


Article

# Deep Learning-Based Remaining Useful Life Prediction Method with Transformer Module and Random Forest

Lefa Zhao <sup>1,\*</sup>, Yafei Zhu <sup>2</sup> and Tianyu Zhao <sup>3,\*</sup> <sup>1</sup> School of General Education, Shenyang Sport University, Shenyang 110115, China<sup>2</sup> International Engineering College, Shenyang Aerospace University, Shenyang 110136, China<sup>3</sup> Key Laboratory of Structural Dynamics of Liaoning Province, College of Sciences, Northeastern University, Shenyang 110819, China

\* Correspondence: larry2012@syty.edu.cn (L.Z.); zhaotianyu@mail.neu.edu.cn (T.Z.)

**Abstract:** This paper focuses on the prognosis problem in manufacturing of the electronic chips for devices. Electronic devices are of great importance at present, which are popularly applied in daily life. The basis of supporting the electronic device is the powerful electronic chip and its manufacturing technology. Chip manufacturing has been one of the most important technologies in recent years. The etching machine is the key equipment in the etching process of the wafers in chip manufacturing. Due to the high demands for precise manufacturing, monitoring the health state and predicting the remaining useful life (RUL) of the etching system is quite important. However, the task is very hard because of the lack of knowledge of exact onset of failure or degradation and the multiple operating conditions, etc. This paper proposes a novel deep learning-based RUL prediction method for the etching system. The transformer module and random forest are integrated in the methodology to identify the health state of the machine and predict its RUL, through training with the complex data of the etching machine's sensors and exploring its underlying features. The experiments are based on the subject of the 2018 PHM Data Challenge—for estimating time-to-failure or RUL of Ion Mill Etching Systems in an online fashion using data from multiple sensors. The results indicate the proposed method is promising for the real applications of the prognosis of the etching system for electronic devices.

**Keywords:** deep learning; remaining useful life prediction; transformer; random forest**MSC:** 68T01

**Citation:** Zhao, L.; Zhu, Y.; Zhao, T. Deep Learning-Based Remaining Useful Life Prediction Method with Transformer Module and Random Forest. *Mathematics* **2022**, *10*, 2921. <https://doi.org/10.3390/math10162921>

Academic Editor: Yolanda Vidal

Received: 10 July 2022

Accepted: 10 August 2022

Published: 13 August 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

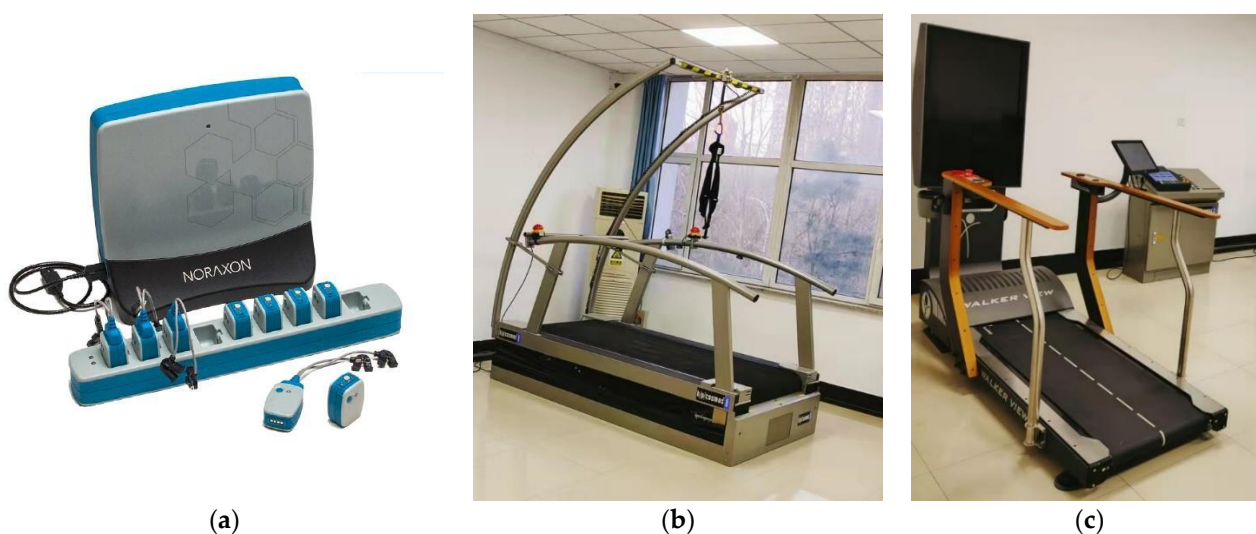


**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

In recent years, deep learning technology has made a great breakthrough in image recognition, natural language processing, etc. In industry, machine condition monitoring has benefited well from deep learning [1–3]. Many researchers have used the deep learning algorithm to monitor the operating situation of all kinds of equipment and investigate the problem of remaining useful life (RUL) prediction, such as the pipeline system of a nuclear power station [4]. Traditionally, researchers could resolve the related problems for some simple structures through mechanical modeling [5,6], but the complexity of more equipment, such as the ion mill etching system, is far beyond the ability of traditional technique.

In daily life, electronic devices such as computers and personal smartphone require the high performance of electric hardware. Electronic chip manufacturing technology has become a nation's core strategy of science and technology. Professional equipment is highly dependent on the electronic chips. For instance, the sensors with chips for training athletes and evaluating their performance require high-level electronic chips. Examples of the electronic devices are shown in Figure 1. The etching machine is one of the most important systems in electronic chip manufacturing. Due to the complex working system of the etching machine, it is difficult to monitor its health state and predict the RUL.



**Figure 1.** Examples of electronic devices with chips in athlete training. (a) Signal measurement device for electromyography. (b) Intelligent running training device. (c) Digital monitoring device for running.

In the current literature, LSTM (long-short-term memory network) of RNN (recurrent neural network) [7] have been popularly used for the RUL prediction problem. Other kinds of machine learning methods have also been used on this issue [8–11]. However, more and more researchers have addressed the related problem by using LSTM [12–14]. For investigating time series data problems, the LSTM network has a good learning capability [15]. The etching machine condition is also a time series data analysis problem.

