IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 3, 2020, accepted June 25, 2020, date of publication July 2, 2020, date of current version July 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3006491

Vibration Signals Analysis by Explainable Artificial Intelligence (XAI) Approach: Application on Bearing Faults Diagnosis

HAN-YUN CHEN AND CHING-HUNG LEE, (Senior Member, IEEE)
Department of Mechanical Engineering, National Chung Hsing University, Taichung 402, Taiwan

Corresponding author: Ching-Hung Lee (chleenchu@dragon.nchu.edu.tw)

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Contract MOST-109-2634-F-005-004, Contract 108-2634-F-005-001, Contract 107-2634-F-005-001, Contract 108-2218-E-150-004, and Contract 106-2221-E-005-010-MY3.

ABSTRACT This study introduces an explainable artificial intelligence (XAI) approach of convolutional neural networks (CNNs) for classification in vibration signals analysis. First, vibration signals are transformed into images by short-time Fourier transform (STFT). A CNN is applied as classification model, and Gradient class activation mapping (Grad-CAM) is utilized to generate the attention of model. By analyzing the attentions, the explanation of classification models for vibration signals analysis can be carried out. Finally, the verifications of attention are introduced by neural networks, adaptive network-based fuzzy inference system (ANFIS), and decision trees to demonstrate the proposed results. By the proposed methodology, the explanation of model using highlighted attentions is carried out.

INDEX TERMS Convolutional neural network, vibration signal, explainable AI, fault diagnosis.

I. INTRODUCTION

The physical phenomena causing vibrations are reflected on the signals acquired by sensors and data acquisition systems. There are several traditional methods to analyze information of signals [1]–[8]. Fast Fourier transform (FFT), which is employed inclusively in the field of signal processing, is applied to find the frequency distribution. Empirical mode decomposition can help decomposing signals into intrinsic mode functions with different frequency ranges to isolate signal components within specific frequency bands. Others are proposed to find the characteristics and decompositions of signal, for instance, wavelet transform, etc. Instead of complex formulation, machine learning can help modeling the relation between features and physical phenomena. The features can be extracted in domains of signals, e.g., statistical analysis, or by models automatically, e.g., autoencoder. These features can be inputs of machine learning models for prediction and classification.

Rolling element bearings (REBs) are crucial components inside rotating machines. Bearing failures can cause serious safety issues, to prevent the damages caused by bearing failures, diagnosis methods using vibration signals of bearings and machine learning methods are proposed by many researches [9]–[14]. Before convolutional neural networks (CNN) and other deep learning methods are popularized, characteristics of vibration signals are computed and utilized to help identifying the signals. By considering rotating motion and relative motion between elements of REBs, characteristic frequencies, which are common characteristics, can be computed and discussed [9], [12]–[14]. There are more information in frequency domain than time-domain signals, therefore, frequency spectra are used for predictions [15], [16]. The statistical features mentioned above also can be applied to build automatic diagnosis models [17]–[19]. Recently, CNNs are applied in some researches for classification and prediction, in which the features can be extracted automatically [20], [21]. If a two-dimensional CNN is utilized for vibration signals analysis, the inputs should be chosen as time-frequency spectra, grey level images of signals, or other two-dimensional data or images [22]–[24].

Deep learning can provide excellent performance in prediction and classification. However, the parameters inside models with network structures are unexplainable and lack of practical meanings. In recent years, explainable AI (XAI) becomes a popular field [25]–[28]. The main objective of XAI is to convince users that machine learning can provide reliable prediction and make machine learning methods more transparent. When the model predicts wrong, explanations can help tracking the reasons and phenomena. The explaining

The associate editor coordinating the review of this manuscript and approving it for publication was Nuno Garcia.

methods are summarized as follows [27]: (a) Explaining the processing of data by networks [26], [29]–[31], e.g., class activation mapping (CAM), gradient class activation mapping (Grad-CAM), and local interpretable model-agnostic explanation. CAM and Grad-CAM are applied to create salience maps and attention maps which help explaining machine learning models by highlighting the attention area of models. (b) Explaining the representation of data inside networks [26], [32], [33]. It explains the models by internal structures of models, such as the role of layers, individual units, and vectors. (c) Create explanation-producing systems designed to simplified explanation of their behaviors [34]. Explanation-producing systems have structures designed to simplify the original models, which make their processing or operation easier to be understood. For instance, attention-based model is a success methodology used in natural language translation and image caption generation, which are sequence problems. The mentioned methods can achieve explanation of algorithms, but the explanations of machine learning for vibration signals are less common.

