Multi-Task Learning and Attention Mechanism Based Long Short-Term Memory for Temperature Prediction of EMU Bearing

Yaohua Chen; Chun Zhang; Ning Zhang; Yiting Chen; Huan Wang
Engineering Research Center of Network Management Technology for High Speed Railway Ministry of Education
Beijing Jiaotong University
Beijing, China
17120349@bjtu.edu.cn

***Abstract*—The traction motor is one of the key components that plays an important role in ensuring the safety and stability of the running EMU (Electric Multiple Units). The running state of the traction motor can be determined through monitoring and predicting the change of EMU bearing temperature. In this paper, we propose a Long Short-Term Memory Neural Network based on Multi-task Learning and Attention Mechanism for the bearing temperature prediction in view of the complex influencing factors of bearing temperature in train operation. The model learns the characteristics of temperature sensors in different positions jointly through multi-task learning. And the Long Short-Term Memory Neural Network based on Attention Mechanism is used to consider the influence of current operating conditions and previous train records on bearing temperature in different degrees. So the model takes various influencing factors and spatial-temporal correlation into consideration. The experimental results with actual EMU datasets show that our method outperforms the baseline approaches.**

Keywords-bearing temperature prediction; Long Short-Term Memory; Multi-Task Learning; Attention Mechanism; EMU

I. INTRODUCTION

According to the statistics of International Union of Rail ways (UIC), by April 2017, the number of high-speed Electric Multiple Units (EMU) in China has reached more than 2700, ranking first in the world, which makes the safety and reliability of EMU more concerned.

The traction motor, as one of the key components of EMU, have a significant impact on the safe and efficient operation of EMU.

Monitoring and predicting the change of the traction motor bearing temperature can effectively judge the running status of traction motor, and provide important guidance for improving the operation efficiency and maintenance. For example, if the bearing temperature of traction motor is too high, alarm will be triggered and emergency operation, including running at a slower speed, emergency braking, stopping operation, etc., will be carried out to prevent hot-cutting shaft accident [1, 2].

Recently, there are many methods to predict bearing temperature. However, these methods are difficult to be directly applied to predicting bearing temperature of the EMU traction motor. And it is necessary to consider the complex conditions under real conditions, mainly including the following 3 points.

1) The bearing temperature is affected by various factors, such as the running state of the train (Traction Force, Speed, Acceleration), environmental factors (Outside Temperature), component performance and so on.

2) The change of the bearing temperature is a cumulative process. It is not only related to current influencing factors, but also related to various factors in the previous process. Therefore, there is a temporal correlation among the states of the sensor in different time steps.

3) There are 12 temperature sensors in a traction motor of the same carriage. Under the same external factors, there is a spatial correlation among different sensors. Fig. 1 shows the data of different temperature sensors in the same traction motor. Although the temperature of some different sensors varies greatly, it can be clearly seen that the change trend of the bearing temperature is similar, which shows that there is spatial correlation among these sensors.

Considering the above situation, we propose a deep learning model named MTL-AM-LSTM Neural Network, which combines Multi-Task Learning (MTL) with Long Short-Term Memory (LSTM) [3] based on Attention Mechanism [4] to predict bearing temperatures under the known influencing factors. Specifically, Attention Mechanism is integrated into LSTM, and the temperature data of sensors in different positions of the same traction motor bearing are learnt to construct an bearing temperature prediction model which includes multiple factors of bearing and spatial-temporal relationship. We highlight our contributions as follows.

1) We propose a deep learning framework to predict bearing temperature, which can provide effective guidance for the operation of traction motor, so as to ensure reliable operation and reduce maintenance costs.

Ministry of Industry and Information Technology of the People's Republic of China Program (No. BZYJ2018-03)

978-1-7281-0860-5/19/$31.00 ©2019 IEEE

Figure 1. Data tendency and relationship

2) LSTM Neural Network is utilized to solve the problem of multi-factor influence and time correlation of the bearing temperature prediction. Meanwhile, attention focus in time series is considered by combining Attention Mechanism.

3) We exploit MTL to learn the bearing temperature characteristics of different positions under the same conditions, thus solving the problem of spatial correlation among different sensors. Besides, other tasks can provide additional evidence for the correlation and irrelevance of features, and reduce the impact of missing data and noise data and the risk of over-fitting.

