Data Augmentation to Improve the Performance of Ensemble Learning for System Failure Prediction with Limited Observations

Guo Shi[1], Bin Liu[1], Lesley Walls[1]
[1]Department of Management, University of Strathclyde, Glasgow, United Kingdom
(guo.shi@strath.ac.uk)

Abstract – **Ensemble learning has been widely used to improve the performance and robustness of machine learning algorithms on time series data. However, in real operational processes where the observed data is limited, it hinders the capability of ensemble learning algorithms. To address the challenge of limited observed data, this paper proposes a novel three-layer ensemble learning framework by use of data augmentation. Firstly, multiple classical time series augmentation methods are applied to increase the size of the data set. Subsequently, after pre-processing, these augmented data is trained by multiple basic learners with K-fold cross-validation as the first layer of the developed ensemble learning framework. The outputs of the first layer are integrated via LASSO to further improve the prediction performance, which serves as the second layer of the developed framework. Finally, the third-layer output is generated by averaging the prediction of the second layer and the output from an improved Long-Short Term Memory model that provides prediction based on the augmented data. A case study on a real wastewater treatment plant is used to illustrate the effectiveness of the proposed method.**

Keywords – **Ensemble learning, data augmentation, failure prediction, time series data, data reweight**

I. INTRODUCTION

As one of the key issues in the research field of Prognostic and Health Management, the prediction of health indicators aims to alert the system faults in advance and assist in maintenance decision-making [1]. During the operational process, the health indicators are collected in the form of time series to reflect the degradation level. Model-based, physical-based and data-driven are three main methods developed in the research field of failure prediction with time series data [2]. Recently, the application of data-driven methods has attracted an increasing attention due to its ability to provide accurate predictions without exploring the mechanism of complex systems [3].

In an early stage, statistical models including the Autoregressive Integrated Moving model (ARIMA) and Exponential Smoothing have gained popularity in the time series prediction [4]. However, these methods have limited capabilities as they fail to capture the feature of high dimension [5]. Machine learning methods such as Support Vector Machine (SVM), Decision Tree and Long-Short Term Memory model (LSTM) have shown high performance in representing latent features and improving prediction accuracy [6].

To improve the applicability of machine learning algorithms, ensemble learning has been proposed to integrate multiple machine learning models in different structures, including bagging, boosting and stacking methods [7-8]. Nevertheless, in reality, there are cases that few inspections/observational data are attainable in industrial engineering systems, which is likely to cause overfitting in machine learning algorithms, and therefore degrade the performance of ensemble learning [9].

In less data-abundant settings, data augmentation methods are useful to generate artificial data sets and avoid overfitting. It can be witnessed from the literature that most of the research focuses on training a specific machine learning algorithm with the augmented data [10-11]. However, to the best of our knowledge, there is no work on enhancing the performance of ensemble learning framework with limited data for system failure prediction.

Considering the insufficiency of training data, this paper proposes a novel framework to improve the prediction of health indicators based on a three-layer ensemble learning framework. The main contributions are summarized as follows:

1) Multi data augmentation methods are adopted to generate artificial data.
2) A three-layer ensemble learning framework is developed to effectively integrate the augmented data set and improve the prediction performance with limited real data.
3) A case study based on a real Wastewater Treatment Plant (WWTP) is conducted to verify the effectiveness of the proposed method.

II. THE PROPOSED ENSEMBLE LEARNING FRAMEWORK

A. Overview of the proposed method

We propose a three-layer ensemble framework integrating multi data augmentation methods to improve the failure prediction accuracy. The first step is generating augmented data and then pre-processing these artificial data by normalization and moving window methods. Subsequently, these pre-processed data will be channeled into the three-layer ensemble learning framework. The first two layers in the ensemble learning framework aim to stack classical machine learning algorithms. The third layer averages the stacking outputs and estimations from an improved LSTM model considering data reweight process. The detailed framework is given in Fig. 1.

