Remaining Useful Life Prediction Using Deep Multi-scale Convolution Neural Networks

Y Ge¹,², J J Liu¹,² and J X Ma¹,²
¹School of Mechanical Engineering, Changshu Institute of Technology, Changshu, PR China
²Jiangsu Key Laboratory for Elevator Intelligent Safety, Changshu Institute of Technology, Changshu, PR China
E-mail:geyang@cslg.edu.cn

Abstract. Accurate Remaining useful life (RUL) prediction is the premise of system prognostics and health management (PHM). In fact, it is often difficult to predict the accurate RUL because of the diversity of working condition and self-state between systems. In this paper, a data-driven method is proposed for RUL prediction using deep multi-scale convolution neural networks (DMSCNN), which is made up of four convolution layers. Two of the four convolution layers are multi-scale convolution layers, which are composed by three different scales convolutions connected by deep concatenation layer. After the four convolution layers, we add fully connected layer and regression layer to build a deep learning structure for predicting RUL. The IEEE PHM 2012 data challenge bearing dataset is employed to verify the proposed method. The experiments result show that the proposed DMSCNN has better performance and higher prediction accuracy than other methods. Additionally, the t-distributed stochastic neighbour embedding (t-SNE) dimension reduction method is used to reduce the computation. Experiments result demonstrate that the features processed by t-SNE have better distinguishability, and the prediction accuracy can be further improved by inputting the t-SNE features into the DMSCNN proposed in this paper for RUL prediction. This paper offers a new and effective approach for RUL prediction.

Key words. Failure mode identification, Remain useful life prediction, Deep Learning, Convolution neural network.

1. Introduction
With the development of deep learning technology, more and more on-line state recognition methods are applied to the systems that demand high reliability, such as aircraft, nuclear power equipment, and so on [1]. Condition-based maintenance (CBM) is a famous strategy of Prognostics and health management (PHM), which uses modern information feature extraction technology and intelligence algorithm for system maintenance or replacement decision. The premise of PHM is to predict the remaining service life (RUL) of the system, so as to carry out maintenance at the right time, reduce the probability of random failure of the system, and improve the reliability of the system [2-3].

Generally, there are three categories of RUL prediction methods, i.e. fault empirical formula methods [4-5], mathematical statistics models [6-7] and data-driven methods [8-9]. The fault
empirical formula methods can explain the nature of system fault, and has high prediction accuracy. However, it is a very complex work to establish an empirical formula, and the working condition requirement of the system is very strict. The prediction method of mathematical statistics model is to use the historical fault data of the system to fit the system life distribution function. Then the fitted distribution function is used to input the parameters of the test system to estimate the RUL of the system. Due to the great influence of RUL on working condition and other factors, the prediction methods based on mathematical statistics model often have large errors. Compared with empirical formula and mathematical statistics methods, data-driven method is a black box method, which does not need to establish complex mathematical formula. It extracts key features from historical data and uses neural networks to learn and predict RUL. With the development of deep learning technology, data-driven method has become a popular method in RUL prediction research [10-11].

In general, feature extraction is the first and very important step in RUL. Common features include time domain, frequency domain, time frequency domain and so on [12]. In practice, extracting valid features is very difficult. Because many characteristics change irregularly during the life cycle of a system. The root mean square (RMS) of the vibration signal is often used to predict RUL. In the literature [13], at the initial stage of equipment degradation, RMS showed an upward trend, but with the deepening of degradation, RMS showed a downward trend. This indicates that RMS cannot directly reflect the degree of equipment degradation. In addition to RMS, many literatures show that many time, frequency, and time-frequency characteristics are also used in RUL prediction [14-15]. Using only one feature for RUL prediction may lead to information loss, and the prediction accuracy is not high. Therefore, researchers use multiple feature fusion technology for RUL prediction. But too many features can lead to information redundancy and even dimensional disaster. Therefore, many dimension reduction methods have been proposed for system fault pattern recognition and RUL prediction, such as PCA [16], LDA [17], ProPCA [18], JADE [19] and so on.

