Explainable Health State Prediction for Social IoTs through Multi-Channel Attention

Yu-Li Chan, Hong-Han Shuai
Department of Electrical and Computer Engineering
National Yang Ming Chiao Tung University, Hsinchu, Taiwan

GLOBECOM 2021 - 2021 IEEE Global Communications Conference | 978-1-7281-8104-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/GLOBECOM46510.2021.9685864

***Abstract*—The core technology of Industry 4.0 is to enable the intelligence of manufacturing. One of the important tasks is anomaly detection. Although existing anomaly detection methods have achieved high accuracy, the basis of judgments cannot provide explainability, which greatly reduces the possibility for improving the model or facilitating human-machine cooperation. Therefore, in this paper, the goal is to provide the explainability for machine fault detection for social IoTs and realize the health monitoring and prognosis of the bearings simultaneously. Specifically, vibration signals from multiple sensors are transformed into spectrograms by short-time Fourier transform. Afterward, the features of frequency-domain data are extracted by the Squeeze-and-Excitation block and self-attention mechanism to assess the degradation of whole system. As such, when the process enters the early degradation, the source of components that causes the abnormality can be identified through the attention weight distribution. Experimental results show that the proposed approach achieves high accuracy in run-to-failure tests. Moreover, the proposed approach shows a better ability to explain the predicted results than the state-of-the-art bearing detection methods.**

I. Introduction

In complex industrial systems, most rotating machinery components are subject to long-term wear and tear, which can lead to failure. The issues of safety and reliability can even cause tremendous economic losses. Nowadays, owing to the development of sensing technology, a large number of sensors are configured on the machines to acquire time-domain or frequency-domain features, and to detect anomalies through data-driven methods. Compared with traditional machine health state prediction methods, deep learning-based methods have gradually gained attention due to their automatic feature extraction and performance improvement. Most of existing bearing inspection methods focus on fault diagnosis and prognostic and health management (PHM). Fault diagnosis can be classified according to different characteristics of individual bearings, *e.g.*, identifying different types of damage [17] or degrees of damage [10]. On the other hand, PHM extracts features from a set of run-to-failure data to predict the remaining useful life (RUL), health index (HI), or the onset of degradation. For example, some studies use condition based monitoring (CBM) to predict the onset of degradation [6].

Although the above methods have been extensively developed and can achieve high detection performance, two main issues arise. (1) Only the characteristics of a single bearing are considered for the prediction. (2) The results of the prediction cannot be explained. When the anomalies are detected from multivariate time series, the predicted results are often not explainable. In other words, the individual components that causes equipment anomalies cannot be directly detected, and thus cannot provide key information for operators to make decisions. In actual application scenarios, the data collected through sensors usually present multiple situations and interact with each other. Moreover, the number of sensors on a machine may be large. Therefore, it is challenging even for the technicians to find abnormal components one by one or find the relations.

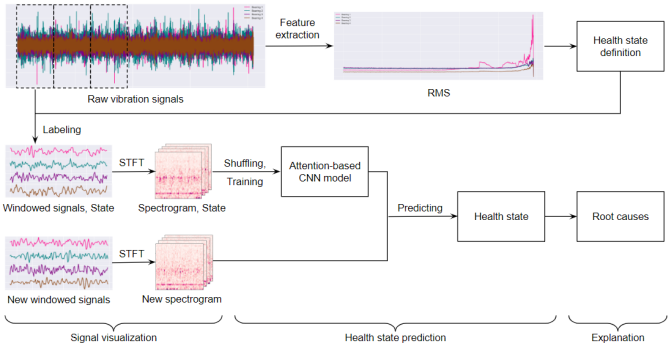


Fig. 1: The overall process.

One promising solution is to leverage the attention mechanism to select the hidden state across time steps in the recurrent neural network (RNN). However, this approach fails to indicate the importance of the input variables. Therefore, a recent line of studies proposes to apply the attention mechanism to the selection of related variables [5], [14]. Moreover, in the field of image vision, more and more architectures based on convolutional neural networks (CNNs) have introduced attention modules for improving the accuracy of the model through feature selection in various dimensions. Attention modules are usually placed in the hidden layer of the network to explore useful information between different features extracted by filters, which may provide the explainability. However, the attention can still scatter in irrelevant channels as long as it attends to the most important one [11].