In this study, explanation of convolutional neural networks (CNNs) for vibration analysis is discussed. CNN is employed as classification model for REB faults. Though CNN can find features automatically, the features cannot be recognized or identified manually. By visualizing the attention of CNN model with Grad-CAM, the explanation of CNN can be carried out. The explanation is verified by neural network and other simple machine learning methods. The proposed explanation process explains model using the processing of data by networks [27]. Also, the proposed process belongs to ''explain to justify'' and ''explain to discover'' [25]. For ''explain to justify'', the model is proved to be legitimate. For ''explain to discover'', a new phenomenon, which is different from traditional cognition and analysis method, is discovered: The characteristics in low-frequency bands are applied to classify and diagnose CWRU bearings using traditional analysis methods. However, CNN tends to focus at high frequency bands more.

In the rest of paper, preliminaries, including CNN, Grad-CAM, short-time Fourier transform, and analyzed dataset, are introduced in Section II. The proposed method and applied CNN model are illustrated in Section III, including the performance and explanation. The results of using proposed method are discussed in Section IV. Finally, the conclusion is given in Section V.

II. PRELIMINARIES

A. CONVOLUTIONAL NEURAL NETWORKS (CNNs)

Convolutional Neural Network (CNN) is first proposed in 1998 [35], it is always utilized for classification and prediction in image processing and other researching fields. Structure of a classical CNN is shown as FIGURE 1, there are three basic types of layers, including convolutional layers, pooling layers, and fully-connected layers. In convolutional layers, the inputs are convolved with filters to bring out features. The hyperparameters are L, W (length and width of inputs), L_C, W_C (length and width of filters), S_{CL}, S_{CW} (stride of filters in different directions), N (filter number). The operation of single filter in convolutional layer is

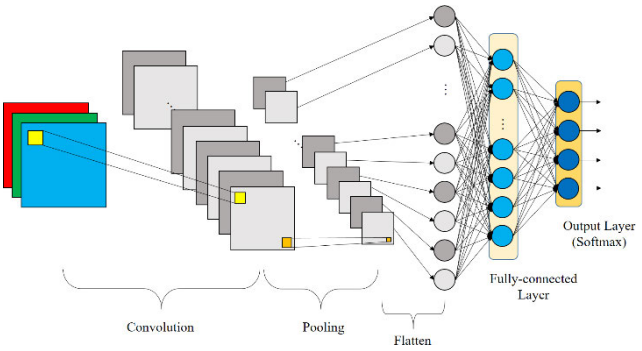

FIGURE 1. Structure of a classical CNN.

$$z_l^k = f(\boldsymbol{\alpha_l} * \boldsymbol{x} + b), \tag{1}$$

where $\boldsymbol{x} \in \mathcal{R}^{\boldsymbol{W}\times\boldsymbol{L}}$ is the input, $l = 1, 2, \ldots, N$ is the index of filters in the convolution layer, k is the index of convolutional layer, f is the nonlinear activation function, $*$ represents the convolutional operation, b is the bias, z_l^k is the feature map generated by the lth filter, $\boldsymbol{\alpha}_l$ is the corresponding kernel matrix of the lth filter. The output (feature map) of the kth convolutional layer is

$$\boldsymbol{z^k} = [z_1^k, z_2^k, \ldots, z_N^k]. \tag{2}$$

The length and width of feature map after convolved by single filter become $L_z = \text{ceil}(\frac{L-L_c}{S_{CL}})$ and $W_z = \text{ceil}\left(\frac{W-W_c}{S_{CW}}\right)$.