The paper is organized as follows. Section II describes the related works. Section III describes the introduction of relevant models and the proposed model. Section IV, we implement the experiments and evaluate the results. Section V follows with the conclusion of the paper.

II. RELATED WORKS

A. *Relevant Research on Bearings*

Temperature sensors are usually used for condition monitoring, which have the advantages of easy deployment and low cost. EMU also utilizes temperature sensors to monitor bearings. It has been proved that the temperature raising model is effective in the diagnosis of bearing faults caused by surface contact fatigue [5], position deviation [6], and lubrication condition [7]. So there are many studies on the bearing prediction, which are generally divided into two categories: the statistical method and the artificial intelligence method.

1) The Statistical Method

For stationary data, the traditional statistical method has a high operational efficiency, and the prediction scheme can be completed in a relatively short time. However, these statistical models usually have strict constraints on adaptability assumptions.

J. Cui et al. [8] designed an optimized Autoregressive Moving Average Model (ARMA) based on Genetic Algorithm (GA) to predict the service life of aeroengine. In [9], the Autoregressive Integrated Moving Average Model (ARIMA) was applied to the prediction of temperature data in non-stationary time series of large thrust bearings. Tobonmejia et al. [10] proposed the combination of Mixture of Gaussian distribution and Hidden Markov Models (HMM) for bearing fault prediction. These statistical methods have good results in experiments, but it is usually difficult to satisfy their conditional assumptions and have greater limitations in real situations.

2) The Artificial Intelligence Method

Artificial Intelligence Method can automatically extract learning features from a large number of input and output, and fit the relationship between input and output without manual intervention.

Saidi, Lotfi et al. [11] proposed a vibration-based prediction of high-speed shaft bearings (HSSB) for wind turbines. The lifetime of HSSB was predicted by Support Vector Regression (SVR) using the time domain characteristics of spectral kurtosis (SK). S Kang et al. [12] put forward the method of combining Multiple Criterions Effectiveness Analysis (MCEA), Kernel Principal Component Analysis (KPCA) and SVR to predict the RUL of rolling bearings. These support vector machine methods have good results in small sample data. But there is no suitable method for selecting the kernel function. The experimental results are sensitive to the penalty factor and have weak adaptive ability.

The neural network is also widely used in bearing prediction [13, 14] because it does not need to establish a mathematical model, and it has strong self-learning, robustness, self-adaptive ability and non-linear mapping ability. In [15], L Mo et al. put forward BP neural network based on GA to predict the bearing temperature of rotary control head. C Luo et al. [16] applied LSTM neural network to predict the hot-axle failure of the locomotive bearing. In reference [17], the correlation between bearing temperature and health index is established by using wavelet packet decomposition and Convolutional Neural Networks (CNN).

B. *Relevant Research on Attention Mechanism and MTL*

In recent years, Attention Mechanism has been applied to the neural network for time series prediction. Y. G. Cinar et al. [18] designed a attention-based Recurrent Neural Network (RNN) to predict periodic time series with a large number of missing values. Y Liang, S Ke et al. [19] designed a multi-level attention network, which based on encoder-decoder, to address the dynamic spatial-temporal correlations in geo-sensory time series prediction. In these attention-based models, dynamic temporal or spatial relationship is considered but lacks output of multiple sensors.

For one task, although we can achieve acceptable performance through training and participation, we may neglect the relevance of related tasks. Multi-Task Learning improves the performance of the model by utilizing the relationship between supervised data and the generalization performance of the model on the original task [20]. Hashimoto K et al. [21] exploited a joint multi-task model and defined a joint multi-NLP task network structure to reflect the language hierarchy.

III. CONSTRUCTION OF MODEL

The main task of this paper is to design a model to predict bearing temperature by fitting the non-linear relationship between influencing factors and bearing temperature. At the same time, the temporal correlation of the same bearing and the spatial correlation among different bearings are used to improve the accuracy and generalization ability of prediction. The following mainly introduces the related models and the models proposed in this paper.

A. The Architecture of LSTM Neural Network

Recurrent Neural Network is designed to process time series data, which has good prediction effect and the capability to capture long-term time-dependent and variable-length observations. LSTM is a special kind of recurrent neural network, which performs better than traditional RNN [3, 22] in many tasks. It effectively solves the problem of RNN not being long-term memory and the existence of gradient disappearance or gradient explosion. Therefore, LSTM is utilized as the infrastructure to capture the feature relations between bearing temperature time series and influencing factors, As shown in Fig. 2.