In this paper, we propose an explainable machine health state prediction framework based on condition monitoring. Figure 1 shows the overall process. Specifically, the vibration signals collected by multiple sensors are first converted to time-frequency features through short-time Fourier transform (STFT). The transformed spectrograms are used as multi-channel inputs. Afterward, we applied the squeeze-and-excitation (SE) [7] operation to the input layer to explore the relationship between the channels for providing the explainability of related channels. However, SE only explores the important sensors, whereas the signals with an abnormal frequency remains unknown. Therefore, to facilitate the explainability of related frequencies, the self-attention mechanism is applied with the sparsity regularization to emphasize the information. Finally, based on the extracted features, we adopt the ResNeXt architecture [20] for the classification task.

Once the model learns how to identify the anomalies or degraded states, the attention weight can be used to identify and explain the source of the anomaly. We evaluate the performance of the proposed machine health state prediction method on two real datasets. The results manifest that the proposed approach outperforms existing methods. The proposed model can not only accurately identify the current health state of the machine, but also provide reasonable explanation of prediction results. The contributions are summarized as follows.

- We propose an explainable machine health state prediction framework, which is able to point out the cause of the anomalies when anomalies are detected.
- Multiple signals are considered at the same time, and multi-channel monitoring is adopted to meet more practical application scenarios. To the best of our knowledge, there is no previous literature that simultaneously considers multi-channel signals and provides explainability of prediction results.
- Experiments show that the classification accuracy reaches 99.88% and 98.86% on the IMS and PHM datasets, respectively. In addition, the model is able to provide a correct explanation and point out the damaged bearing.

II. Preliminary

A. Related Work

a) Explanation: Before deep learning became a prominent approach, Granger Causality (GC) was widely-used to capture the temporal dependencies between variables [15]. With the rapid advances of artificial intelligence (AI) and machine learning (ML) in recent years, apart from the accuracy of prediction, there is an urgent need to understand how the model makes a decision. Therefore, explainable AI (XAI) gradually gains attention. Conventionally, deep learning models are regarded as black boxes, which makes a trade-off between accuracy and explainability. To improve the explainability, [25] developed a training strategy based on saliency periodic feedback to represent the characteristics of objects with saliency maps, and thereby to distinguish domain specific information. Besides, they demonstrated that performance on the original domain does not deteriorate due to explainability. [2] proposed a method to score semantics of individual hidden units by evaluating the alignment to reveal the latent representation features of CNN model and to explain the basis of discriminative. Zhang et al. [24] proposed a method to modify traditional CNNs for encoding more semantically meaningful knowledge in high convolutional layers to enhance a model's interpretability. In the industrial field, [9] made an interpretation for mechanical fault diagnosis by designing a continuous wavelet convolutional (CWConv) layer which is used for the first layer of standard CNN to discover more meaningful filters. [4] used layer-wise relevance propagation (LRP) [1] as an indicator to calculate the contribution of single pixels to the diagnosis results, thereby providing an interpretation for how the CNN predicts fault type. On the other hand, attention mechanism also provides an ability to explain the predicted results. Previous works [5], [14] have investigated the influence of exogenous variables on the prediction of a target series. Nauta et al. [12] presented a deep learning framework that learns a causal graph structure by discovering the causal relationships in observational time-series data. Yeh et al. [21] proposed a multi-task learning model to investigate individual time series with respect to the final quality for early product quality prediction.

To take a trade-off between accuracy and explainability into account, most mainstream XAI methodologies belong to post-modelling explainability. The methods were based on the perturbation mechanism such as LIME. Although it can be applied to any black box model, the local perturbation mechanism must be extended for explaining models with complex outputs. Moreover, as LIME requires running a considerable amount of perturbed inputs through a model to generate an explanation, it is relatively computationally expensive. This is especially a challenge for models with high dimensional input. In contrast, the methods based on the backward propagation mechanism such as LRP only requires less inferences to be explained by the model, therefore it is more computationally efficient. However, the backward propagation mechanism is only applicable to deep networks, and explanations produced may be unreliable. Compared to our method, one of the main differences is that attention mechanisms and regularization belong to explainable modelling. Since attention mechanisms aim to identify the most relevant part of the input for a given task, it is less restricted to the number of input dimensions, and even for high-dimensional input, the computational cost does not increase too much. In addition, models using regularization can be more explainable without sacrificing the performance.

