depthwise Convs	Conv2d, out_channels = 128, kernel_size = (3,3), padding = 1, groups = 3
	Pointwise Convs	Conv2d, out_channels = 128, kernel_size = (1,1), padding = 0, groups = 1
	BatchNorm2d	out_channels = 128
	neuron.IFNode	out_channels = 128
	ECABlock	out_channels = 128
	PulseMaxpool2d	out_channels = 128, kernel_size = (4,1), padding = 0
RUL prediction module	Transformer-Encoder MLP	
Transformer-Encoder	Transformer-Encoder_1 Transformer-Encoder_2	
Transformer-Encoder_1	Multi-head Attention	Input_dim = 3, d_model = 128, heads = 4, dropout = 0.1
	Add & Norm	d_model = 128, Num_layers = 2
	Feed Forward	dim_feedforward = 4* input_dim
	Add & Norm	d_model = 128, Num_layers = 2
Transformer-Encoder_2	Multi-head Attention	Input_dim = 3, d_model = 128, heads = 4, dropout = 0.1
	Add & Norm	d_model=128, Num_layers = 2
	Feed Forward	dim_feedforward = 4* input_dim
	Add & Norm	d_model = 128, Num_layers = 2
	PulseMaxpool2d	out_channels = 128, kernel_size = (8,1), padding = 0
MLP	LINEAR_1	Input = 32 * n; Ouput = 1
Training	Epoch = 1000	batch_size = 32; Optimizers = SGD; lr = 0.001

4.3. Experimental Design

1) This paper devised the experimental procedure based on three distinct aspects. The comparative and analytical examination of the experimental outcomes was conducted. Every experiment's training set and test set were divided in exactly the same way. As shown in Figure 8, there are three kinds of injuries in the rail data set, and each injury has eight vibration signal data. In order to fully extract and analyze the RUL information of each rail injury, the training set and the test set were divided by the ratio of 7:1. When testing the RUL of a certain type of damage, the full life data related to the same type of damage are used for training. And the weights that perform best throughout the training were kept.

2) To ascertain the legitimacy and indispensability of incorporating a modular design into the proposed model, an ablation study is conducted to perform a quantitative evaluation of the proposed approach.

3) To assess the efficacy of alternative activation functions, namely Sigmoid, Leaky RelU, and RelU, a comparative analysis of activation functions was conducted. Please refer to section 4.3.3 for the results of the comparison. To assess the efficacy of the proposed methodology, it was subjected to comparative analysis with a range of established and contemporary techniques for time series data prediction, including SNN, CNN, and TCN-SA [27].

4) In addition, to enhance the model's applicability and generalization capacity, the approach outlined in this study was implemented using the PHM2012 bearing degradation datasets. The results were compared with those of advanced methods performed with the same training data set and testing data set.

4.3.1. Experimental results and performance evaluation

First, according to the experimental design, as depicted in Figure 10, the most representative prediction effects of 3 kinds of injury are selected, the RUL prediction curve of the technique suggested can relatively well follow its real label value in most cases, especially in the second half of the whole life cycle. The above findings demonstrate that the technique suggested in this study achieves favorable fitting outcomes and can provide more accurate RUL prediction results.

4.3.2. Ablation Experiment

To assess the function and impact of various crucial modules within the proposed network, a process of ablation study is conducted whereby said modules are systematically removed or substituted while maintaining a constant number of remaining parameters. Ultimately, the subsequent four models have been developed for the purpose of conducting a comparative analysis.

1) Model-without-ECA: keep the other structure of the network unchanged, remove the attention mechanism, and the present model is utilized for the assessment of the attention mechanism.