In the deep learning-based studies, researchers adopted the Attention Model to solve the disappearance of long-short-term memory network gradients [16–18]. Although the attention machine successfully solved the gradient disappearance issue of LSTM, it also increased more computation to the learning task. Therefore, researchers further proposed the self-attention mechanism, which removes the structure of the long-short-term memory network and uses the attention machine directly to establish a network. In this way, the network can learn the relevance of information in time. Additionally, it can further overlay the network structure into a multi-head self-attention module block. Afterwards, the researchers created a new deep learning module with the structure of an auto-encoder network, which is named as the Transformer network.

Since the Transformer network was proposed, it has shown excellent learning in natural language [19–21], and compared with the basic CNN and RNN networks, the Transformer network outperforms them in the related problems. In addition, the Transformer network could not cause gradient vanishing or gradient exploding as easily as the RNN.

This paper proposes a novel deep learning-based method for the RUL prediction problem of the etching system. The Transformer module is integrated with the random forest algorithm to improve the learning capability of the model for processing the condition-monitoring time series data. Experiments on the real etching machine data are carried out for validations. The results of the experiments show that the proposed method is promising for the RUL prediction problem in real industries.

Section 2 presents the data set and the pre-processing of it. Section 3 introduces the methodologies proposed in this paper. Section 4 provides the experimental results for validations. We close the paper with conclusions in Section 5.

## 2. Data Set

### 2.1. Data Set

The data set used in this study is from the competition topic of the 2018 PHM Data Challenge. The data set is provided by the competition, including the training set and the

prediction task data set. Additionally, it includes three kinds of data of 20 etching machines: physical information of etching machine operation, fault occurrence operation and fault occurrence countdown. The prediction task data set includes 5 data of etching machines, and it has only 2 categories: physical information of the etching machine operation and the countdown to the occurrence of the fault as the predicted answer. Etching machine failures are divided into three parts:

1. FlowCool Presser Dropped Below Limit;
2. FlowCool Presser Too High Check FlowCool Pump;
3. FlowCool Leak.

The sensor that collects physical information of etching machine operation contains the contents shown in the following Table 1.

**Table 1.** Operation susceptor of etching machine.

ID#	Parameter Name	Type	Description
S1	time	Numeric	time
S2	Tool	Categorical	tool id
S3	stage	Categorical	processing stage of wafer
S4	Lot	Categorical	wafer id
S5	runnum	Numeric	number of times tool has been run
S6	recipe	Categorical	describes tool settings used to process wafer
S7	recipe_step	Categorical	process step of a recipe
S8	IONGAUGEPRESSURE	Numeric (Sensor)	pressure reading for the main process chamber when under vacuum
S9	ETCHBEAMVOLTAGE	Numeric	voltage potential applied to the beam plate of the grid assembly
S10	ETCHBEAMCURRENT	Numeric	ion current impacting the beam grid determining the amounts of ions accelerated through the grid assembly to the wafer
S13	FLOWCOOLFLOWRATE	Numeric	rate of flow of helium through the flowcool circuit, controlled by mass flow controller
S14	FLOWCOOLPRESSURE	Numeric (Sensor)	resulting helium pressure in the flowcool circuit
S15	ETCHGASCHANNEL1-READBACK	Numeric	rate of flow of argon into the source assembly in the vacuum chamber
S16	ETCHPBGAS-READBACK	Numeric	rate of flow of argon into the PBN assembly in the chamber
S17	FIXTURETILTANGLE	Numeric	wafer tilt angle setting
S18	ROTATIONSPEED	Numeric	wafer rotation speed setting
S19	ACTUALROTATION-ANGLE	Numeric (Sensor)	measure wafer rotation angle
S20	FIXTURESHUTTER-POSITION	Numeric	open/close shutter setting for wafer shielding
S21	ETCHSOURCEUSAGE	Numeric	counter of use for the grid assembly consumable
S22	ETCHAUXSOURCE-TIMER	Numeric	counter of the use for the chamber shields consumable
S23	ETCHAUX2SOURCE-TIMER	Numeric	counter of the use for the chamber shields consumable
S24	ACTUALSTEPDURATION	Numeric (Sensor)	measured time duration for a particular step

The physical information of the etching machine operation is collected every four seconds, but there are also many irregular collection points. Every data set has a different size, but all of them are above 70 million seconds. The data set file of the time of failure contains the time of occurrence and category of failure.

The fault occurrence countdown file contains countdown time points for multiple occurrence points of three types of faults, and these points correspond to the time points collected by physical information of the etching machine.

In the training network, we use the information collected by all numerical type sensors except S1: time. Among them, the information collected by the water fixture position sensor (S20: FIXTURESHUTTER-POSITION) is the position of the water on the platform during the etching process. There are four different positions, which are represented by five numbers—they are 0,1,2,3,225. To facilitate the study, we change the five positions into a one-hot code represented by a binary vector. Other vectors which were selected in the data set have been normalized by the provider and they are suitable for further study.

Strictly speaking, the time point recorded by the fault occurrence time data set is not an accurate time of fault occurrence. Instead, the time point is when the etching machine shuts down due to the failure. The actual failure should be much earlier than these time points.

## 2.2. Pre-Process

There are three types of etching machine failures. This study assumes that the three types of failure are independent of each other. In the training model, the data is processed into three data sets based on the three types of failure, and three learning networks are trained for each of the three failures.

The etching machine used has running data up to tens of millions of seconds, and among these data, data which is too early can be considered to be in a healthy state. However, the occurrence of failure or shutdown should originate at a later point in time than abnormal operation. So, to directly learn the early data has no meaning for the remaining prediction of efficient life, and it will occupy a significant amount of time and resources of calculation. Therefore, we should train a random forests model which is used to monitor the time point when abnormal operation occurs.

At present, since there are no relatively abnormal operations and scientific theory on the concerned data, the references of the related study set the period at more than 5500 s from shutdown as a healthy time, and considers the period less than 5000 s as the time of abnormal operation. The data between the two steps of time will not be used.

However, the data of healthy time is far more than abnormal time. Such an uneven data set can cause the random forests, which we would use in the methodology, to have difficulty studying the characteristic of abnormal time. Hence, we should obtain the interpolation and fit for the abnormal time data, and then the original data is about sampling 1000 times the number of samples (different data sets have different and uneven numbers of failures, so they need to be adjusted according to the situation) to form new abnormal time data. Furthermore, among the data of healthy time and abnormal data, we choose smaller data, respectively, as a sample. Finally, we merge them into the training set of random forests, with 20% of them as the test set.