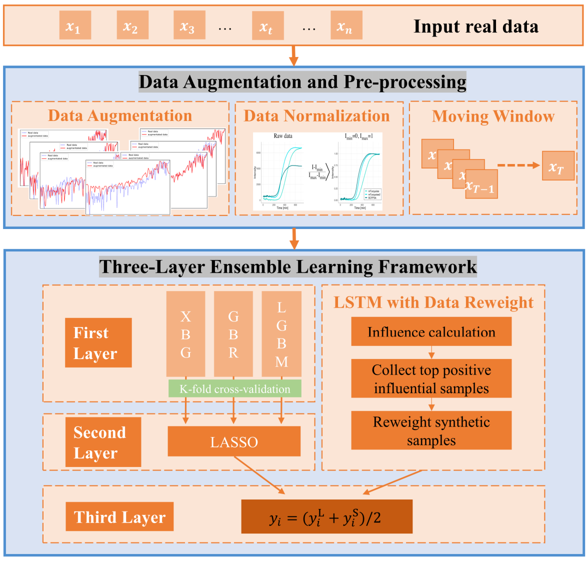


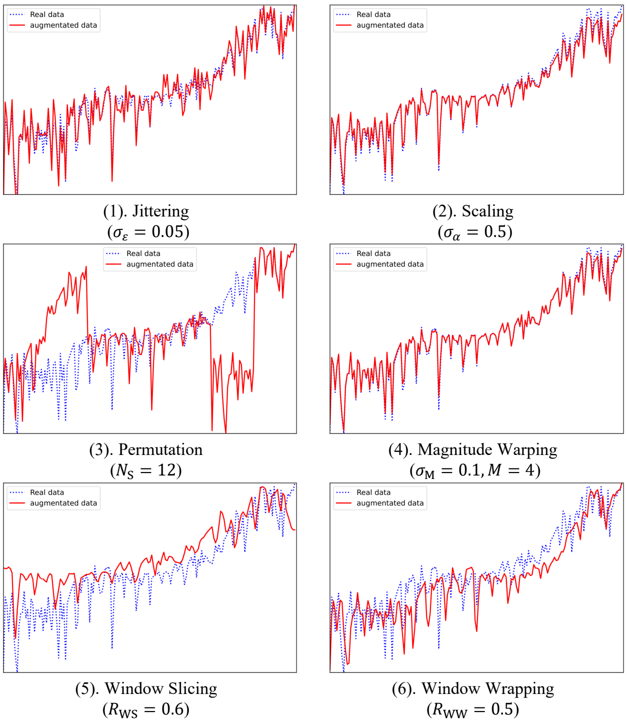
Fig. 1. The overall structure of the proposed method.

B. *Data augmentation and pre-processing*

In order to overcome the overfitting caused by insufficient data, six classical data augmentation methods are applied in this paper, including jittering, scaling, permutation, magnitude warping, window slicing and window warping [12].

Jittering produces artificial data by adding noise ε_i to the real data, and the noise follows the Gaussian distribution $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. Scaling aims to change the magnitude of time series with a random scalar value α, which follows the Gaussian distribution $\alpha \sim N(0, \sigma_\alpha^2)$. Permutation has been proposed to disrupt the order of segments of time series and the numbers of segments N_{S} is the key variable. Magnitude warping is a specific augmentation method to wrap the time series data by a smoothed curve with knots parameters, which are generated randomly and follow the Gaussian distribution $N(1, \sigma_{\mathrm{M}}^2)$. σ_{M}^2 and the total number of knots M are both the hyperparameters of magnitude warping. Window slicing is similar to cropping in the image data augmentation by slicing time steps, and the size of window which can be decided by the wrap ratio R_{WS}. Window warping origins from time wrapping by taking a random window of time series and perturbing a pattern in the temporal dimension. The wrap ratio R_{WW} is a parameter of window warping and can be tuned further in the experiment. The examples of different data augmentation methods are illustrated in Fig. 2.