2) Model-without-separable: Replacing the Separable convolution with a standard convolution and keeping the other structure of the network unchanged, this
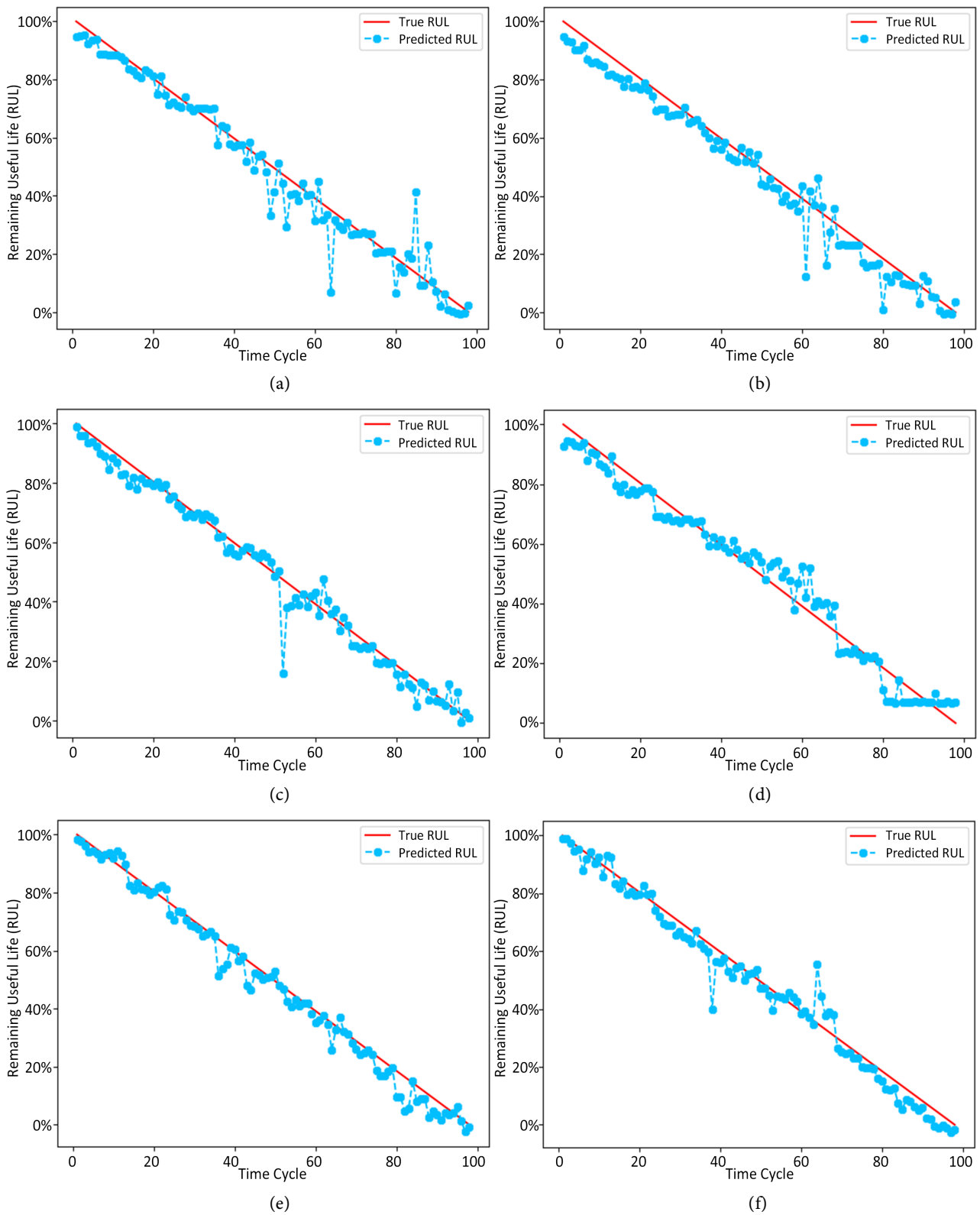
Figure 10. RUL prediction results of some rails: (a) A_1. (b) A_2. (c) B_2. (d) B_3. (e) C_2. (f) C_3.

Model can illustrate the ability of separable convolution to capture the interrelation between different sensor data.