Attention mechanisms selectively focus on relevant parts that increase the representation power of their features and make the DNNs explainable. In particular, self-attention can capture global dependencies of the input by calculating the weighted sum of all positions as the context at the current position, which is first proposed in [16] for the task of machine translation. Another similar work [18] proposed non-local operation to capture interactions in both space and time dimensions. Recently, attention modules have been widely

used in CNNs and demonstrated the effectiveness in image vision fields. In addition to enriching the context by computing similarities between each two pixels [3], [22], [23], some works explored the inter-channel relationships. Moreover, [7] proposed the SE block which uses global average pooling to compute channel-wise attention. [19] added other branches to the SE block to enhance its representations, and [17] applied SENet to emphasise the major features and achieved a higher classification rate on a bearing diagnosis task.

B. Problem Formulation

Given a multivariate time series $\boldsymbol{D} \in \mathbb{R}^{C\times T}$ collected from multiple sensors, C is the number of sensors, and T is the length of the time series. The output of the model has three terms: the first output is the health state $\boldsymbol{state} \in \mathbb{R}^h$, which shows the current degradation phase of the system and is represented by one-hot vector, where h is the number of degradation phase categories. The second output is the channel attention vector $\boldsymbol{s} \in \mathbb{R}^C$, indicating that each signal has a different degree of importance. The third output is the spatial attention matrix $\boldsymbol{A} \in \mathbb{R}^{C\times H\times W}$, which indicates the working frequency region that the model pays attention to when making predictions. H and W are the height and width of the spectrogram, respectively.

III. Method

A. Signal Transformation

The multivariate time series collected by the sensor are first converted into a spectrogram by STFT because it can clearly represent the operation state of bearing and retain the time characteristic. STFT can be regarded as a series of Fourier transform (FT) of the window signal. and defined by Eq. (1) and (2) according to continuous time and discrete time:

$$\boldsymbol{STFT}\{x(t)\}(\tau,\omega) \equiv X(\tau,\omega) = \int_{-\infty}^{\infty} x(t)w(t-\tau)e^{-i\omega t}dt \tag{1}$$

$$\boldsymbol{STFT}\{x[n]\}(m,\omega) \equiv X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n} \tag{2}$$

We only consider the case of discrete time, where $w[n]$ is the window function, $x[n]$ is the signal to be transformed, and $X(m,\omega)$ is the FT of $x[n]w[n-m]$. The square of the STFT is called the spectrogram and represents the power spectral density of the function, which can be expressed as follows.

$$\boldsymbol{spectrogram}\{x[n]\}(m,\omega) \equiv |X(m,\omega)|^2. \tag{3}$$

In this step, each signal is converted by STFT and aggregated along the channel axis. That is, the input of the model is a multi-channel spectrogram, and the number of channels is equal to the number of sensors, as shown in Figure 2.

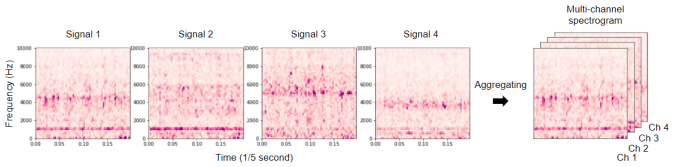


Fig. 2: An example for multi-channel spectrogram.

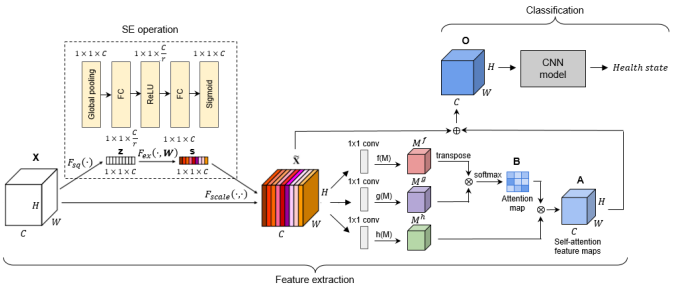


Fig. 3: Model architecture.