Data normalization and moving window method are included during the pre-processing of the augmented data. We combine the synthetic data generated from different augmentation methods to obtain a big dataset. In order to better fit the machine learning algorithms, we utilize the min-max scaling to the range within [0,1].



(1). Jittering
($\sigma_\varepsilon = 0.05$)

(2). Scaling
($\sigma_\alpha = 0.5$)

(3). Permutation
($N_{\mathrm{S}} = 12$)

(4). Magnitude Warping
($\sigma_{\mathrm{M}} = 0.1, M = 4$)

(5). Window Slicing
($R_{\mathrm{WS}} = 0.6$)

(6). Window Wrapping
($R_{\mathrm{WW}} = 0.5$)

Fig. 2. Illustration of data augmentation methods.

Moving window method is applied to generate samples by using the previous l data points to predict the next data points at time t. The synthetic data samples after normalization and moving window are in the form of pairs and denoted as $\{(x_i^{\mathrm{s}}, y_i^{\mathrm{s}}), 1 \le i \le J\}$, where x_i^{s} is the input value of i^{th} synthetic sample after pre-processing, y_i^{s} is the associated output value. We also define the real data points after pre-processing as $\{(x_i^{\mathrm{r}}, y_i^{\mathrm{r}}), 1 \le i \le R\}$ is the set of the pre-processed real data samples, where $(x_i^{\mathrm{r}}, y_i^{\mathrm{r}})$ is the i^{th} data sample in real data set after pre-processing.

C. *Three-layer ensemble learning framework*

A three-layer ensemble learning framework based on staking method is proposed in this paper. In the first layer, three basic learners based on Decision Tree, including eXtreme Gradient Boosting (XGB), Gradient boosting regression (GBR) and Light Gradient Boosting Machine (LGBM), are trained in parallel. Moreover, we also implement K-fold cross-validation in the first layer to improve the generalization of the stacking process.

Suppose in the first layers there are P basic models, $M_1^{\mathrm{B}}, \dots, M_i^{\mathrm{B}}, \dots, M_P^{\mathrm{B}}$. The augmented samples after pre-processing are divided into K folds without overlapping randomly and the k_{th} validation data set is denoted as $A_k = \{(x_i^s, y_i^s), 1 \le i \le N_k\}$, where N_k is the size of each fold. Under the K-fold cross-validation, $K - 1$ folds are set as the training data set to train these basic models and the remaining one is the validation data set to output the estimations of basic learners. After K-fold cross-validation, we obtain the estimation results of each fold from the basic learner M_i as $\{\hat{y}_{i,1}^{B}, \dots, \hat{y}_{i,k}^{B}, \dots, \hat{y}_{i,K}^{B}\}$. When it comes to evaluate the performance of testing samples, the output for the first layer is the average of each trained models with different training samples. The first layer of M_i^{B} considering K-fold cross-validation is illustrated in Fig. 3.

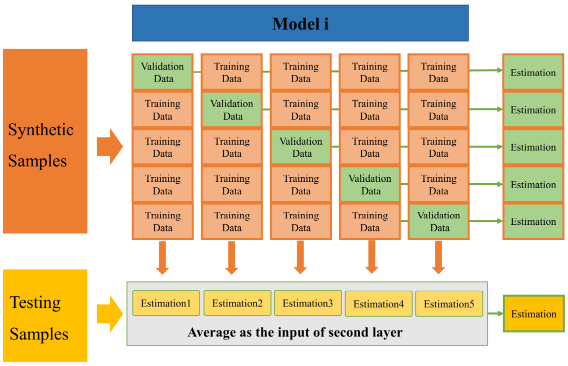


Fig. 3. K-fold cross-validation in the basic learner M_i

The stacking method works in the second layer to integrate the outputs from basic learners in the first layer. In order to avoid overfitting in the second layer, the selection of the meta learner in this layer will be relatively simple [13]. Thus, we use Lasso in this paper as a meta learner in the second layer.