3) Model-without-pulse maximum pooling: Keep other network structures unchanged and remove Pulse maximum pooling. This Model is used to evaluate the effect of the added pulse maximum pooling.

4) Standard-encoder: encoder without separable convolution and attention mechanism.

The experimental results are presented in Table 3. The superior predictive performance of the technique suggested is evident when compared to alternative models. The ablation experiment, the removal of separable convolution achieves a great impact on the network, which means that the incorporation of separable convolution is an effective means of modeling the pulse representation of multi-channel sensor data in the pulse coding module, and enables the preliminary feature extraction stage to capture the interdependence between different sensor data. Removing the ECA attention mechanism also affects the performance of the network, which indicates that the incorporation of the ECA attention mechanism enhances the network's ability to discern pulse features and reduces the interference of noise signals.

4.3.3. Comparative Experiment of Several Activation Functions
The activation function can construct the mapping between input features and output features in a nonlinear relationship, so that the network model has the

Table 3. Comparison of different methods under various operating conditions.

Method	Evaluating Indicator	Condition1 Injury Types			Condition2 Injury Types			Condition3 Injury Types			Condition4 Injury Types		
		Corrugation	Fish scaly injury	Stripping off blocks	Corrugation	Fish scaly injury	Stripping off blocks	Corrugation	Fish scaly injury	Stripping off blocks	Corrugation	Fish scaly injury	Stripping off blocks
Standard-encoder	MAE	11.11	13.02	18.55	12.46	19.29	18.55	15.54	19.29	14.45	12.24	18.56	18.53
	RMSE	13.68	14.94	22.12	14.49	23.58	22.12	20.16	23.58	18.49	18.52	22.12	22.13
	Score	0.73	0.72	0.63	0.72	0.63	0.63	0.68	0.63	0.73	0.70	0.62	0.63
W/o-Separable	MAE	3.26	4.45	10.69	6.22	3.19	6.23	6.67	3.71	4.51	4.20	3.37	7.58
	RMSE	4.40	6.10	15.29	9.03	3.99	7.78	11.02	4.80	5.95	6.67	4.24	12.49
	Score	0.90	0.90	0.84	0.84	0.91	0.87	0.89	0.90	0.92	0.94	0.88	0.90
W/o-ECA	MAE	9.20	12.80	12.00	7.33	7.26	8.39	13.18	5.41	8.38	10.34	3.03	9.57
	RMSE	12.87	15.60	16.03	9.39	11.52	13.64	17.06	7.06	11.43	13.40	3.81	15.38
	Score	0.75	0.80	0.78	0.77	0.80	0.81	0.62	0.84	0.79	0.78	0.90	0.75
W/o-Pulsemaximum pooling	MAE	9.18	9.56	9.41	6.44	10.04	8.28	8.45	9.75	3.63	3.90	9.76	8.31
	RMSE	12.70	12.99	12.83	12.44	13.26	10.27	10.67	12.71	4.64	5.54	12.71	11.27
	Score	0.82	0.80	083	0.88	0.72	0.79	0.78	0.76	0.91	0.92	0.76	0.78
Proposed method	**MAE**	**2.05**	**3.26**	**9.19**	**2.72**	**2.03**	**3.05**	**2.36**	**1.95**	**2.16**	**3.21**	**1.51**	**4.52**
	RMSE	**2.76**	**4.43**	**16.89**	**3.43**	**2.83**	**3.87**	**3.05**	**2.45**	**2.67**	**4.31**	**1.95**	**5.61**
	Score	**0.96**	**0.93**	**0.89**	**0.95**	**0.97**	**0.93**	**0.95**	**0.95**	**0.96**	**0.92**	**0.96**	**0.92**

ability to learn complex function mapping from input data, and can better predict, classify and judge the output results of the network. The alternative activation function gradient descent method is employed to address the challenge of non-differentiability exhibited by spiking neural networks during the back propagation process. To assess the efficacy of alternative activation functions, the activation functions of Sigmoid, Leaky RelU and RelU are employed for the purpose of comparing various activation functions. Taking condition 1 as an example, three experiments are designed for comparison, and the comparative outcomes are presented in Table 4. Figure 11 illustrates the comparison of RUL prediction effects of the three activation functions.