On the basis of determining the appropriate characteristics, another important work of RUL prediction is to establish training and prediction neural networks. At present, many mature methods have been used for RUL prediction. For example, support vector machine (SVM) [20-22], extreme learning machine (ELM) [22], joint approximate diagonalization of eigen-matrices (JADE) [15], particle filter (PF) [23], genetic programming (GP) [24], etc. The deep learning network can combine the low level features to form the abstract high level features, thus reflecting the degradation trend of the system. Therefore, it is highly favored by scholars. In [25] and [26], a RUL prediction method based on long and short term memory (LSTM) neural network is proposed. A semi-supervised learning model based on ladder network (LN) is proposed for parameter dynamics evaluation and RUL prediction of centrifugal pump [27]. In literature [28], the autoencoder and deep neural network (DNN) are used to predict bearing RUL, which achieved better efficiency than other deep learning networks. Literature [29] uses superposition sparse autoencoder and logic regression to predict RUL of aircraft engine. A deep belief feedforward neural network (DBN-FNN) method for RUL prediction of rotating equipment is proposed in literature [30]. In literature [31], a deep belief network integration (MODBNE) method is proposed, which combines multi-objective evolutionary algorithm with traditional DBN training technology to simultaneously evolve multiple DBNS with precision and diversity.

Feature extraction is the key of deep learning RUL prediction method. For the above methods, feature extraction way and scale are single, which leads to some key information that may be omitted. In addition, if the input data dimension is high, feature extraction will consume a lot of time, and if the dimension is low, the extracted features may lose some information. For these reasons, a deep multi-scale convolution neural networks (DMSCNN) method is proposed in this paper for systems RUL prediction. Four different scales convolution layers are used to extract features, and features of different scales are fused through depth connection layer, which makes the extracted features more comprehensive than single scale neural networks. To reduce the input data dimension, the t-distributed stochastic neighbor embedding (t-SNE) method is employed for original features dimensionality reduction.
The rest of the paper is arranged as follows. The t-SNE algorithm is described in Section 2, and a flowchart is designed for dimensionality reduction using t-SNE. DMSCNN is proposed in Section 3. Datasets of the IEEE 2012 PHM prognostic challenge are used to test the proposed method in Section 4, and Section 5 is conclusion.

2. Data dimension reduction using t-SNE
t-SNE is a nonlinear and popular learning algorithm with deep learning, which is proposed by Laurens van der Maaten and Geoffrey Hinton [32]. It has excellent dimensionality reduction effect on the high dimensional nonlinear data set. t-SNE is an improved version of the SNE algorithm. It adopts the t-distribution of low-dimensional space to effectively solve the data congestion problem of SNE algorithm. The principle is to use the joint probability of high dimensional data points and low dimensional space to represent the similarity between data points. Then the optimal low-dimensional simulation data points are obtained by minimizing the Kullback-Leibler divergence.

Let \( \{x_1, x_2, \ldots, x_N\} \) be a \( N \) dimensions sequence, similarity conditional probabilities between any two points in the higher dimensional space are denoted as \( p_{ij} \) and \( p_{ji} \), see formulas (1) and (2) for details.

\[
p_{ij} = \frac{\exp\left(-\frac{d(x_i, x_j)}{2\sigma_j^2}\right)}{\sum_k \exp\left(-\frac{d(x_i, x_k)}{2\sigma_j^2}\right)} \quad (1)
\]

\[
p_{ji} = \frac{\exp\left(-\frac{d(x_j, x_i)}{2\sigma_i^2}\right)}{\sum_k \exp\left(-\frac{d(x_j, x_k)}{2\sigma_i^2}\right)} \quad (2)
\]

where, \( \sigma_j \) denotes the variance of a gaussian function centred on the data point \( x_j \). Joint distribution of any two data points in a high dimensional space is denoted as \( p_{ij} \), which defined as formula (3).

\[
p_{ij} = \frac{p_{ij} + p_{ji}}{2N} \quad (3)
\]

Let \( \{y_1, y_2, \ldots, y_N\} \) be the simulated data points in the low-dimensional space. Using the t-distribution with a degree of freedom of 1, the joint distribution between simulated data points in low-dimensional space is defined as formula (4).

\[
q_{ij} = \frac{(1 + y_i - y_j^2)^{\frac{1}{2}}}{\sum_k (1 + y_i - y_k^2)^{\frac{1}{2}}} \quad (4)
\]

The simulation accuracy of low-dimensional spatial simulation points corresponding to high-dimensional spatial data points is measured by KL divergence. The simulation accuracy is denoted as \( C \), and defined as formula (5).