B. The Architecture of Attention-based LSTM Neural Network

Because of the time series relationship of the bearing temperature prediction, the influence of different time steps on the output of the current time is also different and moves with the times. But LSTM only considers it equally, i.e., the weights of H_1 to H_{t-1} are equal. Therefore, we consider using Attention-based LSTM, which can solve the problem of different influence degree of different time steps.

Attention Mechanism has different attention to different parts of the same thing. Strictly speaking, it is an idea rather than a model, so the way to implement it can be utterly different [4, 23]. Attention Mechanism was originally proposed by Bahdanau et al. [24] in the encoder-decoder structure to solve the problem of the long input sequence. Zhou, P, Yin W et al. [25, 26] have further integrated Attention Mechanism into BiLSTM and CNN, and achieved great results in Neuro-Linguistic Programming (NLP).

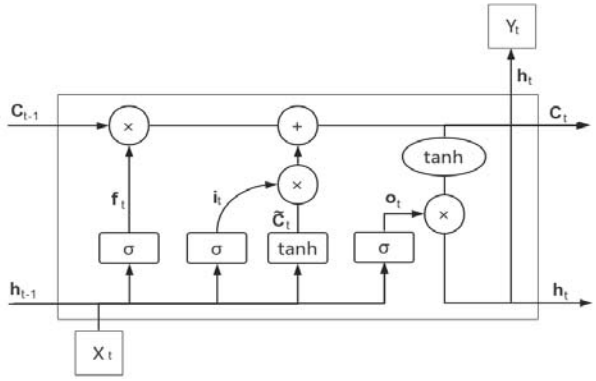

Figure 2. The Architecture of LSTM

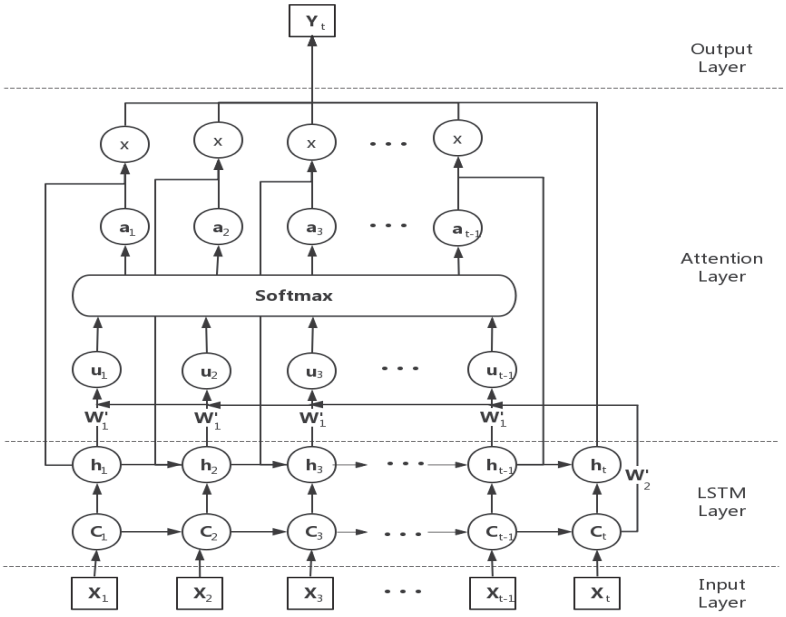

Figure 3. The Architecture of Attention-Based Long Short-Term Memory

This article is not based on Encoder-Decoder, but directly adds Attention Layer to LSTM Neural Network, as shown in Fig. 3. The basic process is to consider the different weights of the first t-1 time series data on the current time data, when processing input for each step. And these information is added to the prediction of this input after mapping weights. Specifically, by matching module, the similarity between the hidden layer state at the first t-1 time and the cell state at time t is calculated. Then we utilize the softmax function to make the sum of all weights 1, i.e. Attention weights of each input. Finally, the sum of all input weighted vectors is calculated as the next input. The equations are shown in (1), (2) and (3).

$$u_i^t = v^T \cdot \tanh(\mathrm{W}_1^{'} h_i + \mathrm{W}_2^{'} c_t) \tag{1}$$

Among them, $\mathrm{W}_1^{'}$ and $\mathrm{W}_2^{'}$ are the learnable parameters of Attention Mechanism, h_i is the hidden layer output of time step i, c_t is the cell state of the current time step (step t), and u_i^t is the similarity between h_i and c_t of time step i.

$$a_i^t = soft\max(u_i^t) \tag{2}$$

where a_i^t is the normalization of u_i^t at each time i, which is the weight in Attention layer.

$$h_t' = \sum_{i=t-n}^{t-1} a_i^t h_i \tag{3}$$

where h_t' is the accumulation of multiplication of the output h_i and Attention weight a_i^t in the first t step.