B. Health State Prediction

Figure 3 shows the architecture of the model, containing three main mechanisms: channel attention, position attention, and sparsity regularization. The classification tasks finally are completed using the existing ResNeXt.

1) Channel Attention: We first introduce the SE block, which learns global information and redistribute weights to the feature maps. As the SE module is used for the input layer in this work, no convolution transform operation $\boldsymbol{F}_{tr}$ is performed. Given a multi-channel spectrogram $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_C] \in \mathbb{R}^{H\times W\times C}$, $\boldsymbol{x}_c \in \mathbb{R}^{H\times W}$, generated by the signal visualization step, it is mapped to a global space 1D feature vector by a $squeeze$ transform operation $\boldsymbol{F}_{sq}$. The actual practice of $\boldsymbol{F}_{sq}$ is to use the global average pool to compress the spatial dimension $H \times W$ of $\boldsymbol{X}$ to generate $\boldsymbol{z} \in \mathbb{R}^C$, where z_c is calculated as follows.

$$z_c = \boldsymbol{F}_{sq}(\boldsymbol{x}_c) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_c(i,j). \tag{4}$$

Next, we perform the $excitation$ transform operation $\boldsymbol{F}_{ex}$ and the weight estimation on each channel through a gating mechanism with two fully-connected layers to perform adaptive feature recalibration, which can be expressed as

$$\boldsymbol{s} = \boldsymbol{F}_{ex}(\boldsymbol{z}, \boldsymbol{W}) = \sigma(g(\boldsymbol{z}, \boldsymbol{W})) = \sigma(\boldsymbol{W}_2\delta(\boldsymbol{W}_1, \boldsymbol{z})), \tag{5}$$

where δ is the ReLU activation function, and σ is the sigmoid function. In addition, since the dimensionality reduction strategy is used here, given the reduction ratio r, the dimensions of $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are $\mathbb{R}^{\frac{C}{r}\times C}$ and $\mathbb{R}^{C\times\frac{C}{r}}$, respectively. Finally, $\boldsymbol{X}$ is rescaled by activating $\boldsymbol{s}$. The output of the SE block can be written as

$$\tilde{\boldsymbol{x}_c} = \boldsymbol{F}_{scale}(\boldsymbol{x}_c, s_c) = s_c \cdot \boldsymbol{x}_c, \tag{6}$$

where $\cdot$ is a multiplication operator and $\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{x}_1}, \tilde{\boldsymbol{x}_2}, \cdots, \tilde{\boldsymbol{x}_C}]$.

2) Position Attention: After selecting the relevant channels, we introduce the self-attention mechanism to emphasize the characteristics of information in the spatial area. The output feature $\tilde{\boldsymbol{X}}$ of the SE operation is treated as $\boldsymbol{M} \in \mathbb{R}^{C\times H\times W}$, and $\boldsymbol{M}$ is first mapped to the $\boldsymbol{f}$ and $\boldsymbol{g}$ feature spaces via a 1×1 convolution operation to generate two new feature maps $\{\boldsymbol{M}^f, \boldsymbol{M}^g\} \in \mathbb{R}^{C\times H\times W}$, and this features are reshaped to the size of $C \times N$, where $N = H \times W$ is the number of pixels. The spatial attention map $\boldsymbol{B} \in \mathbb{R}^{N\times N}$ can be calculated by

$$\beta_{j,i} = \frac{exp(\boldsymbol{M}_i^f \cdot \boldsymbol{M}_j^g)}{\sum_{i=1}^{N} exp(\boldsymbol{M}_i^f \cdot \boldsymbol{M}_j^g)}, \tag{7}$$

where $\cdot$ is a multiplication operator, and $\beta_{j,i}$ represents the degree of influence of the i^{th} position on the j^{th} position. $\boldsymbol{M}$ is also mapped to another feature space $\boldsymbol{h}$ through the same 1×1 convolution operation, and its feature map is reshaped to the size of $C \times N$ to generate $\boldsymbol{M}^h \in \mathbb{R}^{C\times N}$. The output of the attention layer can be calculated as follows.