The objective of pooling layer is to extract and reserve important information in feature maps. Max-pooling operation is mostly used to achieve the desired objective. The hyperparameters are L_P and W_P (length and width of filters). The operation of single filter in max-pooling layer is

$$p^k_{l\,q,r} = \max\left(\begin{bmatrix} z^k_{l\,q,r} & z^k_{l\,q,r+1} & \cdots & z^k_{l\,q,r+L_P} \\ z^k_{l\,q+1,r} & & \ddots & z^k_{l\,q+1,r+L_P} \\ \vdots & & & \vdots \\ z^k_{l\,q+W_P,r} & z^k_{l\,q+W_P,r+1} & \cdots & z^k_{l\,q+W_P,r+L_P} \end{bmatrix}\right), \tag{3}$$

where q and r are the row and column index of features after pooling where $q = 1, 2, \ldots, \text{ceil}\left(\frac{L_z}{L_P}\right)$, $r = 1, 2, \ldots, \text{ceil}\left(\frac{W_z}{W_P}\right)$. If the boundary of filter is beyond the range of signal, zero padding is applied to fit feature map's size with the size of filter. p_l^k is the pooling result of the lth feature map after convolved by the kth convolutional layer. The lth feature map after pooling can be represented as

$$p_l^k = \begin{bmatrix} p^k_{l\,1,1} & \cdots & p^k_{l\,1,\text{ceil}\left(\frac{W_z}{W_P}\right)} \\ \vdots & \ddots & \vdots \\ p^k_{l\,\text{ceil}\left(\frac{L_z}{L_P}\right),1} & \cdots & p^k_{l\,\text{ceil}\left(\frac{L_z}{L_P}\right),\text{ceil}\left(\frac{W_z}{W_P}\right)} \end{bmatrix}. \tag{4}$$

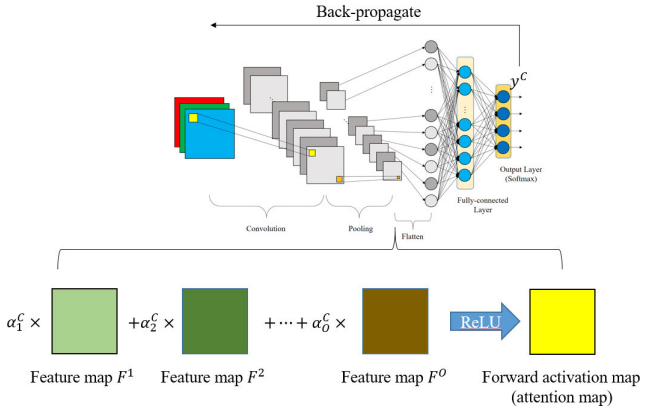

FIGURE 2. An illustration of Grad-CAM.

The feature maps after convolving and pooling are flatten into one-dimensional features which are inputs of artificial neural networks. Fully-connected layers are back-propagation neural networks which are widely applied on prediction and classification. The feedforward operation of neurons in fully-connected layers is

$$y = f\left(\sum_{a=1}^{n} w_a h_a + b\right), \tag{5}$$

where h_a is the input of neuron, w_a is weight of h_a, $a = 1, 2, \ldots, n$, b is the bias, f is the activation function of neuron, y is the output of neuron. Finally, a fully-connected layer with softmax activation function is applied for classification. The outputs of softmax function will be in range of 0 and 1. Softmax function is denoted as

$$\sigma(y) = \frac{e^y}{\sum e^y}. \tag{6}$$

In this study, the popular CNN structures are not applied due to the huge number of parameters, e.g., VGG16 has over 10 million of parameters [36]. Therefore, a simple structure of CNN is introduced to classify the bearing faults.