For the Transformer, as in random forest, the data set is processed into three data sets based on three failures for model training. To predict the RUL, it is necessary to input a time sequence. To reduce the amount of calculation, the data will be down-sampled. According to the prediction strategy, the Transformer network of this subject only learns from the data of the abnormal running time.

In down-sampling, for a sequence of failures, taking the downtime as the starting point, retrospective sampling is carried out. One sample point is taken for each R point, and M sample points are taken for one sampling. After one sampling, the first sample point after the starting point is moved to the starting point is repeated the above operation, and the operation is repeated N times. In this way, the sampling will overlap the same piece of fault information, but there is a certain degree of difference, which enhances the learning information and gives the trained model a certain degree of robustness.

During the experiment, the data set exposed some problems. For example, when the data pre-processing adopts the sampling method of  $R = 15$ ,  $M = 300$ ,  $N = 15$ , theoretically, the method  $M \times R + N = 4515$  requires retrospective sampling from the location of the

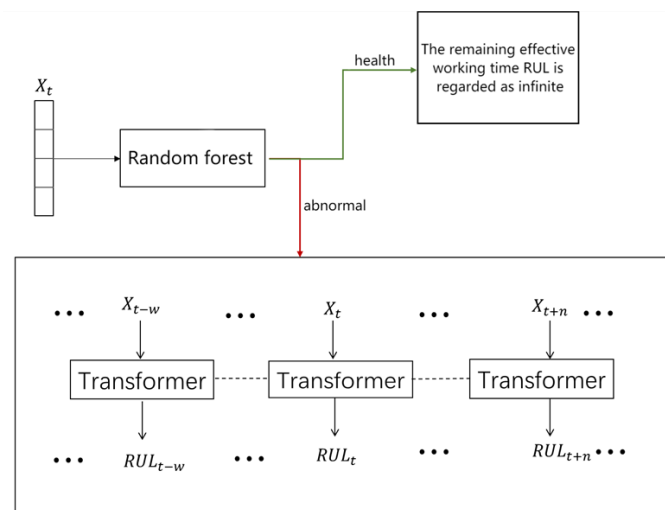
failure, and it is required to be able to collect data points, which represents a time series of  $4515 \times 4 = 18,060$  s. However, in actual operation, it is found that a lot of time is skipped in the time before the data set is close to the fault location. In other words, the time point before the fault location is not evenly collected at the frequency of collecting a data point every 4 s. The overall span of some samples can reach more than 30,000 s, and the maximum can reach 80,000 s, which is far greater than 18,060 s. Therefore, in order to ensure that there are sufficient and effective samples, in the actual data pre-processing, the data can only be required to include as many data points as possible within 18,000 s according to the specific conditions of the data set. Taking the data pre-processing of the failure type of “coolant pressure falling below the limit” in the data of the etching machine training set numbered 03\_M01\_DC as an example, in this training set, such failures occurred 43 times in total. According to the collection scheme described above, when the requirement is within 18,000 s, it contains 2000 data points, and 255 samples can be obtained.

### 3. Methodology

#### 3.1. Overall Structure

Because early operating time can be considered as healthy running time, this stage has no relationship with the eventual failure. To strengthen efficiency, this study first uses the random forest algorithm to examine the abnormal operation time point in the operation of the etching machine, and then the data in the abnormal operation stage is used to predict the RUL.

In the stage of RUL, the Transformer network is used. However, the original Transformer network is faced with the problem of processing natural language and a complex structure and a huge amount of calculation. Based on referring to the improved network of the existing Transformer, only the original Transformer can be used in the part of enclosed code. After being improved, it will be used in the prediction tasks. The detailed process is shown in Figure 2, where  $X_t$  is a high-dimension input data at time t, while w and n represent the time steps.



**Figure 2.** The overall process of random forest detection of abnormal running time and Transformer prediction of RUL.

#### 3.2. Random Forest

In this paper, because the Transformer leads to a huge calculation amount for every input time-series data, we cannot input all the fragments of the raw data. Additionally, the fragments of the raw data could be divided into two kinds, such that one is the health data and the other one is the abnormal data. In addition, just the abnormal fragments would cause the reduction in the machine’s RUL. Therefore, the Transformer could just train with the abnormal fragments of the raw data.

In order to identify the healthy state of the etching machine, the random forest is used to process the raw data [22–24], which is one of the most popular and efficient machine learning methods. Although the machine learning methods are not as effective as the deep learning methods for many tasks with complex data, their simple structures are more efficient than the deep learning methods. So, a machine learning method should always be used to pre-process data before a complicated deep learning method.

The random forest method is based on the decision tree method, and the random forest, simply speaking, is a combination of many decision trees [25,26]. As the basic machine learning method, it has three kinds, including ID3, C4.5 and CART (Classification and Regression Tree). The main difference between them is the different algorithm for the split of every node. Generally, CART has the best effect, which is selected in this paper.

Random forest is an integrated learning method in machine learning, which connects many decision trees. An intuitive explanation is that a decision tree is a classifier, and then n decision trees will have n classification results. Random forest is to integrate the classification results of these n trees. If the result of the decision tree is regarded as a vote for a certain category, the category with the most votes is finally selected as the result.

The study uses scikit-learn library in Python 3.8 to build a random forest algorithm which can be used to detect the abnormal running time point. Additionally, the MAE (Mean Absolute Error) can be used to calculate the correct rate in the testing phase, which is shown as Equation (1).

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \tag{1}$$

In Equation (1),  $y_i$  represents the true value of data category,  $\hat{y}_i$  is category which is predicted by the random forest, and n is the number of samples.

### 3.3. Transformer

The Transformer is a powerful and advanced method of deep learning, which is an important technology of artificial intelligence. Additionally, deep learning can be roughly classified into three categories according to the structure of deep artificial neural networks, includes convolution neural networks (CNN), recurrent neural network (RNN) and auto-encoder (AE). They have great success in resolving the problems with time-series data, such as natural language processing tasks, video identification problems and instrument diagnosis [27–29]. The panorama of Transformer structure is displayed in Figure 3. It is obvious that the kernel of the Transformer network is the multi-headed self-attention mechanism.