1) Model-with-Piecewise LeakyRelU: the selected Model.

2) Model-with-RelU: The pulse piecewise LeakyRelU activation function has been substituted for the standard RelU activation function, and the other structures of the network are kept unchanged. The efficacy of the pulse piecewise LeakyRelU activation function in augmenting the expressive capacity of deep neural networks can be validated by the Model.

3) Model-with-Sigmoid: The impulsive piecewise LeakyRelU activation function has been substituted for the Sigmoid activation function, and the other structures of the network are kept unchanged. This Model can verify the advantage of impulsive LeakyRelU activation function in enhancing the expression ability of deep neural networks.

Based on the findings presented in Table 4, it is evident that the prediction effect of using Sigmoid activation function is poor, mainly reflected in the gap between the evaluation indicators MAE, RMSE and Score values, and the pulse convolutional network is easy to disappear the gradient when the gradient is back propagated. In the case of RelU activation function, the evaluation index is slightly better than that of Sigmoid activation function. The model using Piecewise EleakyRelU activation function achieved the best prediction effect. Although the MAE and RMSE indexes of Piecewise EleakyRelU doubled compared

Table 4. Comparison of different activation functions under condition 1.

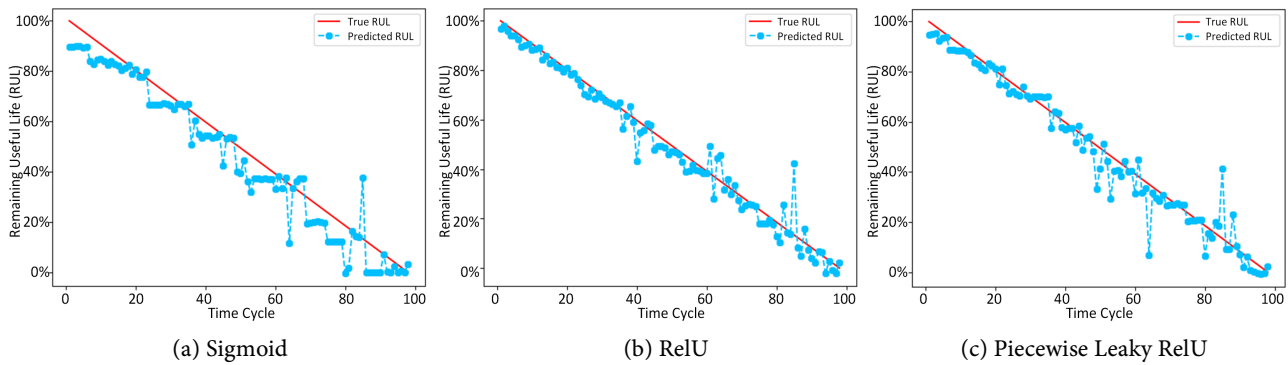| Method	Evaluating Indicator	Injury Types		
		Corrugation	Fish scaly injury	Stripping off blocks
Model-With-Sigmoid	MAE	3.15	3.78	7.90
	RMSE	4.30	4.83	10.43
	Score	0.93	0.93	0.86
Model-With-RelU	MAE	2.40	3.22	4.04
	RMSE	2.94	4.47	5.17
	Score	0.96	0.94	0.86
Model-With-Piecewise LeakyRelU	MAE	2.05	3.26	9.19
	RMSE	2.76	4.43	16.89
	Score	0.96	0.93	0.89

Figure 11. Prediction effect of different activation functions for Corrugation damage under condition 1.

with RelU, the score of the improved scoring function increased by 3.49%. The higher the score index, the more reliable the network RUL prediction in the later stage, thereby aligning more closely with practical engineering applications. Therefore, considering the performance evaluation index, the Piecewise Eleaky-RelU activation function is selected as a solution to address the issue of non-differentiability of spiking neural network in the process of back propagation.