B. GRADIENT CLASS ACTIVATION MAPPING (GRAD-CAM)

Grad-CAM is proposed to understand the attention of CNN model for classification [31]. By computing the weights of each feature maps with respect to classification scores, the heat maps (or attention maps) are generated. An illustration of Grad-CAM is shown in FIGURE 2. In the following, the major operation of Grad-CAM is presented. At first, the oth rectified feature map F^o can be represented as

$$F^o = \frac{1}{Z}\sum_m \sum_n A^o_{m,n}, \tag{7}$$

where Z is the number of pixels in the feature map, m and n are the index of row and column of the feature map, $A^o_{m,n}$ is the value of pixel in the mth row and nth column.

Since Grad-CAM is an improved method of class activation mapping (CAM) [30], the attention map of CAM is the beginning of proving Grad-CAM. The score of class C (y^C), also known as the value of class C before softmax layer, generated by CAM can be represented as

$$y^C = \sum_o \alpha_o^C F^o, \tag{8}$$

where α_o^C is the weight of the oth feature map. The forward activation map ($L^C_{Grad-CAM}$) generated using Grad-CAM is

$$L^C_{Grad-CAM} = \text{ReLU}\left(\sum_o \alpha_o^C F^o\right). \tag{9}$$

The weights of single pixel can be evaluated using gradient method. The process can be represented as

$$\alpha_o^C = \frac{\partial y^C}{\partial F^o} = \frac{\partial y^C}{\partial A^o_{m,n}}\frac{\partial A^o_{m,n}}{\partial F^o} = \frac{\partial y^C}{\partial A^o_{m,n}} \cdot Z. \tag{10}$$

The weight of the oth feature map is the average of weights of every pixel in the feature map. The weight of the oth feature map is

$$\frac{1}{Z}\sum_m \sum_n \alpha_o^C = \frac{1}{Z}\sum_m \sum_n \frac{\partial y^C}{\partial A^o_{m,n}} \cdot Z. \tag{11}$$

Since the weights of each pixel in the oth feature map should be the same, (11) is simplified as

$$\frac{1}{Z}\cdot Z \cdot \alpha_o^C = \frac{1}{Z}\sum_m \sum_n \frac{\partial y^C}{\partial A^o_{m,n}} \cdot Z = \sum_m \sum_n \frac{\partial y^C}{\partial A^o_{m,n}}, \tag{12}$$

$$\alpha_o^C = \sum_m \sum_n \frac{\partial y^C}{\partial A^o_{m,n}}. \tag{13}$$

Finally, the normalized localization map (heat map or attention map) using Grad-CAM can be represented as

$$S = \frac{1}{Z}\sum_m \sum_n \sum_o \alpha_o^C A^o_{m,n}. \tag{14}$$

Before applying Grad-CAM, the model must be trained. Then the attentions of model with specific inputs can be computed using Grad-CAM. The advantage of using Grad-CAM is that the structure of model will not affect the process since the whole operation is based on feature maps and classification results. By observing and analyzing attention maps, the attention of models can be explained and verified.

C. SHORT-TIME FOURIER TRANSFORM (STFT)

The frequency spectra of signals can be computed using Fourier transform (FT). In order to apply FT to real-time application, discrete Fourier transform (DFT) is introduced and represented as

$$\text{FFT}(x[n]) \equiv X\left(e^{-j\omega}\right) = \sum_{n=0}^{N-1} x[n]\, e^{-j\omega n}. \tag{15}$$

From (15), a discrete signal x with size N can be understood as a series of frequencies ω with different magnitudes of energy. Another type of representation for DFT is known as

$$\text{FFT}(x[n]) \equiv X_k = \sum_{n=0}^{N-1} x[n]\, e^{-\frac{j2\pi kn}{N}}, \quad k = 0, 1, \ldots, N-1. \tag{16}$$

Though DFT works well to evaluate the frequency components of signal, the computational effort of DFT is considerable.

In this study, short-time Fourier transform (STFT) is adopted to transform vibration signal to image [37]. Although frequency spectra provide information in frequency domain,

TABLE 1. Characteristic frequencies under different rotating speed.