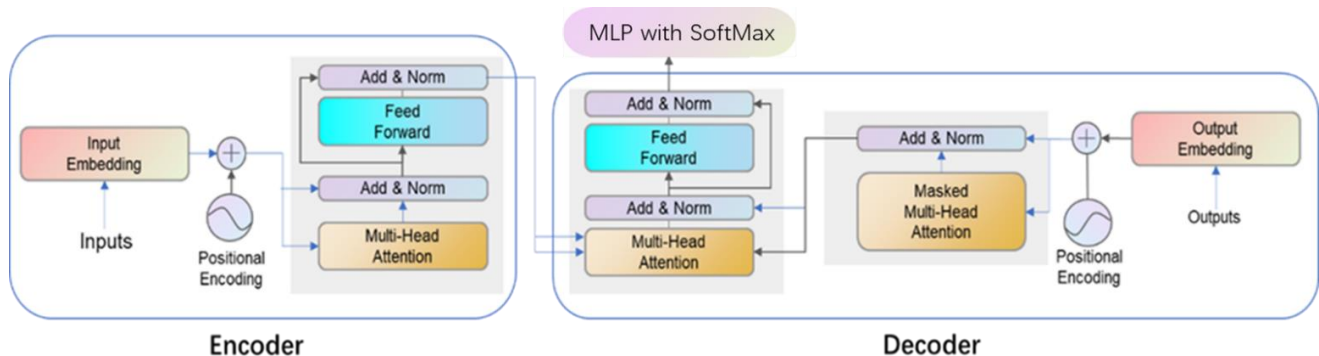


Figure 3. Transformer network structure.

The multi-head attention structure is a stacked self-attention mechanism. Additionally, self-attention is a variant of the attention mechanism. It reduces dependence on external information and is better at capturing the internal correlation of data. Its calculation includes three vectors,  $Q, K, V$ , named query, key, value, respectively, which have the same dimensions. They are obtained by Equation (2). The matrix  $X = \{x_1, x_2, \dots, x_n\}$

represents a sample of input data for the Transformer, and every  $x_i$  is one of the points in the time-series data, which contains the signals of the etching machine’s sensors and means its dimension is 20 in the study. Additionally, the  $Q, K, V$  are the information about  $X$  in three kinds of state, which are the  $X$  itself in other words. In the way of the attention mechanism, which is the source of the self-attention mechanism, the  $Q$  means what we want to know, the  $K$  means the crucial information we already have, and  $V$  is the raw information where we want to explore some underlying content. In the attention mechanism, the three matrixes could come from different information; however, in the self-attention mechanism, they all come from the  $X$  mechanism, which means what we explore is the  $X$  itself. In addition, the three matrixes could be defined as  $Q = \{q_1, q_2, \dots, q_n\}$ ,  $K = \{k_1, k_2, \dots, k_n\}$  and,  $V = \{v_1, v_2, \dots, v_n\}$ , where  $n$  is the number of time steps in the sequence and equals to 2000 in the study. In Equation (2),  $W_Q, W_K, W_V$  are three trainable parameter matrixes.

$$\begin{aligned} Q &= X^T \cdot W_Q \\ K &= X^T \cdot W_K \\ V &= X^T \cdot W_V \end{aligned} \tag{2}$$

Then, the calculation of the self-attention mechanism can be expressed as Equation (3), and the  $d_k$  is the dimension of the three matrixes  $Q, K, V$ , which depends on the dimension of input data and the head number of multi-headed self-attention mechanism. The multi-head self-attention mechanism is averagely dividing the  $d_k$  into few parts then continuing the calculation of the self-attention mechanism for each part.

$$Self-Attention(Q, K, V) = softmax\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V \tag{3}$$

The early Transformer networks’ task is natural language processing [30–33]. Their input training data and label data language sentences with rich meanings, so we should encode the training data information by encoder and decode label data information. However, the task in this study is to predict the RUL of the etching machine based on the physical information sequence running. The input training data is the physical information of the working condition with rich information, but the label data is only a countdown sequence to the downtime, and it has no profound information.

After referring to the structure of the Vision Transformer (ViT), only the encoder part of the original Transformer network will be used, and the word vector-encoding part will be removed [34–37]. This study uses TensorFlow and Keras library in Python 3.8 to achieve the encoder part of the predictive network, which also means the original Transformer network. The Hyperparameters that the network needs to determine include the number of training times, the number of heads of the multi-head attention mechanism, the number of layers superimposed by the autoencoder, and the dimensions of the Feed Forward network. The structure of our proposed method is showed in Figure 4.

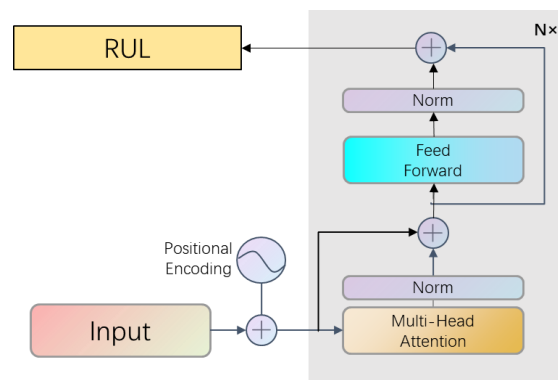


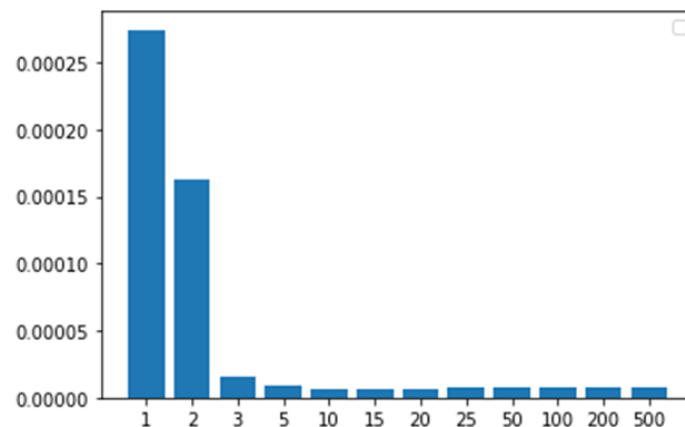
Figure 4. The improved Transformer structure used in this study.

The model uses Adam optimizer as the loss function. Take 80% of the data as training data, 20% as testing data. In the test, the mean absolute error is used to calculate the error between the true value and the predicted value of the proposed method [38].