4.3.4. Comparison Experiments of Different Networks

Secondly, to enhance the objectivity of the research findings, a comparative analysis is conducted between the proposed model and other time series forecasting models. Furthermore, a comprehensive examination of the predictive outcomes of the model is carried out. Table 5 compares the outcomes of rail RUL prediction using the proposed technique with the established CNN, SNN, and TCN-SA methods in the literature [27]. The present study demonstrates that the prediction model proposed in this paper exhibits superior performance in comparison to several alternative models, as evidenced by various performance evaluation metrics.

4.3.5. Research on Model Generalization

Finally, to evaluate the efficacy and generalizability of the proposed approach in predicting the remaining lifespan of the entire life cycle, it was implemented on the PHM2012 rolling bearing dataset [28] and compared against methods that have demonstrated exceptional predictive performance on the same dataset. The data was gathered utilizing the accelerated aging platform known as PRONOSTIA, and the PHM2012 rolling bearing dataset provided by PRONOTIA contains data on rolling bearings during the period from normal operation to failure. Two accelerometers were positioned in the horizontal and vertical orientations to acquire vibration signals along these two axes, utilizing a sampling frequency of 25.6 kHz.

There are similarities between the PHM2012 bearing dataset and the rail dataset in this paper, specifically in the following aspects: 1) Similarly, vibration signals in both directions were collected in horizontal and vertical directions for

Table 5. Comparison of prediction performance of different methods under various operating conditions.

Method	Evaluating Indicator	Condition1 Injury Types			Condition2 Injury Types			Condition3 Injury Types			Condition4 Injury Types		
		Corrugation	Fish scaly injury	Stripping off blocks	Corrugation	Fish scaly injury	Stripping off blocks	Corrugation	Fish scaly injury	Stripping off blocks	Corrugation	Fish scaly injury	Stripping off blocks
Regular-CNN	MAE	9.32	10.9	10.26	10.3	11.8	11.28	9.56	10.58	10.45	10.69	11.45	10.56
	RMSE	11.2	13.01	11.36	12.0	14.13	12.38	11.36	13.12	12.38	11.03	14.13	12.38
	Score	0.80	0.78	0.77	0.79	0.77	0.76	0.79	0.77	0.76	0.79	0.77	0.76
Regular-SNN	MAE	5.41	8.62	8.18	6.8	8.07	10.16	7.71	5.82	11.06	8.35	6.27	15.93
	RMSE	6.9	11.35	11.5	8.51	11.51	12.82	9.77	7.96	15.21	10.99	8.69	20.83
	Score	0.9	0.86	0.84	0.88	0.87	0.8	0.83	0.83	0.77	0.86	0.85	0.73
TCN-SA [25]	MAE	4.56	6.23	5.68	5.98	6.67	7.12	5.02	6.13	7.03	4.51	5.02	10.89
	RMSE	5.66	6.9	7.22	7.12	8.6	7.89	6.85	8.64	8.5	6.29	7.01	13.65
	Score	0.92	0.89	0.88	0.9	0.86	0.88	0.91	0.85	0.86	0.9	0.88	0.79
Proposed method	MAE	2.05	3.26	9.19	2.72	2.03	3.05	2.36	1.95	2.16	3.21	1.51	4.52
	RMSE	2.76	4.43	16.89	3.43	2.83	3.87	3.05	2.45	2.67	4.31	1.95	5.61
	Score	0.96	0.93	0.89	0.95	0.97	0.93	0.95	0.95	0.96	0.92	0.96	0.92

processing and analysis.

2) Also according to the experimental conditions, a variety of conditions were set up, which is convenient for a large number of network model tests and results analysis. The advantages of this option are: 1. Verify the generalization ability of the model: By testing the model on similar data sets, it is possible to better evaluate whether the model can handle previously unseen data, which is an important indicator of the generalization ability of the model. 2). Reduce the risk of overfitting: Using different datasets can help identify if the model is overfitting the characteristics of the training data, thus ensuring that the model can maintain good performance on new, unseen data. 3). Improve the applicability of the model: If the model performs well on multiple similar data sets, then it is more applicable and can be applied to real problems with more confidence.