\[
C = KL(PQ) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (5)
\]

The smaller the value of \( C \) is, the higher the correctness of the simulation of the low-dimensional space to the high-dimensional space will be. If \( C = 0 \), that means they have the same probability distribution. To maximize the accuracy of the simulation and obtain the minimum KL divergence, gradient descent method is used to optimize KL divergence. The specific formula is as formula (6).

\[
\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij}) \left(y_i - y_j\right) \left(1 + y_i - y_j^2\right)^{\frac{1}{4}} \quad (6)
\]
The result of dimension reduction can be obtained through the above calculation, which is denoted as \( \{y_1, y_2, \ldots, y_N\} \). To improve the effect of dimensionality reduction, the original data can be iterated many times to improve the accuracy of low-dimensional space simulation data. Set \( S \) as the maximum number of iterations, and the dimensionality reduction process of high-dimensional data based on the t-SNE algorithm is shown in Fig. 1.

3. Structure of Deep Multi-scale Convolution Networks

The convolutional neural network (CNN) performs well in solving many problems such as visual recognition, speech recognition and natural language processing. CNN has also been employed in mechanical fault identification and RUL prediction. CNN is mainly used to identify two-dimensional graphics with constant size zoom distortion. The feature extraction layer of CNN learns implicitly from training data, instead of explicitly extracting features. Due to the same neurons equal weights, on the surface of the feature mapping network can parallel learning, which is a convolution network main advantages. The weight sharing of CNN reduces the complexity of the network. Multi-dimensional input vectors can be directly input into the network, thus avoiding the complexity of data reconstruction in feature extraction and multi-classification processes.

As shown in Fig. 2, the typical CNN structure is mainly composed of the input layer, convolution layer, pooling layer, full connection layer, and output layer.

Let \( M' \) be the original input matrix, \( X_i \) represents the features of \( i \)’th layer, and \( X_i \) can be
obtained from formula (7).

\[ X_i = R(X_{i-1} \odot W_i + b_i) \]  

(7)

where \( W_i \) represents the convolution kernel weight of \( i \)-th layer, \( \odot \) is the convolution operation symbol. After the convolution output is added with the bias \( b_i \) of the \( i \)-th layer, the features \( X_i \) of the \( i \)-th layer can be obtained by entering the result into the activation function \( R(x) \).

The pooling layer is usually after the convolution layer. Its main functions are: (1) reducing the dimension of features matrix, combining low-level local features to higher-level features, (2) compressing the features matrix, simplifying the network and reducing the computational complexity. Let \( X_i \) represent the \( i \)-th pooling layer, it can be expressed as formula (8).

\[ X_i = P(X_{i-1}) \]  

(8)

where \( P(x) \) represents pooling operation, and common pooling operations include average pooling and maximum pooling. The full connection layer is connected in multiple alternate convolutional layers and pooling layers, and the learned depth features are mapped to the sample label space.

A deep neural network can obtain the representation information from the original input signal adaptively through multiple nonlinear transformations and approximate complex nonlinear functions, which is the main structure of this study. Generally, the common deep learning method includes two substructures, namely, multiple convolutional neural networks and a fully connected regression layer. In this paper, a DMSCNN method is proposed for systems RUL prediction, its architecture is shown in Fig. 3 and Fig. 4.

1. Input layer. To facilitate convolution operation, input data samples are prepared in a two-dimensional format. The size of input data is \( N_t \times N_f \), where \( N_t \) denotes time series and \( N_f \) denotes the number of original features. Original features are often extracted from the vibration signal of equipment, and the detailed data are given in section 4.1.

2. Convolution layer. A total of four convolution layers are stacked in the network for feature extraction. The first and third convolution layers are multi-scale convolution layers, its specific structure is shown in Fig. 5. Three convolution kernels of different scales are used to extract different features, and the sizes of filters are \( 1 \times 1, 3 \times 1, 5 \times 1 \) respectively. The second and fourth convolution layers are single scale convolution layers, and the size of filter is \( 4 \times 1 \). All convolution layers use the zero-point padding operation to keep the dimension of feature mapping unchanged, which is the same as the dimension of the original features of the input. In particular, the multi-layer convolutional neural network is mostly used for image recognition, and the two-dimensional filter is generally adopted. But this paper uses one-dimensional filters because there is no obvious relationship between features that are compressed by t-SNE. Relu activation function is employed for all convolution layers in this paper. Relu function can increase the nonlinearity of the convolutional neural network. In addition, the batch normalization (BN) method is employed in this paper to optimize the convolution layer. BN algorithm is an optimization method proposed by Google company, to join in the convolution layer BN vector algorithm can improve the network. After two single-scale convolution layers, the average pooling layer is used to accelerate convergence and improve the prediction accuracy.