#### 4. Experimental Results

##### 4.1. Study on the Number of Decision Trees in Random Forest

In the random forest algorithm, the most important parameter is the number of decision trees in random forest. Too few decision trees will impact the accuracy of the category, and many more algorithms will waste time and calculation resources. Due to the failure of each data set and the number of differences being huge, some of the data sets are too small, and thus, this study chooses the data of Number 03\_M01\_DC, which is in the etching training data regarding the drop in cooling fluid pressure to below the limit failure types of data. Random forest uses it to train algorithms to determine the number of decision trees. In that training set, such faults occurred a total of 40 times, with a large number of occurrences and abundant available data. After pre-processing, we obtained 151,500 samples, half of which are healthy data points, while the others are abnormal data points. Additionally, we then disorder them randomly. The preset numbers of decision trees are 1, 2, 3, 5, 10, 15, 20, 25, 50, 100, 200 and 500, and then, we conduct several experiments, respectively. Finally, the MAE of the trained random forest algorithm in the validation set is recorded under these parameters. The results are shown in Figure 5.



**Figure 5.** The relationship between the accuracy of random forest classification and the number of decision trees.

The horizontal axis is the number of decision trees, and the vertical axis is the average absolute error of the random forest algorithm corresponding to the number of decision trees in the verification set. We can see intuitively that when the number of decision trees is less than or equal to 10, the average decision error of the random forest algorithm reaches the lowest level, and the average absolute error is 0.000006. Additionally, we can ensure that the number of random forest decisions is 10.

##### 4.2. Correlation Analysis of Data Set Multidimensional Variables and Abnormal Working Conditions

After the random forest has learned the training set, according to the structure of the decision tree, the various dimensional variables of the input data can be analyzed, and the correlation between each dimension and the learning result can be obtained. After learning all 20 etching machine training sets and combining the results of 20 random forest algorithms, the correlation between the input data and the learning results is shown in Figure 6.



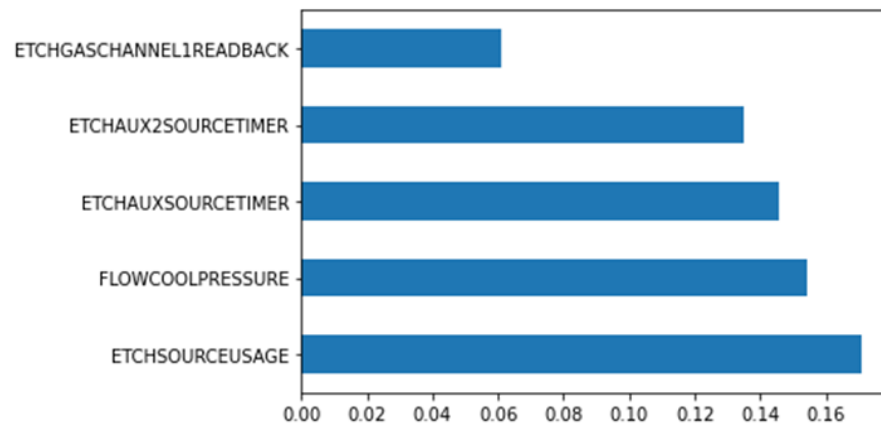


Figure 6. The 5 variables with the highest correlation between input data and results.

The vertical axis is the five variables with the highest correlation with the learning results, and the horizontal axis is the correlation score corresponding to the five variables in deep learning. According to this score, follow-up manual research on the law of the relationship between etching machine failure and working conditions, or related research on maintenance work, can refer to this evaluation, focusing on variables with high correlation scores.

4.3. Research on Hyperparameters of Transformer Prediction Model

In this section, using the training set of the etching machine numbered 03\_M01\_DC obtained in Section 2.2, we first simply set up the network: we set the number of heads of the multi-head attention mechanism to 1, the number of layers superimposed by the auto-encoder is set to 1, and the dimension of the feed-forward network is set to 8. At this time, the prediction network has a simple structure, and low computational complexity. It is used only to determine whether the prediction network is effective when learning the data set. We take reading all samples as one learning, and observe whether the prediction network loss function drops and converges. The result is displayed in Figure 7.

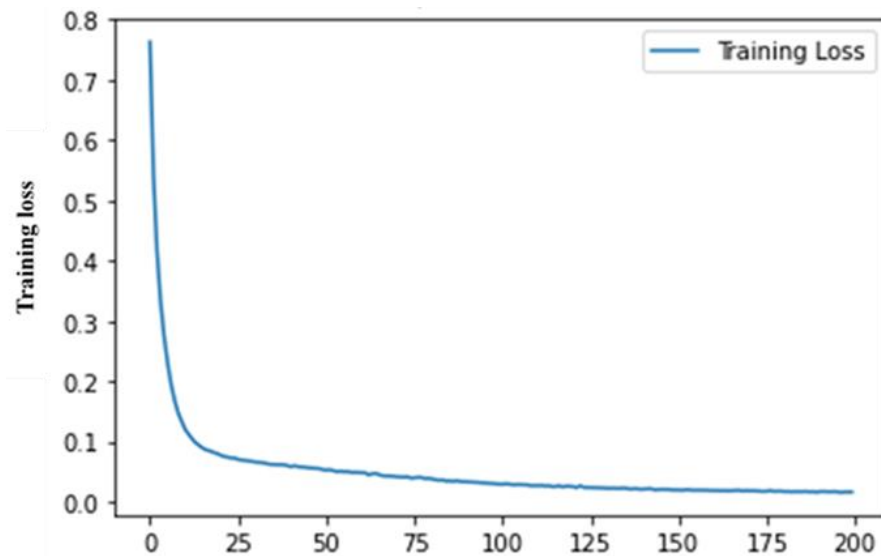


Figure 7. Forecasting the network loss function to change with the number of training steps.

As shown in Figure 7, it is obvious that the loss function of the prediction network has dropped significantly in the first 25 training steps, and its loss function gradually decreases

and tends to be flat with the increase in the number of training times. It can be seen that the prediction network has effectively learned the data set.

According to the results, the number of training times is determined to be 100 times, and only the number of heads of the multi-head attention mechanism is changed to explore the influence of the number of heads of the multi-head attention mechanism on the predictive network learning effect. When setting the head number of the multi-head attention mechanism, it must be set as a factor of the input data dimension. In this experiment, the input data dimension is 20, so the head number of the multi-head attention mechanism can only be set to 1, 2, 4, 5, 10, or 20. The average absolute error of the prediction network in the test is observed under six different parameters.