$$\boldsymbol{A}_j = \sum_{i=1}^{N} \beta_{j,i} \boldsymbol{M}_i^h. \tag{8}$$

Then, we multiply the output of the attention layer by a scale parameter γ and add it to the original input features. The final output $\boldsymbol{O} \in \mathbb{R}^{C\times H\times W}$ can be calculated by

$$\boldsymbol{O}_i = \gamma \boldsymbol{A}_i + \boldsymbol{M}_i, \tag{9}$$

where γ is a trainable parameter with an initial value of 0.

3) Attention with Sparsity Regularization: The probability distribution output via the sigmoid function or softmax function is usually not limited. As described in [7], the SE operation allows information from multiple channels to be emphasized, so the distribution of attention weights may be consistent. However, this does not help to explain the prediction results. Therefore, we applied a regularization method in the model to generate a reasonable channel attention distribution. That is, only a few of the relevant weights are positive, and the rest of the weights are close to zero. To make the weights sparse, the L1 norm is the most commonly used method. The automatic selection of features is achieved by setting useless features to 0, which can be written as

$$||\boldsymbol{\alpha}||_1 = \sum_{i=1}^{K} |\alpha_i|, \tag{10}$$

where α is the parameter to be constrained, which can be the attention weights or the coefficients of the variable, and K is the number of parameters.

4) Objective Function: In addition, L1 norm is added during the training phase thus the overall loss of our model is

$$\boldsymbol{L} = \frac{1}{N}\sum_{n=1}^{N}(-log(\boldsymbol{\eta}(\boldsymbol{scores}^{(n)})_{pos}) + \lambda \boldsymbol{R}_1(\boldsymbol{s}^{(n)})) + \frac{\mu}{2}\boldsymbol{R}_2(\boldsymbol{w}), \tag{11}$$

where N is the number of samples in a batch size, λ and μ are respectively the hyperparameter to balance between log-likelihood and regularization term, η is the softmax function, and pos indicates the positive class. The first term represents the classification loss function, while the second term is L1 norm for attention weight distributions, *i.e.*,

$$\boldsymbol{R}_1(\boldsymbol{s}^{(n)}) = ||\boldsymbol{s}^{(n)}||_1 = \sum_{i=1}^{C} |s_i^{(n)}|, \boldsymbol{R}_2(\boldsymbol{w}) = ||\boldsymbol{w}||_2 = \sum w^2. \tag{12}$$

C. Explanation

The output of the model includes the health state $\boldsymbol{state} \in \mathbb{R}^h$, the channel attention vector $\boldsymbol{s} \in \mathbb{R}^C$, and the spatial attention matrix $\boldsymbol{A} \in \mathbb{R}^{C\times H\times W}$. When the model predicts the system is in a health state, we can know from which signal the anomaly comes by the distribution of the channel attention. The spatial attention also shows the working frequency region to which the model pays attention and explains the basis of its prediction result.

IV. Experiment

A. Dataset Description

	IMS dataset	PHM dataset
Channel 1	Bearing1	Bearing3_1_H
Channel 2	Bearing2	Bearing3_1_V
Channel 3	Bearing3	Bearing3_2_H
Channel 4	Bearing4	Bearing3_2_V
Channel 5		Bearing3_3_H
Channel 6		Bearing3_3_V

TABLE I: Channel Arrangement for Two Datasets.