3. Fully-connected layer. Each node of the fully connected layer is connected to all nodes of the previous layer, integrating the features extracted from the previous layer. The dropout layer is used on the last features map to eliminate overfitting. Adam optimization algorithm is used to speed up learning speed. Finally, a regression layer is added to the end of the proposed network, to calculate the semi-mean square error loss of the regression.
4. Experiment and Analysis

4.1. Feature Extraction

In this experiment, the bearing vibration signal is the research object. Firstly, time domain and frequency domain features are extracted from bearing vibration signals, and then, t-SNE is used to reduce the dimensionality of the extracted features. A total of 8 time-domain features and 8 frequency domain features are extracted from the vibration signal, as shown in Table 1.

The data from Case Western Reserve University Bearing Data Center is used to test the effect of t-SNE. The motor load is 2 HP, the motor speed is 1750r/min, and the sampling frequency is 12kHz. A total of 10 kinds of state data, such as in Table 2.
Table 1 The time domain and frequency domain features

| Time domain features | Frequency domain features |
|----------------------|--------------------------|
| $F_i = \frac{1}{N} \sum_{i=1}^{N} x_i$ | $F_i = \frac{1}{N} \sum_{i=1}^{N} \sqrt{x_i}$ |
| $F_2 = \frac{1}{N} \sum_{i=1}^{N} \sqrt{x_i}$ | $F_2 = \frac{1}{N} \sum_{i=1}^{N} |x_i|$ |
| $F_3 = \frac{1}{N} \sum_{i=1}^{N} \frac{x_i}{\max_i} (\frac{1}{N} \sum_{i=1}^{N} x_i)$ | $F_3 = \frac{1}{N} \sum_{i=1}^{N} \frac{x_i}{\max_i} (\frac{1}{N} \sum_{i=1}^{N} x_i)$ |
| $F_4 = \frac{1}{N} \sum_{i=1}^{N} \frac{x_i}{\max_i} (\frac{1}{N} \sum_{i=1}^{N} x_i)^2$ | $F_4 = \frac{1}{N} \sum_{i=1}^{N} \frac{x_i}{\max_i} (\frac{1}{N} \sum_{i=1}^{N} x_i)^2$ |

where $S_i$ is the amplitude of $x_i$ after the Fourier transform, and $f_i$ represents the frequency value after transformation.

Table 2 Classification of Bearing Failure Data

| No. | Failure parts | Failure diameter |
|-----|---------------|------------------|
| data 1 | ball | 0.007'' |
| data 2 | ball | 0.014'' |
| data 3 | ball | 0.021'' |
| data 4 | inner ring | 0.007'' |
| data 5 | inner ring | 0.014'' |
| data 6 | inner ring | 0.021'' |
| data 7 | outer ring | 0.007'' |
| data 8 | outer ring | 0.014'' |
| data 9 | outer ring | 0.021'' |
| data 10 | normal | |

Set the sample size as 1024, 120 sets of data were collected for each type. Sixteen kinds of time domain and frequency domain features are extracted for each set data. Then, the extracted features are reduced to a 3-dimensional vector using t-SNE. To facilitate comparison, PCA, LDA, ProbPCA, SNE and JADE six methods are used for dimensionality reduction concurrently. The dimensionality reduction results are shown in Fig. 6.
Fig. 6. Bearing features dimension reduction results

That can be seen from Fig. 6, among the six dimensionality reduction methods, t-SNE has the best feature aggregation and the most obvious feature differentiation. For more, 60 groups of training data were randomly selected from each class of feature vector after dimensionality reduction, and the remaining 60 groups were used as test data. We select 10 groups of training and test data and use SVM for training and recognition. The recognition results are shown in Fig. 5. It is obvious that the highest recognition accuracy is the features which are reduced by t-SNE.