In this paper, the utilization of data obtained from two distinct working conditions is demonstrated in Table 6, and both horizontal and vertical vibration signals are employed as input. The data collected at each time is divided into a sample, that is, the input data shape is 2 * 2560. In each condition, a single bearing data is designated as the test set while the remaining data are utilized as the training set. The output labels were established in a manner akin to the rail data, wherein the output label was determined by the lifetime percentage of the bearing, and the RUL of the bearing was normalized to a scale ranging from 0 to 100%.

The efficacy of the proposed methodology is evaluated using the bearing dataset and juxtaposed with alternative methodologies. Table 7 displays partial

Table 6. Operating conditions and bearing numbers in the PHM2012 dataset.

Operating condition	Loading force	Rotating speed	No. tested bearings
Condition 1	4000 N	1800 rpm	B1-1 - B1-7
Condition 2	4200 N	1650 rpm	B2-1 - B2-7

Table 7. Comparison of RUL prediction performance test results for PHM2012 bearing data set.

Method	Evaluating Indicator	Bearing number					
		B1-1	B1-2	B1-3	B2-4	B2-5	B2-6
Standard-encoder	MAE	22.64	22.88	22.51	22.81	22.73	23.07
	RMSE	26.79	27.10	26.69	27.21	26.62	27.40
	Score	0.59	0.60	0.58	0.57	0.58	0.59
Model-Without-Separable	MAE	13.02	10.50	10.53	14.12	14.63	15.16
	RMSE	15.95	13.76	14.66	17.91	19.45	18.10
	Score	0.71	0.79	0.78	0.71	0.67	0.71
Model-Without-ECA	MAE	14.55	17.47	15.58	17.23	18.14	17.86
	RMSE	18.55	21.91	19.31	21.21	21.77	23.34
	Score	0.71	0.66	0.67	0.63	0.65	0.65
Model-Without-Pulse maximum pooling	MAE	9.24	7.77	10.68	6.84	18.19	10.96
	RMSE	11.20	10.96	13.67	8.33	22.69	13.79
	Score	0.79	0.81	0.78	0.84	0.64	0.76
DANN [29]	MAE	13.3	15.3	28.4	12.0	22.7	20.6
	RMSE	16.0	17.2	33.5	14.3	27.5	23.3
TCN-SA [27]	MAE	10.4	13.0	9.7	5.18	13.3	10.9
	RMSE	11.3	16.9	11.7	6.42	16.1	13.8
	Score	0.87	0.80	0.81	0.91	0.78	0.83
Proposed method	**MAE**	**6.00**	**5.93**	**4.54**	**6.70**	**12.42**	**10.03**
	RMSE	**7.59**	**7.43**	**5.81**	**9.10**	**14.95**	**12.36**
	Score	**0.85**	**0.85**	**0.85**	**0.85**	**0.70**	**0.77**

findings. The findings indicate that the proposed model exhibits enhanced MAE, RMSE, and Score metrics when applied to the PHM2012 bearing dataset. Moreover, the proposed model demonstrates a favorable predictive performance.

5. Conclusions

The present study introduces a novel model for predicting RUL that utilizes an enhanced pulse separable convolution technique to augment the feature output of a Transformer encoder. In addition, ECA attention mechanism is incorporated to readjust the pulse feature map to highlight the useful pulse features. Then, based on the effective features, the multi-head self-attention mechanism is employed to comprehensively control the dependency of fault features at each

time step in the long time span of the entire lifespan. Finally, the final RUL result is obtained through the improved Transformer encoder algorithm and the linear layer. Through the generalization experiment with the widely used PHM2012 bearing data set, the findings indicate that the predictive capacity of the model proposed in this study surpasses that of other sophisticated algorithms currently available. This confirms that the model presented in this paper can significantly enhance the precision of rail Remaining RUL prediction.