C. *The Architecture of MTL-AM-LSTM Neural Network*

When predicting bearing temperature faults, the temperature difference of temperature sensors in different positions of bearings is usually considered. Thus, there is a certain correlation among these bearing temperatures, as shown in Fig. 1. But the Attention-based LSTM does not consider the spatial correlation among these bearing temperatures.

Multi-task learning is an inductive transfer mechanism, which trains the original tasks and related tasks. By sharing the characteristics of related tasks, the model can better generalize the original tasks [19, 27]. There are many forms of multi-task learning, including learning to learn, learning with auxiliary task, joint learning and so on.

This paper combines joint learning with the Attention-based LSTM above and proposes a MTL-AM-LSTM Neural Network, as shown in Fig. 4. The model is trained based on the same input characteristics (the input characteristics of different sensors are the same for the same traction motor under the same operating conditions), and the spatial correlation among bearing temperatures is learned by combining the information of bearing temperature sensors at different positions.

In Fig. 4, the input of our model X is the characteristics of bearing temperature, including speed, acceleration, traction, brake feedback energy, outside temperature, etc. The output Y_1, Y_2, Y_3 and Y_4 are bearing temperature at different positions respectively. The model utilizes the hard sharing mechanism of MTL, and all tasks share the hidden layer, while each task retains the special hidden layer and its output layer. It is mainly divided into the following two layers.

$$L_{joint} = \frac{1}{n} \sum_{i=1}^{n} L_i \tag{4}$$

IV. Experiments and Results

In this section, we implement MTL-AM-LSTM Neural Network and experiment on real EMU running data sets. At the same time, the performance of the model is compared with that of the existing methods.

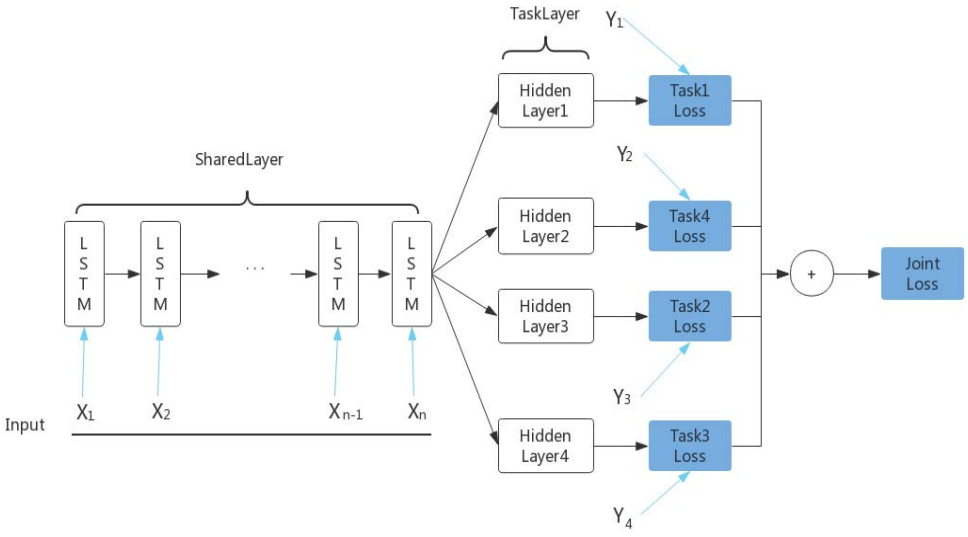

Figure 4. The Architecture of MTL-AM-LSTM

V. EXPERIMENTS AND RESULTS

A. Evaluation Index

This paper uses RMSE and MAE to evaluate the prediction results. The smaller the value, the higher the prediction accuracy and the better the model effect. The formula is as follows.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

B. Datasets

Our experiments utilize time series data monitored from two CRH2 trains during the period from March 2017 to April 2017. In each carriage, 12 J-type thermocouple temperature sensors are used to monitor the running status of the traction motor. Other operational information, such as speed, acceleration, traction and so on, are also recorded. The data set consists of more than 500 thousand rows of records. Some variables are described in Table Ⅰ.