Generally, there are only two layers in the classical stacking methods. In this paper, we add an additional layer by choosing a strong learner to further improve the stability and effectiveness of ensemble learning framework. With the third layer, the final estimation result is obtained as

$$y_i = \frac{y_i^{\mathrm{L}} + y_i^{\mathrm{S}}}{2} \tag{6}$$

where y_i is the final estimation of the i^{th} sample, y_i^{L} is the estimation of i^{th} sample based on LSTM that considers data reweight, y_i^{S} is the estimation of the i^{th} sample based the stacking process in the first two layers.

D. *An improved LSTM considering data reweight*

Traditionally, LSTM is used to capture the dynamic of time series by minimizing the expected loss in the training data set, where all the samples are of the same importance. However, the synthetic data generated from different data augmentation approaches have different influences on the performance of predicting the real data. This paper proposes an improve LSTM model to reweight the artificial samples.

Firstly, we apply an influence function to collect the top positive samples, following the procedure of the developed influence function [14]. We can then sort the influence values in an ascending order and collect the top I positive ones as influential data set, which is denoted as $\{(x_i^{\mathrm{I}}, y_i^{\mathrm{I}}), 1 \le i \le I\}$, to facilitate the subsequent data reweight.

Secondly, we will reweight the input synthetic sample by optimizing the prediction accuracy of the influential and synthetic samples. Instead of mixing influential and synthetic samples, the influential samples are applied as a validation set. The objective for the training model is to minimize the weighted loss in the synthetic data.

In addition, to improve the prediction accuracy on the influential data, the optimal weights can be obtained by minimizing the loss on the influential data set [15].

III. CASE STUDY

A. *Description of the case and data*

A data set from a real WWTP with a modified activated sludge process in China is applied to validate the prediction of sludge bulking, which is one of the most common failures during the operational process. Sludge volume index (SVI) is the health indicator to imply the degrading level of wastewater.

In this case, the value of SVI is collected daily with 213 data points. In the moving window method, the length of the input data l is 4, which means the previous 4-day SVI is applied to predict the value of the fifth day. 100 influential samples are selected to improve the LSTM model. Moreover, in order to test the effectiveness of the proposed method, we use 80% data to generate synthetic data and then train the ensemble learning model, and 20% data is used to test the effectiveness.

The parameters of the six augmentation methods are given in Table I. Besides, the LSTM used in this paper involves 3 BSTM layers and 2 dense layers. There are 64 neurons in each layer. Rectified Linear Units set as the active function and Adaptive Moment Estimation is the optimization method to minimize the square loss during the training process.

TABLE I
PARAMETERS OF DATA AUGMENTATION METHODS

Data augmentation methods	Parameters
Jittering	$\sigma_\varepsilon \in \{0.03,0.06,0.09\}$
Scaling	$\sigma_\alpha \in \{0.05,0.10,0.15,0.20\}$
Permutation	$N_S \in \{5,6,7,8,9\}$
Magnitude wrapping	$\sigma_{\mathrm{M}} \in \{0.1,0.2,0.3\}$ $M \in \{4,5\}$
Window slicing	$R_{\mathrm{WS}} \in \{0.5,0.6,0.7,0.8,0.9\}$
Window wrapping	$R_{\mathrm{WW}} \in \{0.05,0.10,0.15,0.2\}$