The first dataset is from NSF I/UCRC on Intelligent Maintenance Systems (IMS) [8] and consists of three independent test-to-failure experiments. The main shaft of the test stand is equipped with four bearings, and the rotation speed is maintained at a constant value of 2,000 RPM. Moreover, a radial load of 6,000 lbs is applied. An accelerometer is mounted on each bearing box to record vibration signals. Since our main goal is to predict the health states, we focus on the whole process of a certain fault. Specifically, the first experimental data include multi-class faults but the mixed defects is not used. In addition, as a high-level noise present in the third experimental data as described in [6], we only used the complete record of the second experiment in this dataset and considered all bearings simultaneously, and the outer race fault eventually occurred in Bearing 1. The second dataset is from the 2012 IEEE PHM Data Challenge [13] and

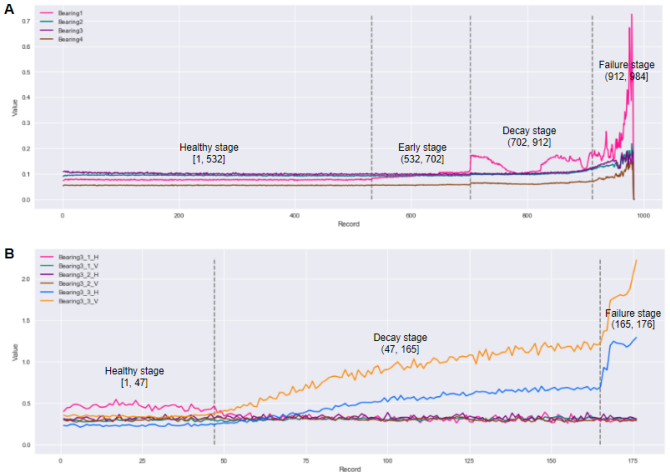


Fig. 4: Degradation states definition by RMS of the vibration signal. A) IMS. B) PHM.

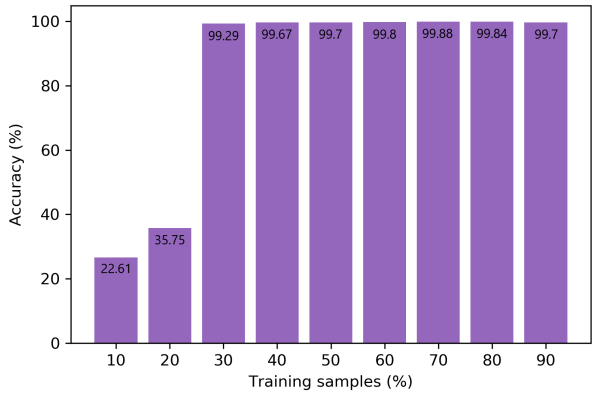


Fig. 5: Accuracy in different percentage of training samples.

consists of 17 sets of data. Each set describes a bearing test-to-failure experiment. The start and end times of the records are different, including horizontal and vertical vibration signals and temperature values. We only consider vibration signals, take three sets of data as Bearing3_1, Bearing3_2, and Bearing3_3, and merge the time zone together into six channels to simulate a multi-channel situation. The failure eventually occurs at the earliest end of the original record, recorded in Bearing3_3. Table I lists the detailed channel arrangements.

To train the CNN model, we use the time domain feature RMS to define appropriate health states for the run-to-failure process, such as the healthy stage, early stage, decay stage, and failure stage. The distinction of states mainly depends on the large changes. Figure 4 shows the detailed definitions and intervals of degradation states for the two datasets, IMS and PHM, respectively.

B. Implementation Details

All samples are shuffled during training. The first 70% is used as the training set and the remaining 30% is used as the validation set. We also train the model with different percentage of training samples and compare their accuracy, as shown in Figure 5. The initial learning rate is 0.1 with a decay rate of 10% every 30 epochs. We use Momentum to optimize our network and the whole training process goes through 100 epochs with a weight decay of 0.0005. The batch size is 32, and the reduction ratio r of the SE operation is set to 2.