C. Experimental Results

All features are re-sampled at a frequency of 30 seconds in order to ensure the timing interval of data and reduce noise and redundancy. In order to ensure the timing interval of data and reduce noise and redundancy, all features are re-sampled at a frequency of 30 seconds. Then delete the data of abnormal operation state to ensure that the data used is the data of the healthy operation of EMU, such as warehouse stop, maintenance, etc. Finally, standardize the data and zoom it between [0,1]. All models are trained on 70% of the data and validated on 15%, and the remaining 15% is used as the test set.

The input of the model contains 16-dimensional characteristics and the output is 1-dimensional temperature label. We select 4 driving side bearing temperatures for joint training of four tasks. As shown in Table Ⅱ, we finally set batch size to 128, stride size to 1, time step to 90 and the initial learning rate to 0.006 after hyper-parameter tuning. We exploit a single hidden layer and the number of hidden units per layer is 32. and the range of Attention is also 32.

In Fig. 5, the experiment utilizes data for March 2016 to conduct a joint experiment on 4 bearing temperatures. There are 64000 training data and 14000 test data. From the view of RMSE and MAE of each task, good prediction results have been achieved.

As shown in Table Ⅲ, We compare our model with the four baselines and some improved algorithms in this paper. The results are as follows.

1) ARIMA has a strong demand on the stability of time series data, and can not consider the impact of external factors, so the performance of the algorithm prediction is the worst.

2) SVR is not suitable for processing time series data, and it is sensitive to error boundaries and noise, so the experimental prediction is not very optimistic.

3) From the results, it is obvious that LSTM has a good performance. LSTM not only considers the external factors on bearing, but also takes the time dependence into account. However, it does not capture the spatial dependence.

4) ConvLSTM, which extracts features by convolution before LSTM, has better performance than traditional LSTM. But it does not capture the spatial dependence.

5) Attention-Based LSTM has a better effect than traditional LSTM because of adding Attention Mechanism and considering the weight ratio of time series data in different time steps. But the effect is not as good as ConvLSTM.

6) Compared with the prediction results of other models in the experiment, our proposed model has the best performance. Based on Attention-based LSTM, joint learning is exploited to consider the temporal and spatial correlation of the bearing temperature. The experiment combines 4 task learning to discover their common features, reduce the impact of missing data and noise data, and improve the accuracy of prediction.

TABLE I. THE DATA DESCRIPTION

Name	***Description***
train_no	train number
time	time to collect data
speed	train speed (km/h)
out_temp	outside temperature (°C)
traction	traction of motor (kN)
brake_energy	brake feedback energy (kwh)
drive bearing 1	left bearing temperature 1 (°C)
drive bearing 2	left bearing temperature 2 (°C)
drive bearing 3	right bearing temperature 1 (°C)
drive bearing 4	right bearing temperature 2 (°C)
stator1	stator temperature 1 (°C)
stator2	stator temperature 2 (°C)
...	...

TABLE II. RMSE FOR OUR METHOD UNDER DIFFERENT TIME STEPS AND HIDDEN CELL NUMBERS

Time Step \ Hidden Layer Units	**16**	**20**	**30**	**50**	**70**
10	2.73	2.74	2.65	2.92	3.14
20	3.4	3.27	3.23	3.41	3.40
30	3.61	3.12	2.86	3.13	3.44
50	2.94	2.99	2.88	2.64	3.05
70	2.73	2.70	2.78	2.80	2.95
100	2.51	2.44	1.98	2.04	2.17

TABLE III. COMPARISON RESULTS OF DIFFERENT METHODS

Method	RMSE	MAE
ARIMA	9.06	7.88
SVR	7.90	5.68
LSTM	3.85	2.99
ConvLSTM	2.12	2.09
Attention LSTM	2.39	2.16
MTL-AM-LSTM	1.96	1.47