B. *Performance criteria*

Root Mean Square Error (RMSE) and mean absolute percentage error (MAPE) are two criteria used in this paper to evaluate the effectiveness of different prediction methods. The RMSE is defined as follows:

$$RMSE = \sqrt{\frac{1}{N^{\mathrm{T}}}\left(\sum_{i=1}^{N^{\mathrm{T}}}(y_i^{\mathrm{T}} - \hat{y}_i^{\mathrm{T}})^2\right)} \tag{12}$$

where N^{T} is the total number of test data samples, y_i^{T} is the i^{th} real value and $\hat{y}_i^{\mathrm{T}}$ is output of the i^{th} estimation. The definition of MAPE is given as

$$MAPE = \frac{1}{N^{\mathrm{T}}}\sum_{i=1}^{N^{\mathrm{T}}}\left|\frac{y^T-\hat{y}_i^{\mathrm{T}}}{y^T}\right| * 100\% \qquad (13)$$

The model performs more accurate when the values of RMSE and MAPE are smaller.

C. *Experiment results and comparison*

The mean and Std. of RMSE and MAPE after running the proposed methods 20 times are illustrated in TABLE II. Moreover, the results of classical prediction methods, including ARIMA, SVM, LASSO, Random Forest Regressor (RFR), GBR, LGBM, XGB, and LSTM, are also given in TABLE II. It shows that the proposed method outperforms compared with other commonly used methods. Moreover, the standability of the proposed method shows the best among these methods.

TABLE II
COMPARISON OF RMSE AND MAPE AMONG DIFFERENT PREDICTION METHODS

Methods	RMSE		MAPE (%)	
	Mean	Std.	Mean	Std.
ARIMA	13.1248	\	4.6731	\
LASSO	13.1752	0.1938	4.6319	0.0934
SVM	14.9113	0.6091	5.2918	0.2599
RFR	30.4910	0.5160	11.5360	0.262
GBR	25.1023	1.0954	8.6603	0.4772
LGBM	30.1568	1.3502	11.3648	0.6755
XGB	30.1681	2.0662	10.9646	0.9135
LSTM	14.9384	0.2567	5.3520	0.0633
Three-layer Ensemble Learning	**12.4619**	**0.0900**	**4.4070**	**0.0302**

The detailed estimations based on different prediction methods are illustrated in Fig. 4. It can be observed that XGB, LGBM, RFR and GBR fail to predict the trend of the test samples. When taking the value of SVI 230mg/L as the failure threshold, the failure data point predicted by the proposed three-layer ensemble learning method is closest to the real failure point.

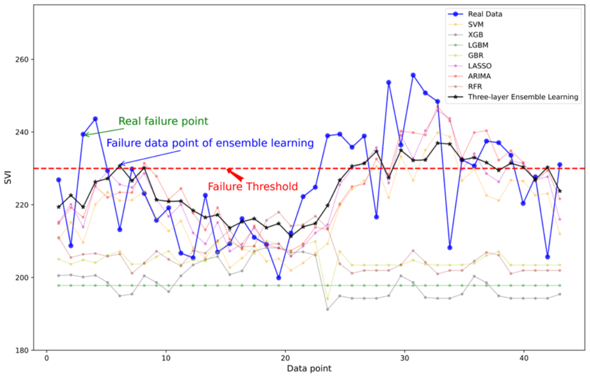
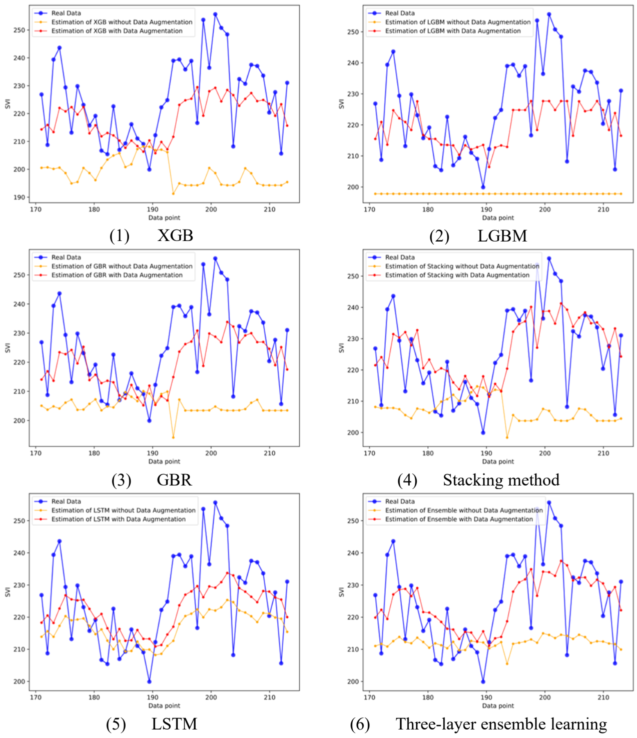