Fig. 8 shows the recognition results of 16 original features and t-SNE dimension reduction features. We select 5 groups of training and test data from original features and t-SNE dimension reduction features respectively, and use SVM for training and recognition. The data selection method is the same as Fig. 7. It can be seen from Fig. 8, the recognition accuracy is about 80% when using the 16 original features. But the recognition accuracy is almost 100% when using the features reduced by t-SNE. This indicates that t-SNE can retain valid information in the original features and filter out redundant and interference information.

4.2. RUL Prediction
Experimental Data Introduction: Datasets of the IEEE 2012 PHM prognostic challenge are used in this paper, which is provided by the FEMTO-ST institute in France. There are three operating conditions in the experiment. The whole life experiment data of 6 bearings are provided for learning and training, and 11 bearings are provided as test data for life prediction, the detail is shown in Table 3. Each set of data includes vibration and temperature sensor data. The vibration signal comes from the horizontal and vertical sensors. In this paper, only the vibration signals are used for RUL prediction, and the signals are collected every 10 seconds.

| Datasets | Working Conditions |
|----------|--------------------|
| Training | Bearing1-1 Bearing2-1 Bearing3-1 |
|          | Bearing1-2 Bearing2-2 Bearing3-2 |
|          | Bearing1-3 Bearing2-3 Bearing3-3 |
| Test     | Bearing1-4 Bearing2-4 |
|          | Bearing1-5 Bearing2-5 |
|          | Bearing1-6 Bearing2-6 |
|          | Bearing1-7 Bearing2-7 |

Data Preprocessing: According to the formula in Table 1, we extract the time domain and frequency domain features from vibration signals of each bearing, and there are 16 features. Besides, the wavelet packet decomposition method is used to decompose the vibration signal into three layers, and the energy of the decomposition coefficient of the third layer packet is used as the time-frequency domain features, whose number is 8. In this way, 24 features can be extracted from each set of vibration signals. Because each bearing has both horizontal and vertical sensors to collect signals. Therefore, a total of 48 original features can be extracted from each bearing. Then, the t-SNE method in section 2 is used to reduce the 48 original features to \( N_f \) dimensions features, the dimension reduction results are denoted as \( \left[ FD_{1}, FD_{2}, \ldots, FD_{N_f} \right] \). In order to adapt to the two-dimensional convolution network, it is necessary to construct a two-dimensional features matrix. In this paper, we divide 8 consecutive samples into a group, which is denoted as a matrix \( X_i \), and \( X_i \) can be expressed as formula (9). The rows of \( X_i \) represent the t-SNE features, and the number is \( N_f \). The columns of \( X_i \) represent the sample size, and the number is 8.

\[
X_i = \begin{bmatrix}
FD_{(i-7)1} & \cdots & FD_{(i-7)N_f} \\
\vdots & \ddots & \vdots \\
FD_{i1} & \cdots & FD_{iN_f}
\end{bmatrix}
\]

(9)

For training data, labels are required. In this paper, the RUL of sample collection is directly used as the training labels. The training labels are denoted as \( Y_{i} \), and \( Y_{i} \) can be calculated by formula (10).

\[
Y_{i} = T_f - It^{*i}
\]

(10)

where \( T_f \) represents the total life of the bearing, \( It \) represents the interval time between the signal collection, and \( It = 10s \).
Feature Extraction (Time Domain, Frequency Domain, Time-Frequency Domain)

Feature Reduction (t-SNE)

Learning Set

Test Set

RUL Labels

Normalization

Deep MSCNN

Trained network

Smoothing Method

Estimating RUL

Fig. 9. RUL prediction process of DMSCNN

Network Training and Prediction: First of all, normalization processing is required for the t-SNE features and labels. The t-SNE features are normalized to (0, 1). The normalization of labels uses the maximum empirical value method. The empirical value of the label upper limit is set as 2810. Each label is normalized by dividing 2810. After normalization, all labels’ values are less than 1. Then, the normalized features and labels are input into the DMSCNN established in section 3 for training, and a detailed flow chart is shown in Fig. 9. The number of filters for each convolution layer is set to 80, the number of iterations is 200. The Piecewise learning rate is adopted, the initial learning rate is 0.001, the learning rate drop factor is 0.1, and the learning rate drop period is 40. The probability of the dropout layer is set to 0.1. Considering that the prediction result of RUL may have large fluctuation, the smoothing method is used for shrinking the error. The prediction results are smoothed with 10 consecutive points, and the estimated value of RUL is finally output.