In the future work, we are interested in further improving the prediction ability from the following two aspects: First, it is to search for an optimized tag setting method to better explain the RUL. Second, the combination of the end-to-end RUL prediction method and the traditional degradation mechanism of rail damage needs further exploration. In addition, the network proposed in this paper has the potential to be applied to aero-engine RUL prediction and other fields to further improve RUL prediction.

Funding Statement

Conflicts of Interest

The authors declare that they have no known conflicts of interest or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Shang, P.P., Liu, X.B. and Ma, S. (2022) Study on Rail Service Life of Shuohuang Heavy Haul Railway. *Railway Engineering*, **62**, 68-71+85.

[2] Ni, J.N., Zhang, W., Liu, N.X., *et al.* (2022) Demand Forecast and Development Countermeasures of Heavy Haul Railway Transportation in the "14th Five Year" Plan. *Railway Freight Transport*, **40**, 7-11+19.

[3] He, B., Liu, L. and Zhang, D. (2021) Digital Twin-Driven Remaining Useful Life Prediction for Gear Performance Degradation: A Review. *Journal of Computing and Information Science in Engineering*, **21**, Article ID: 030801.
https://doi.org/10.1115/1.4049537

[4] Khazeiynasab, S.R. and Qi, J. (2021) Generator Parameter Calibration by Adaptive Approximate Bayesian Computation with Sequential Monte Carlo Sampler. *IEEE Transactions on Smart Grid*, **12**, 4327-4338.
https://doi.org/10.1109/TSG.2021.3077734

[5] Muneer, A., Taib, S.M., Naseer, S., *et al.* (2021) Data-Driven Deep Learning-Based Attention Mechanism for Remaining Useful Life Prediction: Case Study Application to Turbofan Engine Analysis. *Electronics*, **10**, Article 2453.
https://doi.org/10.3390/electronics10202453

[6] Wang, C. and Kou, P. (2020) Wind Speed Forecasts of Multiple Wind Turbines in a Wind Farm Based on Integration Model Built by Convolutional Neural Network and Simple Recurrent Unit. *Transactions of China Electrotechnical Society*, **35**, 2723-2735.

[7] Kang, S.Q., Xing, Y.Y., Wang, Y.J., et al. (2023) Rolling Bearing Life Prediction Based on Unsupervised Deep Modeltransfer. *Acta Automatica Sinica*, **49**, 2627-2638.

[8] Shifat, T.A., Yasmin, R. and Hur, J.W. (2021) A Data Driven RUL Estimation Framework of Electric Motor Using Deep Electrical Feature Learning from Current Harmonics and Apparent Power. *Energies*, **14**, Article 3156. https://doi.org/10.3390/en14113156

[9] Dong, S., Wang, P. and Abbas, K. (2021) A Survey on Deep Learning and Its Applications. *Computer Science Review*, **40**, Article ID: 100379. https://doi.org/10.1016/j.cosrev.2021.100379

[10] Janiesch, C., Zschech, P. and Heinrich, K. (2021) Machine Learning and Deep Learning. *Electronic Markets*, **31**, 685-695. https://doi.org/10.1007/s12525-021-00475-2

[11] Sherstinsky, A. (2020) Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. *Physica D: Nonlinear Phenomena*, **404**, Article ID: 132306. https://doi.org/10.1016/j.physd.2019.132306

[12] Zhao, Z.H., Li, Q., Yang, S.P., *et al.* (2022) Research on Remaining Service Life Prediction Based on Bilstm and Attention Mechanism. *Journal of Vibration and Shock*, **41**, 44-50+196.

[13] Vaswani, A., Shazeer, N., Parmar, N., *et al.* (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems* (*NIPS* 2017), Long Beach, 4-9 December 2017, 6000-6010.