Fig. 4. Estimations of SVI value based on different prediction methods.

D. *Effectiveness of data augmentation and ensemble learning*

We also conduct comparison experiments, which include training each basic learner, the stacking method involving the first two layers, the LSTM model and the proposed three-layer ensemble learning with the real data set and the augmented data set respectively. The estimations are given in Fig. 5 to illustrate the effectiveness of data augmentation. It can be seen that training these models with a generated synthetic data set can enhance the prediction accuracy significantly.



(1) XGB

(2) LGBM

(3) GBR

(4) Stacking method

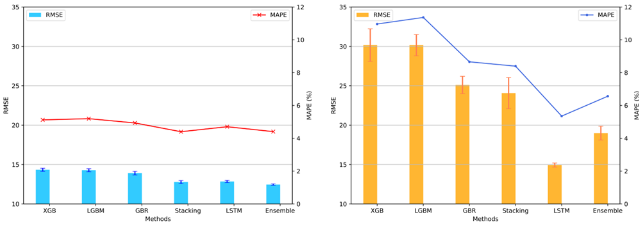
(5) LSTM

(6) Three-layer ensemble learning

Fig. 5. Estimations of basic learners, stacking method and three-layer ensemble framework trained with real data and augmented data.

E. *Effectiveness of ensemble learning*

We compared the RMSE and MAPE values of each basic learners, the stacking structure include the first two layers and the whole three-layer ensemble learning framework. Each of the experiments is trained by real limited data and augmented data respectively. The detailed results are given in Fig. 6. The results show that ensemble learning framework can decrease the errors when trained by augmented data. However, when trained by real data set, the performance of ensemble learning is worse than the LSTM model since the performance gap between the LSTM model and the stacking method is relatively large.



(1) With augmented data

(2) With real data

Fig. 6. RMSE and MAPE of XGB, LGBM. GBR, stacking method, LTSM and proposed ensemble learning training trained with real data and augmented data.

F. *Effectiveness of data reweight*

In order to verify the effectiveness of applying data reweight in the LSTM model, we make comparisons between the scenarios of applying influential samples and randomly selected samples. Moreover, we also adjust the size of influential samples and random samples from 0 to 500. The result is shown in Fig. 5. It shows that reweighting according to influential samples outperform compared randomly selected samples, which fail to improve the prediction performance in most cases.

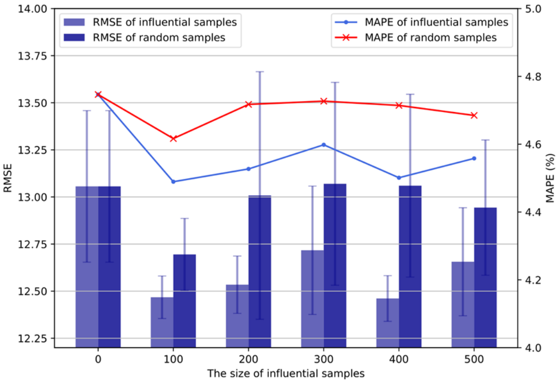


Fig. 7. RMSE and MAPE of LSTM reweighted by different size of influential samples and random samples.