Prediction results and analysis: According to the data in Table 3, there are 6 training data sets and 11 test data sets. After feature extraction, dimension reduction, training process, test data sets are input to the trained DMSCNN for RUL prediction.

To measure the prediction accuracy, Root mean square error (RMSE) is used as an evaluation basis. The smaller the RMSE is, the higher the prediction accuracy is. The calculation formula of RMSE is as formula (11).

\[
RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} d_i^2}
\]

where \(d_i = RUL_i - RUL_t\), \(RUL_i\) is the \(i\)-th RUL predictive value, and \(RUL_t\) is the real value.

To show the superiority of the proposed method, using the same input data, several mainstream prediction methods are used for RUL prediction, the results are shown in Table 4. The first column
in the table lists four methods, there are DCNN, LSTM, SVM and DMSCNN, and DMSCNN is the method proposed in this paper. The second column of the table shows the RUL RMSE results which are predicted using the 48 original features. The third to ninth columns are the RUL RMSE results which are predicted using dimension reduction features by t-SNE, and the dimensions are reduced to 4 to 16 respectively. Where the structure of the DCNN method is similar to the DMSCNN proposed in this paper, except that the multi-scale convolution layer is replaced by the ordinary convolution layer, as shown in Fig. 10. The LSTM method consists of four LSTM layers and a full connection layer, and 80 hidden cells are set for each LSTM layer.

![Fig. 10. The structure of DCNN](image)

![Fig. 11. RUL predicted results of 8D t-SNE features using DMSCNN](image)

The following two conclusions can be drawn from Table 4.

1. Among the four prediction methods, the accuracy of the t-SNE features prediction is higher than the original features. However, it can also be seen that it is necessary to select the appropriate feature dimension to obtain the best prediction effect. For example, DMSCNN has the best prediction accuracy when the feature dimension is reduced to 8, and the predicted results are shown in Fig. 11.

2. Regardless of any type of features, the DMSCNN method has the highest prediction accuracy among the four methods. This indicates that the prediction network proposed in this paper has a certain superiority.

| Methods       | Original features | 4 D t-SNE features | 6 D t-SNE features | 8 D t-SNE features | 10 D t-SNE features | 12 D t-SNE features | 14 D t-SNE features | 16 D t-SNE features |
|---------------|-------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| SVM [33]      | 0.4365            | 0.3465             | 0.3334             | 0.3209             | 0.3006             | 0.2985             | 0.2773             | 0.2644             |
| LSTM [34]     | 0.3394            | 0.3004             | 0.2863             | 0.2929             | 0.3190             | 0.2829             | 0.3060             | 0.3095             |
| DCNN [28]     | 0.3228            | 0.2792             | 0.2468             | 0.2973             | 0.2512             | 0.2960             | 0.2673             | 0.2343             |
| DMSCNN        | 0.3136            | 0.2762             | 0.2371             | 0.2252             | 0.2347             | 0.2306             | 0.2662             | 0.2331             |

5. Conclusions

In this paper, a deep multi-scale convolution network is proposed for bearing RUL prediction. First, different features are extracted from the original features by using convolution operation with
different scales, and then the extracted features are fused by the deep connection layer as the input of the next convolution layer. Finally, the learning and prediction network is constructed by four convolution layers. Two of them are multi-scale convolution layers, and the other two are ordinary convolution layers. Experimental results show that the proposed network structure has a certain superiority. In addition, the t-SNE dimension reduction method is employed into bearing feature extraction in this paper, which significantly improves the failure recognition accuracy. From the RUL prediction results, the prediction accuracy of features extracted by t-SNE is better than the original features. In order to adapt to the two-dimensional convolution network, this paper also gives the construction method of a two-dimensional features matrix.

Due to the limited training samples used in this paper, only the features of 6 bearings are used as the training samples, which has certain limitations. If there are more training and test samples, it is believed that the prediction accuracy will be better. The DMSCNN proposed in this paper has certain superiority over the popular prediction network and has a certain prospect of promotion in other prediction problems or classification problems.