[14] Mo, Y., Wu, Q., Li, X., *et al.* (2021) Remaining Useful Life Estimation via Transformer Encoder Enhanced by a Gated Convolutional Unit. *Journal of Intelligent Manufacturing*, **32**, 1997-2006. https://doi.org/10.1007/s10845-021-01750-x

[15] Zhang, Z., Song, W. and LI, Q. (2022) Dual-Aspect Self-Attention Based on Transformer for Remaining Useful Life Prediction. *IEEE Transactions on Instrumentation and Measurement*, **71**, Article No. 2505711. https://doi.org/10.1109/TIM.2022.3160561

[16] Chen, Z.H. (2023) Research on Residual Life Prediction Technology of Electromechanical Actuator Based on Multimodaltransformer. *Acta Armamentarii*, **44**, 2920-2931. http://kns.cnki.net/kcms/detail/11.2176.TJ.20221024.1556.003.html

[17] Maas, W. (1997) Networks of Spiking Neurons: The Third Generation of Neural Network Models. *Neural Networks*, **10**, 1659-1671. https://doi.org/10.1016/S0893-6080(97)00011-7

[18] Hong, C. and Wang, J. (2020) Short Term Load Forecasting Model Based on Pulse Neural Network. *Proceedings of the CSU-EPSA*, **32**, 139-144.

[19] Wang, B., Lei, Y., Li, N., *et al.* (2019) Deep Separable Convolutional Network For-remaining Useful Life Prediction of Machinery. *Mechanical Systems and Signal Processing*, **134**, Article ID: 106330. https://doi.org/10.1016/j.ymssp.2019.106330

[20] Mamalet, F. and Garcia, C. (2012) Simplifying Convnets for Fast Learning. *International Conference on Artificial Neural Networks* (*ICANN*), Lausanne, 11-14 September 2012, 58-65. https://doi.org/10.1007/978-3-642-33266-1_8

[21] Kong, L., Min, Y., He, J., *et al.* (2022) Research on Steel Surface Defect Recognition

Based on Pulse Neural Network. *Packaging Engineering*, **43**, 13-22.

[22] Zhang, C., Huang, C., He, J., *et al.* (2022) Defects Recognition of Train Wheelset Tread Based on Improved Spiking Neural Network. *Chinese Journal of Electronics*, **32**, 941-954. https://doi.org/10.23919/cje.2022.00.162

[23] Holden, A.V. (2013) Models of the Stochastic Activity of Neurones. In: Levin, S., Ed., *Lecture Notes in Biomathematics*, Vol. 12, Springer, Heidelberg.

[24] Wang, Q., Wu, B., Zhu, P., *et al.* (2020) ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 11534-11542. https://doi.org/10.1109/CVPR42600.2020.01155

[25] LI, X., Wang, W., Hu, X., *et al.* (2020) Selective Kernel Networks. *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 510-519. https://doi.org/10.1109/CVPR.2019.00060

[26] Sengupta, A., Ye, Y., Wang, R., *et al.* (2019) Going Deeper in Spiking Neural Networks: VGG and Residual Architectures. *Frontiers in Neuroscience*, **13**, Article 95. https://doi.org/10.3389/fnins.2019.00095

[27] Wang, Y., Deng, L., Zheng, L., *et al.* (2012) Temporal Convolutional Network with Soft Thresholding and Attention Mechanism for Machinery Prognostics. *Journal of Manufacturing Systems*, **60**, 512-526. https://doi.org/10.1016/j.jmsy.2021.07.008

[28] Chen, Y.Y., Peng, G., Zhu, Z., *et al.* (2020) A Novel Deep Learning Method Based on Attention Mechanism for Bearing Remaining Useful Life Prediction, *Applied Soft Computing*, **86**, Article ID: 105919. https://doi.org/10.1016/j.asoc.2019.105919

[29] Li, X., Zhang, W., Ma, H., *et al.* (2020) Data Alignments in Machinery Remaining Useful Life Prediction Using Deep Adversarial Neural Networks, *Knowledge-Based Systems*, **197**, Article ID: 105843. https://doi.org/10.1016/j.knosys.2020.105843