IV. CONCLUSION

This study aims to improve the prediction accuracy of health indictor based on an ensemble learning framework with insufficient observational data. Multi data augmentation methods are applied to increase the size of the data set. To improve the prediction accuracy, a three-layer framework is proposed based on the stacking method to effectively integrate the augmented data. Multi basic learners are trained using augmented data set by K-fold cross-validation methods in the first layer and then the second layer integrates these estimations by the LASSO model. In the third layer, an improved LSTM considering the sample reweighting is averaged with the output from second layer. Experiments conducted on a real wastewater treatment process illustrate the effectiveness of the proposed method. In the future work, we will focus on making maintenance policy considering the uncertainty of prediction and establishing an effective health management scheme for practical use.

REFERENCES

[1] D. Wang, K.-L. Tsui, and Q. Miao, “Prognostics and health management: A review of vibration based bearing and gear health indicators,” *IEEE Access*, vol. 6, pp. 665–676, 2017.

[2] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, “Machinery health prognostics: A systematic review from data acquisition to rul prediction,” *Mechanical Systems and Signal Processing*, vol. 104, pp. 799–834, 2018.

[3] Z. Han, J. Zhao, H. Leung, K. F. Ma, and W. Wang, “A review of deep learning models for time series prediction,” *IEEE Sensors Journal*, vol. 21, no. 6, pp. 7833–7848, 2019.

[4] J. Contreras, R. Espinola, F. J. Nogales, and A. J. Conejo, “Arima models to predict next-day electricity prices,” IEEE Transactions on Power Systems, vol. 18, no. 3, pp. 1014–1020, 2003.

[5] M. Qin, Z. Li, and Z. Du, “Red tide time series forecasting by combining Arima and deep belief network,” *Knowledge Based Systems*, vol. 125, pp. 39–52, 2017. [Online]. Available:https://www.sciencedirect.com/science/article/pii/S0950705117301569

[6] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. *MIT press*, 2016.

[7] T. Xia, Y. Song, Y. Zheng, E. Pan, and L. Xi, “An ensemble framework based on convolutional bi-directional lstm with multiple time windows for remaining useful life estimation,” *Computers in Industry*, vol. 115, p.103182, 2020.

[8] Z. Li, D. Wu, C. Hu, and J. Terpenny, “An ensemble learning-based prognostic approach with degradation-dependent weights for remaining useful life prediction,” Reliability Engineering & System Safety, vol. 184, pp. 110–122, 2019.

[9] J. Zhang, F. Wu, B. Wei, Q. Zhang, H. Huang, S. W. Shah, and J. Cheng, “ Data augmentation and dense-lstm for human activity recognition using wifi signal,” *IEEE Internet of Things Journal*, vol. 8, no. 6, pp. 4628–4641, 2020.

[10] Q. Wen, L. Sun, F. Yang, X. Song, J. Gao, X. Wang, and H. Xu, “Time series data augmentation for deep learning: A survey,” *arXiv preprint* arXiv:2002.12478, 2020.

[11] Y. Peng, Y. Wang, and Y. Shao, “New data augmentation-driven RUL prognosis approach for cumulative damage model using incomplete observations,” IEEE Transactions on Instrumentation and Measurement, 2021.

[12] B. K. Iwana and S. Uchida, “An empirical survey of data augmentation for time series classification with neural networks,” *Plos one*, vol. 16, no. 7, p.e0254841, 2021.

[13] X. Fei, Q. Zhang, and Q. Ling, “Vehicle exhaust concentration estimation based on an improved stacking model,” *IEEE Access*, vol. 7, pp. 179 454–179 463, 2019.

[14] P. W. Koh and P. Liang, “Understanding black-box predictions via influence functions,” in *International Conference on Machine Learning. PMLR*, 2017, pp. 1885–1894.

[15] M. Ren, W. Zeng, B. Yang, and R. Urtasun, “Learning to reweight examples for robust deep learning,” in *International Conference on Machine Learning. PMLR*, 2018, pp. 4334–4343.