Of course, some aspects of this research need further improvement. Such as the following deficiencies. Although the t-SNE algorithm has excellent feature extraction ability, it needs a long operation time. The evaluation function of prediction results is relatively simple. Whether the original feature extraction is reasonable, and so on. These problems will be improved in the next study.

References
[1] Rocchetta R, Broggi M, Huchet Q, and Patelli E, “On-line Bayesian model updating for structural health monitoring,” Mechanical Systems and Signal Processing, vol. 103, pp. 174-195, Mar. 2018.
[2] Walter G and Flapper SD, “Condition-based maintenance for complex systems based on current component status and Bayesian updating of component reliability,” Reliability Engineering and System Safety, vol. 168, pp. 227-239. Dec. 2017.
[3] Shin JH and Jun HB, “On condition based maintenance policy,” Journal of Computational Design and Engineering, vol. 2, no. 2, pp. 119-127, Apr. 2015.
[4] Kan MS, Tan ACC, and Mathew J, “A review on prognostic techniques for non-stationary and non-linear rotating systems,” Mechanical Systems and Signal Processing, vol. 62–63, pp. 1–20. Oct. 2015.
[5] Zhang JL; Shang DG; Sun YJ, and Wang XW, “Multiaxial high-cycle fatigue life prediction model based on the critical plane approach considering mean stress effects,” International Journal of Damage Mechanics, vol. 27, no. 1, pp. 32-46. Jan. 2018.
[6] Hu JQ, Zhang LB, Ma L, and Liang, W. “An integrated safety prognosis model for complex system based on dynamic Bayesian network and ant colony algorithm,” Expert Systems with Applications, vol. 38, no. 3, pp. 1431-1446. Mar. 2011.
[7] Yang L and Lee J, “Bayesian belief network-based approach for diagnostics and prognostics of semiconductor manufacturing system,” Robotics and Computer Integrated Manufacturing, vol. 28, no. 1, pp. 66-74. Feb. 2012.
[8] Zhao D. F., Liu S. L., Gu D., Sun X., Wang L., Wei Y., and Zhang H. L., “Enhanced data-driven fault diagnosis for machines with small and unbalanced data based on variational auto-encoder,” Measurement Science and Technology, vol. 31, no. 3, pp. 12, Mar. 2020.
[9] Djeziri MA, Benmoussa S, and Sanchez R, “Hybrid method for remaining useful life prediction in wind turbine systems,” Renewable Energy, vol. 116, pp. 173-187. Feb. 2018.
[10] Sikorska JZ, Hodkiewicz M, and Ma L, “Prognostic modelling options for remaining useful life estimation by industry,” Mechanical Systems and Signal Processing, vol. 25, no. 5, pp. 1803-1836. Jun. 2010.
[11] Si XS, Wang W, Hu CH, and Zhou DH, “Remaining useful life estimation—a review on the statistical data driven approaches,” European Journal of Operational Research, vol. 213, no. 1,
pp. 1-14, Jun. 2011.

[12] Kumar T. P., Saimurugan M., Haran R. B. H., Siddharth S., and Ramachandran K. I., “A multi-sensor information fusion for fault diagnosis of a gearbox utilizing discrete wavelet features,” Measurement Science and Technology, vol. 30, no. 8, pp. 11, Aug. 2019.

[13] Liao L, “Discovering prognostic features using genetic programming in remaining useful life prediction,” IEEE Transactions on Industrial Electronics, vol. 61, pp. 2464-472, May 2014.

[14] Scanlon P., Kavanagh D. F., and Boland F. M., “Residual Life Prediction of Rotating Machines Using Acoustic Noise Signals,” IEEE Transactions on Instrumentation and Measurement, vol. 62, no. 1, pp. 95-108, Jan. 2013.

[15] Liu F, Liu YB, Chen FL, and He B, “Residual life prediction for ball bearings based on joint approximate diagonalization of eigen matrices and extreme learning machine,” Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, vol. 231, no. 9, pp. 1699-1711, May 2017.

[16] Ou L and Yu DJ, “Rolling Bearing Fault Diagnosis Based on Supervised Laplacian Score and Principal Component Analysis,” Journal of Mechanical Engineering, (in Chinese), vol. 50, no. 5, pp. 88-94, Jul. 2014.