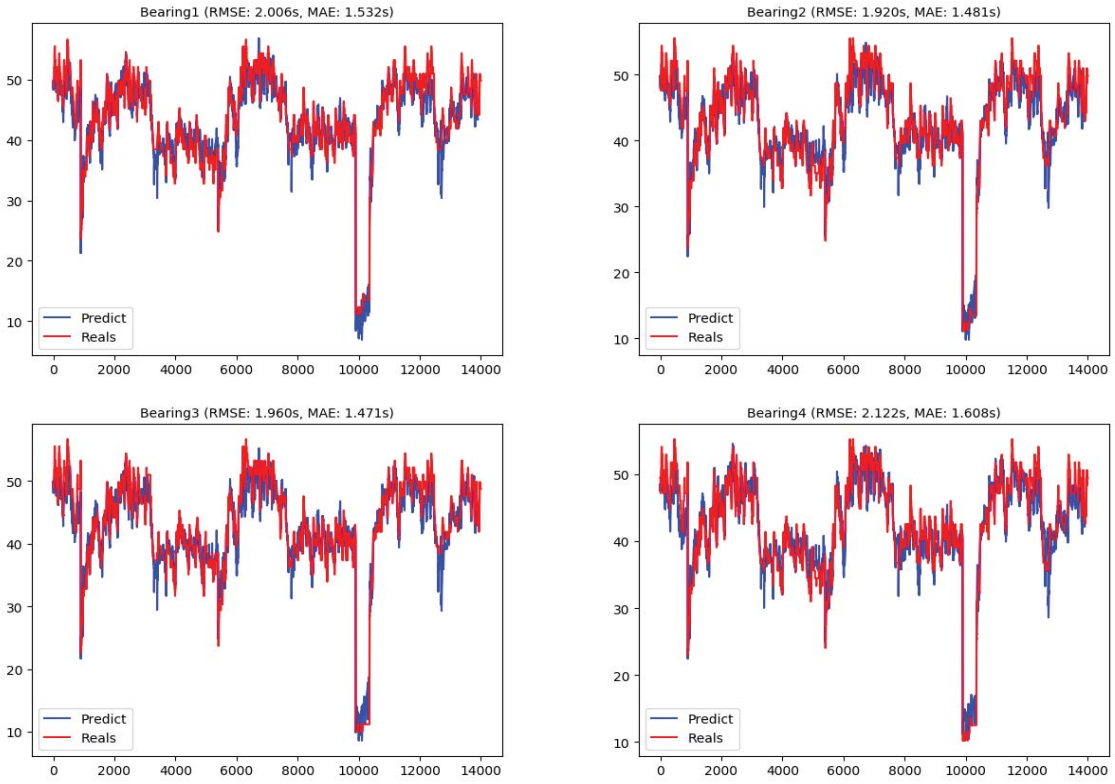

Figure 5. Comparison of forecast and real data for 4 bearing sensors

VI. CONCLUSION

We propose a MTL-AM-LSTM Neural Network to predict the bearing temperature of EMU. Firstly, based on LSTM neural network, the relationship between external factors and the bearing temperature is captured, and the correlation between the bearing temperature and time is obtained. Secondly, LSTM combines Attention mechanism to analyze the different weights of different time steps' information on the current bearing temperature. Finally, the bearing temperature at different positions of the same traction motor bearing is studied jointly, considering the spatial correlation among different sensors, so as to reduce the impact of missing data and noise data and the risk of over-fitting. The experiments show that our model has a good prediction effect and achieves the best of the six models in RMSE and MAE.

This paper only considers short-term prediction (1-2 months). Future research can consider long-term data prediction and periodic variation of bearing temperature. At the same time, feature extraction can be further studied before time series prediction to improve the prediction accuracy.

ACKNOWLEDGMENT

This work is supported by Ministry of Industry and Information Technology of the People's Republic of China Program (No. BZYJ2018-03).

REFERENCES

[1] R. W. Ngigi, C. Pislaru, A. D. Ball and F. Gu, "Modern techniques for condition monitoring of railway vehicle dynamics," 2012.

[2] X. Wang, "Research on Axial Temperature Monitoring and Logic Control of CRH380B," Railway Engineering Cost Management, 2015, vol. 30.

[3] Hochreiter S, Schmidhuber, Jürgen. "Long Short-Term Memory," Neural Computation, 1997, vol. 9(8), pp.1735-1780.

[4] Vaswani, Ashish, et al. "Attention Is All You Need," 2017.

[5] F. Sadeghi, B. Jalalahmadi, T.S. Slack, N. Raje and N. K. Arakere, "A review of rolling contact fatigue," Journal of tribology, 2009, vol. 131, pp. 041403.

[6] O. Tonks, Q. Wang, "The detection of wind turbine shaft misalignment using temperature monitoring," CIRP Journal of Manufacturing Science and Technology, 2017, vol. 17, pp. 71-79.