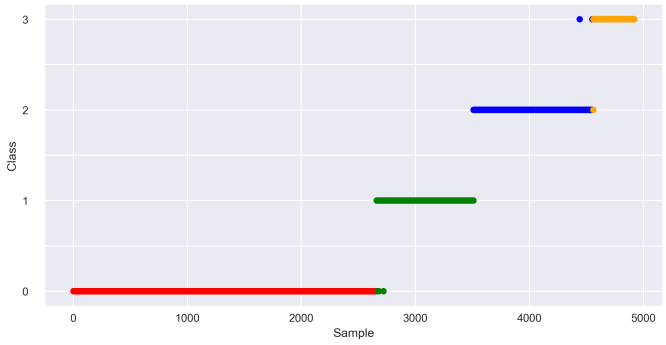
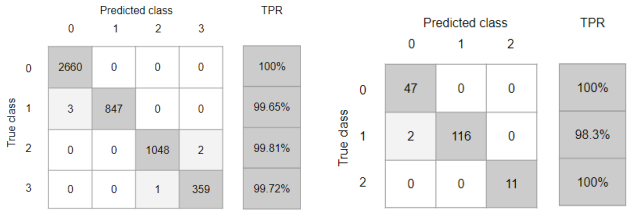


Fig. 6: Actual prediction on IMS dataset.



(a) IMS dataset (b) PHM dataset

Fig. 7: Confusion matrix of the classification results.

C. Quantitative Results

IMS Dataset. We cut the signal recorded per second into 5 samples, giving a total of $4,920$ samples. Figure 6 shows that most of the incorrect prediction occurs where the state changes, which may relate to the label definition. However, the model predicts correctly after the 2,685-th sample. That is, after the 537-th record of the dataset, which is only 5 records away from the interval we previously defined. This finding proves that the model can detect anomalies at an early stage. The overall classification accuracy is 99.88%. Figure 7(a) shows the confusion matrix of the classifier, where class 0 is healthy stage, class 1 is early stage, class 2 is decay stage, and class 3 is failure stage. Once the model detects that the current system is unhealthy, the channel attention is used to identify which signal caused the anomaly. Figure 8 shows the channel attention during state monitoring. The source of the anomaly is Bearing 1. This result is consistent with the results shown by the RMS characteristics, *i.e.*, the outer race fault finally occurred in Bearing 1 from dataset description.

PHM Dataset. We consider the signals recorded every 0.1 second as a sample, resulting in a total of 176 samples. As shown in Figure 9, the incorrect prediction only occurs to the first two samples of degradation phase. The classification

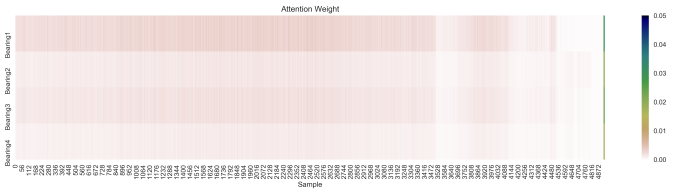


Fig. 8: Channel attention on IMS Dataset.

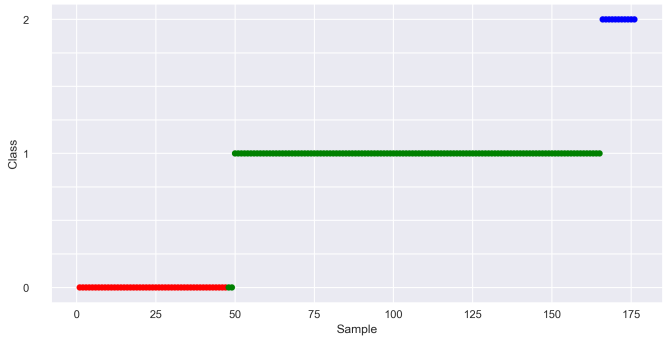


Fig. 9: Actual prediction on PHM dataset.

accuracy for all stages is 98.86%. Figure 7(b) shows the confusion matrix of the classifier, where class 0 is healthy stage, class 1 is decay stage, and class 2 is failure stage.
According to the RMS characteristics, we expect the channel attention to focus on horizontal or vertical vibration signals of Bearing3_3. Figure 10 shows the actual results. The prediction is mainly based on Bearing3_3_V. Bearing3_3_H also has large changes in vibration value but is not considered. This is because the characteristics of remaining channels are considered as redundant features if the characteristics of one channel can predict the results.

ACKNOWLEDGEMENT

This work is supported in part by the Ministry of Science and Technology (MOST) of Taiwan under the grants MOST-109-2221-E-009-114-MY3 and MOST-110-2221-E-001-001. This work was also supported by the Higher Education Sprout Project of the National Yang Ming Chiao Tung University and Ministry of Education (MOE), Taiwan.