[17] Park WJ, Lee SH, Joo WK, and Song JI, “A mixed algorithm of PCA and LDA for fault diagnosis of induction motor,” in Proceedings of Third International Conference on Intelligent Computing, 2007, pp. 934–942.

[18] Bellas A, Bouveyron C, Cottrell M, and Lacaille J, “Model-based clustering of high-dimensional data streams with online mixture of probabilistic PCA”, Advances in Data Analysis and Classification, vol.7, no. 3, pp. 281-300, Sep. 2013.

[19] Liu YB, He B, Liu F, Lu SL, and Zhao YL, “Feature fusion using kernel joint approximate diagonalization of eigen-matrices for rolling bearing fault identification,” Journal of Sound and Vibration, vol. 385, pp. 389-401, Dec. 2016.

[20] Mary J. S., Balaji M. A. S., Krishnakumari A., Nakandhrakumar R. S., and Dinakaran D., “Monitoring of drill runout using Least Square Support Vector Machine classifier,” Measurement, vol. 146, pp. 24-34, Nov. 2019.

[21] Sun C, Zhang Z, and He Z, “Research on bearing life prediction based on support vector machine and its application,” Journal of Physics: Conference Series, vol. 305, pp. 012028, Jul. 2011.

[22] Fu A, Wang XZ, He YL, and Wang LS, “A study on residence error of training an extreme learning machine and its application to evolutionary algorithms,” Neurocomputing, vol. 146, pp. 75-82, Dec. 2014.

[23] Qian YN and Yan RQ, “Remaining useful life prediction of rolling bearings using an enhanced particle filter,” IEEE Transactions on Instrumentation and Measurement, vol. 64, no. 10, pp.2696-2707, Oct. 2015.

[24] Chen C., Xu T. H., Wang G., and Li B., "Railway turnout system RUL prediction based on feature fusion and genetic programming," Measurement, 2020, p. 9.

[25] Mao W. T, He J L, Tang J M, and Li Y, "Predicting remaining useful life of rolling bearings based on deep feature representation and long short-term memory neural network," Advances in Mechanical Engineering, vol. 10, no. 12, pp. 1-18, Dec. 2018.

[26] Wu YT, Yuan M, Dong SP, Lin L, and Liu YQ, “Remaining useful life estimation of engineered systems using vanilla LSTM neural networks,” Neurocomputing, vol. 275, pp. 167-179, Jan. 2018.

[27] He R, Dai YY, Lu JC, and Mou CL, “Developing ladder network for intelligent evaluation system: Case of remaining useful life prediction for centrifugal pumps,” Reliability Engineering & System Safety, vol. 180, pp. 385-393, Dec. 2018.

[28] Ren L, Sun YQ, Cui J, and Zhang L, “Bearing remaining useful life prediction based on deep autoencoder and deep neural networks,” Journal of Manufacturing Systems, vol. 48, pp. 71-77, Jul. 2018.

[29] Ma J, Su H, Zhao WL, and Liu B, “Predicting the Remaining Useful Life of an Aircraft Engine
Using a Stacked Sparse Autoencoder with Multilayer Self-Learning,” *Complexity*, 2018.

[30] Deutsch J and He D, “Using Deep Learning-Based Approach to Predict Remaining Useful Life of Rotating Components,” *IEEE Transactions on Systems Man Cybernetics*–*systems*, vol. 48, no. 1, pp. 11-20, Jan. 2018.

[31] Zhang C, Lim P, Qin A K, and Tan K C, “Multi-objective Deep Belief Networks Ensemble for Remaining Useful Life Estimation in Prognostics,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28 no. 10, pp. 2306-2318, Oct. 2017.

[32] Laurens V D M, Hinton G, Hinton V D M G, “Visualizing Data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, pp.2579-2605, Nov. 2008.

[33] Louen C, Ding SX, and Kandler C, “A new framework for remaining useful life estimation using support vector machine classifier,” *In: Proceedings of IEEE Conference on Control and Fault-Tolerant Systems*; pp. 228–233, 2013.

[34] Malhotra P., TV V., Ramakrishnan A., Anand G., Vig L., Agarwal P., et al. Multi-sensor prognostics using an unsupervised health index based on LSTM encoder-decoder. *arXiv preprint arXiv:160806154* 2016.