As shown in Table 2, when the number of hearts of the multi-head attention mechanism is 2, the average absolute error of the prediction network is the lowest. After determining the number of hearts of the multi-head attention mechanism as 2, and then exploring the influence of predicting the dimension of the feed-forward neural network of the network, according to various Transformer research recommendations, the dimension of the feed-forward neural network should be set to the power of 2 as much as possible. The experimental results are as follows in Table 3.

**Table 2.** The number of heads of the multi-head attention mechanism and the average absolute error of the prediction network.

Number of Heads of Multi-Head Attention Mechanism	Mean Absolute Error (Unit: Second)
1	2840.07
2	2561.87
4	3598.99
5	2807.26
10	3042.27
20	2971.62

**Table 3.** The number of heads of the multi-head attention mechanism and the average absolute error of the prediction network.

Feedforward Neural Network Dimensions	Mean Absolute Error (Unit: Second)
8	5716.82
16	4076.40
32	3736.63
64	3892.85
128	3786.88
256	2844.60
512	2248.19
1024	2968.11
2048	3006.22

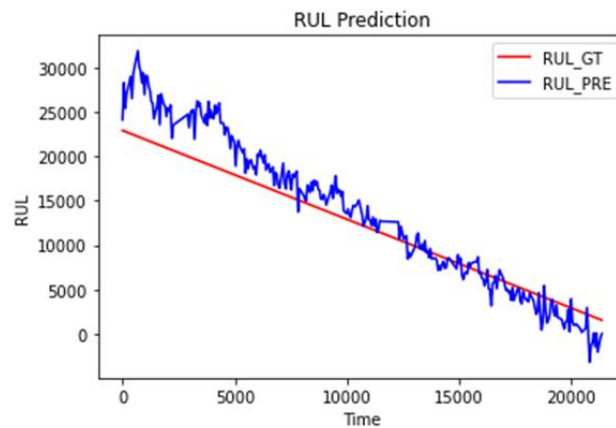
According to the current research experience of various Transformer networks, when the dimension of the feed-forward neural network is slightly larger than the length of the input sequence, the learning effect of the Transformer network is the best. The input sequence length in this experiment is 300 steps of time, as shown in Table 3. When the feed-forward neural is 512 dimensions, the training effect is the best, which is consistent with this experience. Finally, the influence of the number of auto-encoder stacks on the learning effect of the prediction network is studied. The results are shown in Table 4. When the number of auto-encoder stacks is 4, the average error is the lowest.

**Table 4.** The number of superimposed layers of the auto-encoder and the average absolute error of the prediction network.

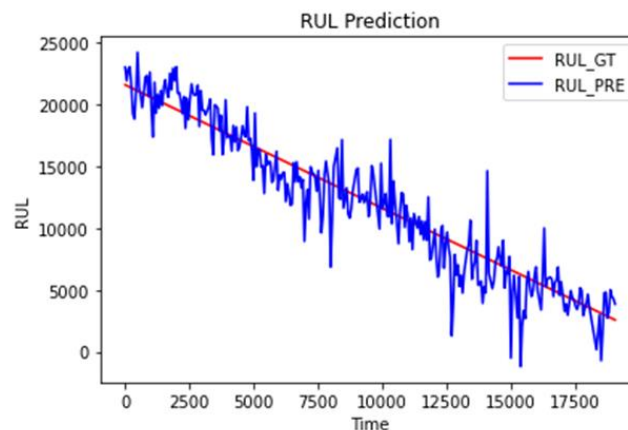
Number of Superimposed Layers of Autoencoder	Mean Absolute Error (Unit: Second)
1	4832.86
2	2615.82
3	2235.90
4	2138.53
5	2452.54

4.4. The Complete Network Performs Prediction Tasks

In this experiment, three sets of training sets were obtained by pre-processing part of the training set data of the etching machine numbered 03\_M01\_DC concerning the three types of failures, in which 204 training samples and 51 test samples were obtained for the “Flow Cool Pressure Dropped Below Limit” fault type, 60 training samples and 15 test samples were obtained for “Flow cool Pressure Too High Check Flow cool Pump” fault type, the Transformer model of this study was used for training, and 120 training samples and 30 test samples were obtained for the “Flow cool leak” fault type. Figures 8–10 show the prediction results of the RUL of a certain sample of the corresponding fault type test sample after the model has learned the three training sample data sets. The red broken line “RUL\_GT” is the actual RUL, and the blue broken line “RUL\_PRE” is the RUL predicted by the model.



**Figure 8.** Prediction of RUL of the “Coolant pressure drops below the limit” fault type.



**Figure 9.** Prediction of RUL of “cooling pressure too high” fault type.

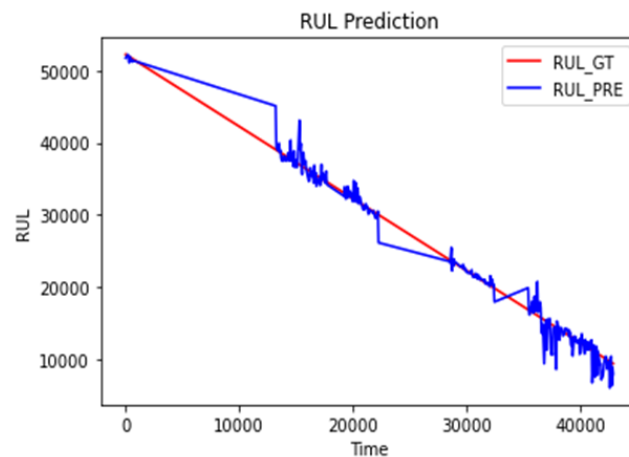


Figure 10. Prediction of RUL of “coolant leakage” fault type.

It can be seen from the following three figures that although the same sampling method is used, the period of the samples is very different, which directly reflects that the original data sampling is not sampled once and evenly every four seconds. It also brings great difficulties to the learning task. It can be seen from the three graphs that the blue line representing the predicted value still fluctuates tightly along the red line representing the true value, which shows that the model used in this topic has better prediction ability.