[7] S. Jiang and H. Mao, "Investigation of the high speed rolling bearing temperature rise with oil-air lubrication," Journal of Tribology, 2011, vol. 133, pp. 021101.

[8] J. Cui, Y. Zhao and S. Dong, "Life Prognostics for Aero-generator Based on Genetic Algorithm and ARMA Model," Acta Aeronautica Et Astronautica Sinica, 2011, vol. 32, pp.1506-1511.

[9] Li-Hui. S and Zhi-Gang. W, "ARIMA Model to Predict Transient Oil Film Temperature of Thrust Bearings," Acta Simulata Systematica Sinica, 2002.

[10] D. A. Tobon-Mejia, K. Medjaher, N. Zerhouni and G. Tripot, "Hidden Markov Models for failure diagnostic and prognostic," 2011 Prognostics and System Health Managment Confernece, Shenzhen, 2011, pp. 1-8.

[11] L. Saidi, J. B. Ali, E. Bechhoefer and M. Benbouzid, "Wind turbine high-speed shaft bearings health prognosis through a spectral Kurtosis-derived indices and SVR," Applied Acoustics, 2017, vol. 120, pp. 1-8.

[12] S. Kang, L. Ye , Y. Wang, J. Xie and V. I. Mikulovich, "Remaining useful life prediction of rolling bearing based on MCEA-KPCA and combined SVR, " Journal of Electronic Measurement and Instrumentation, 2017.

[13] P. Bangalore and L. B. Tjernberg, "An Artificial Neural Network Approach for Early Fault Detection of Gearbox Bearings," in IEEE Transactions on Smart Grid, vol. 6, no. 2, pp. 980-987, March 2015.

[14] A. Zaher, Stephen McArthur, David Infield and Y. Patel, "Online wind turbine fault detection through automated SCADA data analysis," Wind Energy, vol. 12, pp. 574-593, 2009.

[15] L. Mo, J. Wang, L. Wang and L. Y. Wang, "A Rotary Control Head Bearing Temperature Prediction Model Based on GA-BP Algorithm in Underbalanced Drilling," Journal of Southwest Petroleum University (Science & Technology Edition), 2016, pp. 164-169.

[16] C. Luo, D. Yang, J. Huang and Y. D. Deng, " LSTM-Based Temperature Prediction for Hot-Axles of Locomotives," ITM Web of Conferences, vol. 12, pp. 01013, 2017.

[17] D. Belmiloud, T. Benkedjouh, M. Lachi and A. Laggoun, " Deep convolutional neural networks for Bearings failure predictionand temperature correlation," Journal of Vibroengineering 2018, vol. 20.

[18] Y. G. Cinar, H. Mirisaee, P. Goswami, E. Gaussier, A. Ait-Bachir and F. V. Strijov, " Time Series Forecasting using RNNs: an Extended Attention Mechanism to Model Periods and Handle Missing Values," CoRR, 2017.

[19] Y. Liang, S. Ke, J. Zhang, X. Yi and Y. Zheng, "GeoMAN: Multi-level Attention Networks for Geo-sensory Time Series Prediction," IJCAI-18, 2018.

[20] S. Ruder, " An Overview of Multi-Task Learning in Deep Neural Networks," 2017.

[21] Hochreiter S, Schmidhuber, Jürgen. "Long Short-Term Memory". Neural Computation, 1997, vol. 9(8), pp.1735-1780.

[22] Ian Goodfellow, Yoshua Bengio and Aaron Courville, " Deep Learning," Cambridge, MIT Press, 2016, pp. 227-253.

[23] Y. Zhang, P. Zhang and Y. Yan, " Long short-term memory with attention and multitask learning for distant speech recognition," Journal of Tsinghua University (Science and Technology), 2018, vol.58, pp. 249-253.

[24] D. Bahdanau, K. Cho, Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," Computer Science, 2014.

[25] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, et al. "Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification," Proc. 54th Annu. Meet. Assoc. Comput. Linguist, vol. 2, pp. 207‐212, 2016.

[26] W. Yin, H. Schütz, B. Xiang and B. Zhou, "Abcnn: Attention-based convolutional neural network for modeling sentence pairs," Computer Science, 2015.

[27] R. Caruana, "Multitask learning: A knowledge-based source of inductive bias," Machine Learning, vol. 28, pp. 41-75, 1997