V. CONCLUSIONS

In this paper, we propose an explainable health state prediction model, which consists of SE block and self-attention, together with the proposed sparsity regularization for further enhancing the explainability. Experimental results manifest that the proposed model can not only predict the correct states but also provide the explainability.

REFERENCES

[1] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Muller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 2015.
[2] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. *CVPR*, 2017.

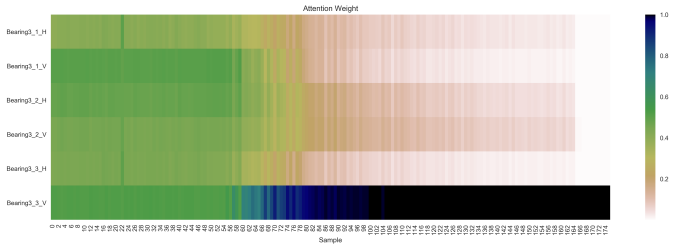


Fig. 10: Channel attention on PHM Dataset.

[3] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. Dual attention network for scene segmentation. *CVPR*, 2018.
[4] J. Grezmak, J. Zhang, P. Wang, K. Loparo, and R. X. Gao. Interpretable convolutional neural network through layer-wise relevance propagation for machine fault diagnosis. *IEEE Sensors Journal*, 2019.
[5] T. Guo and T. Lin. Multi-variable lstm neural network for autoregressive exogenous model, 2018.
[6] R. Hasani, G. Wang, and R. Grosu. A machine learning suite for machine components' health-monitoring. *AAAI*, 2019.
[7] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. Squeeze-and-excitation networks. *CVPR*, 2018.
[8] J. Lee, H. Qiu, G. Yu, and J. Lin. Rexnord technical services. *Bearing data set*, 2007.
[9] T. Li, Z. Zhao, C. Sun, L. Cheng, X. Chen, R. Yan, and R. X. Gao. Waveletkernelnet: An interpretable deep neural network for industrial intelligent diagnosis, 2019.
[10] X. Li, S. Wang, W. Zhou, Q. Huang, B. Feng, and L. Liu. Research on fault diagnosis algorithm based on structure optimization for convolutional neural network. *IJCAI*, 2019.
[11] A. K. Mohankumar, P. Nema, S. Narasimhan, M. M. Khapra, B. V. Srinivasan, and B. Ravindran. Towards transparent and explainable attention models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 4206–4216, July 2020.
[12] M. Nauta, D. Bucur, and C. Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, pages 312–340, 2019.
[13] P. Nectoux, R. Gouriveau, K. Medjaher, E. Ramasso, B. Morello, N. Zerhouni, and C. Varnier. Pronostia: An experimental platform for bearings accelerated degradation tests. *IEEE International Conference on Prognostics and Health Management*, 2012.
[14] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. Cottrell. A dual-stage attention-based recurrent neural network for time series prediction. *IJCAI*, 2017.
[15] H. Qiu, Y. Liu, Niranjan, Subrahmanya, and W. Li. Granger causality for time-series anomaly detection. *ICDM*, 2012.
[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017.
[17] H. Wang, J. Xu, R. Yan, and R. X. Gao. A new intelligent bearing fault diagnosis method using sdp representation and se-cnn. *IEEE Transactions on Instrumentation and Measurement*, 2019.
[18] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. *CVPR*, 2018.
[19] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module, 2018.
[20] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *CVPR*, 2017.
[21] C.-H. Yeh, Y.-C. Fan, and W.-C. Peng. Interpretable multi-task learning for product quality prediction with attention mechanism. *ICDE*, 2019.
[22] Y. Yuan and J. Wang. Ocnet: Object context network for scene parsing, 2018.
[23] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks, 2018.
[24] Q. Zhang, Y. N. Wu, and S.-C. Zhu. Interpretable convolutional neural networks. *CVPR*, 2018.
[25] A. Zunino, S. A. Bargal, R. Volpi, M. Sameki, J. Zhang, S. Sclaroff, V. Murino, and K. Saenko. Explainable deep classification models for domain generalization, 2020.