This paper randomly selects three sequences about the failure type “Flow Cool Pressure Dropped Below Limit” from any etching machine as the data set of the complete network prediction task. The specific operation is to detect the occurrence point of the abnormal operation in the data by the trained random forest model, and then use the occurrence point as the starting point. The sampling is started along the time direction of the data, and the Transformer model is used as input to predict the RUL until the stop point is repeated. The Data Challenge provides two scoring methods for prediction tasks, as in Tables 5 and 6:

Table 5. Scoring method 1.

Ground Truth RUL (GT)	Submission RUL (SUB)	Score
Number	Number	$ GT - SUB  \times \exp(-0.001GT)$
NaN	Number	$ SUB  \times \exp(-0.001SUB)$
Number	NaN	$ GT  \times \exp(-0.001GT)$
NaN	NaN	0

Table 6. Scoring Method 2.

Ground Truth RUL (GT)	Submission RUL (SUB)	Score
Number	Number	$0.1 \times (GT - SUB)^2$
NaN	Number	$5 / ( SUB  + 3)$
Number	NaN	$20 \times \exp[-1 / ( GT  + 0.1)]$
NaN	NaN	0

The sum of the two is the total score, and the smaller score means better. To visually demonstrate the effect of the method of this subject, a single-layer simple recurrent neural network, a double-layer simple recurrent neural network, a single-layer long-term memory network, and a double-layer long-term memory network are added to the experiment to conduct prediction experiments after training on the same data set; according to the scoring rules, the effects of the method of this subject and the comparison method can be obtained as shown in Table 7 below:

**Table 7.** Comparison of prediction results with common deep learning methods.

Method	Total Score	Mean Absolute Error (Unit: Second)
Single-layer RNN	$5.90 \times 10^{15}$	6843
Double-layer RNN	$6.22 \times 10^{15}$	7632
Single-layer LSTM	$5.96 \times 10^{15}$	6500
Double-layer LSTM	$5.67 \times 10^{15}$	6753
Proposed Method	$1.77 \times 10^{15}$	2240

As shown in the table, the total score of the ordinary single-layer long-term and short-term memory network is much higher than the method used in this subject, which proves that the method proposed in this subject has a good effect on predicting the RUL of the multi-dimensional time series. It is worth noting that the average absolute error in the prediction of the method used in this topic is much lower than that of other methods. It can be seen that the prediction results of the method in this topic have better accuracy. In addition, during the experiment, the training time of the contrast method used was more than one hour, while the training time of the method used in this topic was only about half an hour, and the computational efficiency was greatly enhanced.

## 5. Conclusions

In this paper, two models of random forest and transformer are constructed to solve the task of predicting the RUL of the etching machine. The random forest model is used to detect the time point of abnormal operation of the etching machine during operation, and the Transformer model is used to predict the abnormal operation stage until shutdown.

In the experiment, the random forest shows ultimate efficiency, and when we use it to identify the health state of the machine, it is very fast. Additionally, the transformer method is obviously better than the compared methods. The effect of the Transformer would be significantly improved when the rise of the parameters of the Transformer method, but the calculated amount of it also increases rapidly, which means the training needs more time or better devices. Therefore, the method would be labored when it faces tasks with a high dimension and long-length data, and we would reform the method based on the short-length data in the future.

From the experimental results, both models have achieved the preset functions, which proves the feasibility of the prediction method of this subject, but it is worth reflecting that in this method, the three types of failures are separately trained when predicting the RUL. That is, it is assumed that the three are independently affecting the downtime of the etching machine. However, in practice, it is usually not the case. Therefore, in follow-up research, the RUL labels of the three faults can be combined for learning, and the relationship between faults and the RUL can be further studied.

In summary, the method proposed in this topic achieves a good effect on predicting the RUL of the etching machine, which is well suited for practical applications. For instance, the proposed method is promising for manufacturing the electronic chips of the sport devices for training athletes and evaluating their performance. It can be also used for other areas such as the aerospace industry, automation, etc.

**Author Contributions:** Formal analysis, L.Z.; Methodology, Y.Z.; Project administration, T.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study was funded by the Key R&D Plan of China for Winter Olympics (2021YFF0306401) and the Key Special Project of the National Key Research and Development Program “Technical Winter Olympics” (2018YFF0300502 and 2021YFF0306400).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data is contained within the article.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Smith, W.A.; Randall, R.B. Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. *Mech. Syst. Signal Process.* **2015**, *64*, 100–131. [[CrossRef](#)]
2. Zhang, W.; Li, X.; Li, X. Deep learning-based prognostic approach for lithium-ion batteries with adaptive time-series prediction and on-line validation. *Measurement* **2020**, *164*, 108052. [[CrossRef](#)]
3. Li, X.; Li, X.; Ma, H. Deep representation clustering-based fault diagnosis method with unsupervised data applied to rotating machinery. *Mech. Syst. Signal Process.* **2020**, *143*, 106825. [[CrossRef](#)]
4. Li, X.; Fu, X.-M.; Xiong, F.-R.; Bai, X.-M. Deep learning-based unsupervised representation clustering methodology for automatic nuclear reactor operating transient identification. *Knowl.-Based Syst.* **2020**, *204*, 106178. [[CrossRef](#)]
5. Zhao, T.; Li, K.; Ma, H. Study on dynamic characteristics of a rotating cylindrical shell with uncertain parameters. *Anal. Math. Phys.* **2022**, *12*, 97. [[CrossRef](#)]
6. Zhao, T.Y.; Yan, K.; Li, H.W.; Wang, X. Study on theoretical modeling and vibration performance of an assembled cylindrical shell-plate structure with whirl motion. *Appl. Math. Model.* **2022**, *110*, 618–632. [[CrossRef](#)]
7. Li, X.; Zhang, W.; Ma, H.; Luo, Z.; Li, X. Domain generalization in rotating machinery fault diagnostics using deep neural networks. *Neurocomputing* **2020**, *403*, 409–420. [[CrossRef](#)]
8. Li, X.; Zhang, W.; Ma, H.; Luo, Z.; Li, X. Data alignments in machinery remaining useful life prediction using deep adversarial neural networks. *Knowl.-Based Syst.* **2020**, *197*, 105843. [[CrossRef](#)]
9. Li, X.; Zhang, W.; Ma, H.; Luo, Z.; Li, X. Partial transfer learning in machinery cross-domain fault diagnostics using class-weighted adversarial networks. *Neural Netw.* **2020**, *129*, 313–322. [[CrossRef](#)]
10. Zhang, W.; Li, X.; Ma, H.; Luo, Z.; Li, X. Transfer learning using deep representation regularization in remaining useful life prediction across operating conditions. *Reliab. Eng. Syst. Saf.* **2021**, *211*, 107556. [[CrossRef](#)]
11. Li, X.; Zhang, W.; Ma, H.; Luo, Z.; Li, X. Degradation alignment in remaining useful life prediction using deep cycle-consistent learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**. [[CrossRef](#)] [[PubMed](#)]
12. Hsu, C.-S.; Jiang, J.-R. Remaining useful life estimation using long short-term memory deep learning. In Proceedings of the 2018 IEEE International Conference on Applied System Invention, Chiba, Japan, 13–17 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 58–61.
13. Yuan, M.; Wu, Y.; Lin, L. Fault diagnosis and remaining useful life estimation of aero engine using LSTM neural network. In Proceedings of the 2016 IEEE International Conference on Aircraft Utility Systems, Beijing, China, 10–12 October 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 135–140.
14. Zhang, J.; Wang, P.; Yan, R.; Gao, R.X. Long short-term memory for machine remaining life prediction. *J. Manuf. Syst.* **2018**, *48*, 78–86. [[CrossRef](#)]
15. Malhotra, P.; Vig, L.; Shroff, G.; Agarwal, P. Long short-term memory networks for anomaly detection in time series. *Proceedings* **2015**, *89*, 89–94.
16. Zhang, G.; Bai, X.; Wang, Y. Short-time multi-energy load forecasting method based on CNN-Seq2Seq model with attention mechanism. *Mach. Learn. Appl.* **2021**, *5*, 100064. [[CrossRef](#)]
17. Wang, X.; Tang, M.; Yang, T.; Wang, Z. A novel network with multiple attention mechanisms for aspect-level sentiment analysis. *Knowl.-Based Syst.* **2021**, *227*, 107196. [[CrossRef](#)]
18. Wang, Y.; Huang, M.; Zhao, L.; Zhu, X. Attention-based LSTM for aspect-level sentiment classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 606–615.
19. Li, Z.H.; Zhang, Y.; Abu-Siada, A.; Chen, X.; Li, Z.; Xu, Y.; Zhang, L.; Tong, Y. Fault diagnosis of transformer windings based on decision tree and fully connected neural network. *Energies* **2021**, *14*, 1531. [[CrossRef](#)]
20. Yun, H.; Kang, T.; Jung, K. Analyzing and controlling inter-head diversity in multi-head attention. *Appl. Sci.* **2021**, *11*, 1548. [[CrossRef](#)]
21. Savini, E.; Caragea, C. Intermediate-task transfer learning with BERT for sarcasm detection. *Mathematics* **2022**, *10*, 844. [[CrossRef](#)]
22. Wang, N.; Fan, X.; Fan, J.; Yan, C. Random forest winter wheat extraction algorithm based on spatial features of neighborhood samples. *Mathematics* **2022**, *10*, 2206. [[CrossRef](#)]
23. Kovalnogov, V.; Fedorov, R.; Klyachkin, V.; Generalov, D.; Kuvayskova, Y.; Busygin, S. Applying the random forest method to improve burner efficiency. *Mathematics* **2022**, *10*, 2143. [[CrossRef](#)]
24. Pal, M. Random forest classifier for remote sensing classification. *Int. J. Remote Sens.* **2005**, *26*, 217–222. [[CrossRef](#)]
25. Shinkevich, A.I.; Ershova, I.G.; Galimulina, F.F.; Yarlychenko, A.A. Innovative mesosystems algorithm for sustainable development priority areas identification in industry based on decision trees construction. *Mathematics* **2021**, *9*, 3055. [[CrossRef](#)]
26. Al Hamad, M.; Zeki, A.M. Accuracy vs. cost in decision trees: A survey. In Proceedings of the 2018 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies, Sakhier, Bahrain, 18–20 November 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–4.
27. Zhang, W.; Li, X.; Ma, H.; Luo, Z.; Li, X. Universal domain adaptation in fault diagnostics with hybrid weighted deep adversarial learning. *IEEE Trans. Ind. Inform.* **2021**, *17*, 7957–7967. [[CrossRef](#)]

28. Zhang, W.; Li, X. Data privacy preserving federated transfer learning in machinery fault diagnostics using prior distributions. *Struct. Health Monit.* **2022**, *21*, 1329–1344. [[CrossRef](#)]
29. Zhang, W.; Li, X. Federated transfer learning for intelligent fault diagnostics using deep adversarial networks with data privacy. *IEEE/ASME Trans. Mechatron.* **2021**, *27*, 430–439. [[CrossRef](#)]
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–7 December 2017; Volume 30.
31. Dai, Z.; Yang, Z.; Yang, Y.; Cohen, W.W.; Carbonell, J.; Le Quoc, V.; Salakhutdinov, R. Transformer-XL: Language modeling with longer-term dependency. In Proceedings of the ICLR 2019, New Orleans, LA, USA, 6–9 May 2019.
32. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.; Salakhutdinov, R. Transformer-XL: Attentive language models beyond a fixed-length context. *arXiv* **2019**, arXiv:1901.02860.
33. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le Quoc, V. XLNet: Generalized autoregressive pretraining for language understanding. In Proceedings of the NeurIPS 2019, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
34. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. ViViT: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2021, Montreal, QC, Canada, 10–17 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 6836–6846.
35. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2021, Montreal, QC, Canada, 10–17 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 10012–10022.
36. Sun, Z.; Liu, C.; Qu, H.; Xie, G. A novel effective vehicle detection method based on Swin Transformer in hazy scenes. *Mathematics* **2022**, *10*, 2199. [[CrossRef](#)]
37. Ju, X.; Zhao, X.; Qian, S. TransMF: Transformer-based multi-scale fusion model for crack detection. *Mathematics* **2022**, *10*, 2354. [[CrossRef](#)]
38. Qi, J.; Du, J.; Siniscalchi, S.M.; Ma, X.; Lee, C.-H. Analyzing upper bounds on mean absolute errors for deep neural network-based vector-to-vector regression. *IEEE Trans. Signal Process.* **2020**, *68*, 3411–3422. [[CrossRef](#)]