Rotating speed (rpm)	Freqs. (Hz)	F_{BPO}	F_{BPI}	$2F_{BS}$	F_C
1797		107.364	162.186	141.168	11.929
1772		105.870	159.930	139.204	11.763
1750		104.556	157.944	137.468	11.617
1730		103.361	156.139	135.904	11.485

TABLE 2. Four types of fault and corresponding label.

Class label	0	1	2	3
Type of fault	Normal	Inner ring	Outer ring	Ball

there is no any time-domain information. If arranging the frequency spectra of short signal segments together, the variety of frequency in time domain can be observed. In other word, the information of both time domain and frequency domain are shown at the same time. The signal is broken up into frames by window functions, each frame is Fourier transformed. The equation can be represented as

$$\mathrm{STFT}\left(x\left[n\right]\right) \equiv X\left(m, e^{-j\omega}\right) = \sum_{n=0}^{N-1} x\left[n\right] w[n-m] e^{-j\omega n} \tag{17}$$

where w is discrete window function, m is discrete index in the window w. FFT can also be used in STFT to enhance efficiency and reduce requirement in computation.

D. BEARING DATASET

Bearing dataset used in this paper is provided by Case Western Reserve University (CWRU), the dataset is widely discussed and analyzed in many researches [38]–[42]. The signals applied in this study are collected using accelerometers mounted at the drive end of the motor with 12 kHz of sampling frequency. There are four different damage conditions of bearings, (1) normal; (2) inner ring faults; (3) outer ring faults; and (4) ball faults. These faults with different diameters are man-made using electrical-discharge machine. The diameters are not discussed in this study, for example, a 7-mil inner ring fault and a 14-mil inner ring fault are considered as the same class. There are four different types of load on the motor which lead to different rotating speeds. The characteristic frequencies under different rotating speeds are evaluated and shown in TABLE 1, and the labels are manually set as TABLE 2.

There are 64 data in the dataset, sliding window is applied to slice original signals into one-second signals and increase the number of training and testing data. The length and stride of window are 12000 data points and 3000 data points respectively. There are 2368 data after applying sliding window with the four mentioned bearing conditions under different rotating speeds. The training sets and testing sets are selected randomly, 1657 data (70% of data) for training, 711 data (30% of data) for testing.

III. METHODOLOGY FOR VIBRATION SIGNALS ANALYSIS USING XAI APPROACH

In this section, the proposed methodology is introduced, the corresponding scheme of proposed method is shown

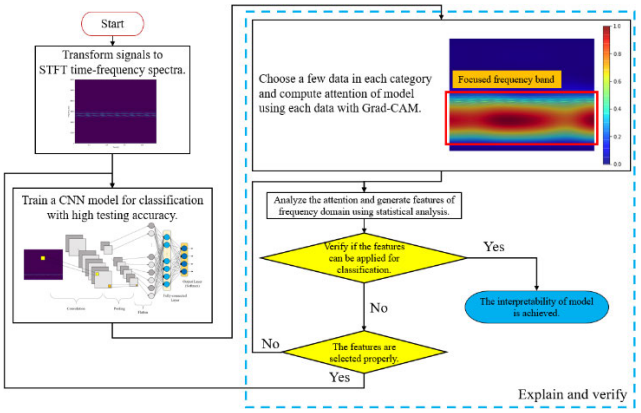

FIGURE 3. Scheme of explaining CNN models for vibration signals analysis.

in FIGURE 3. At first, the signal segmentations are transformed into STFT time-frequency spectra. Then, a CNN model is trained to classify the bearing conditions. The testing accuracy is recommended to be as high as possible. After the model is trained, a few data are applied to help explaining the model and verifying the explanation. The attentions of model are computed using Grad-CAM. The attention can show the frequency band which model focuses at the most. After analyzing the frequency band with high attention, the features of the focused frequency band are generated using statistical analysis methods. Finally, these features are adopted to verify the attention using other simple model, e.g., NN with simple structure, ANFIS, and decision trees utilized in this study. If the models can provide great performance, the features contain information for classification. The frequency band where the model pays more attention is legitimate.

The process of training CNN, analyzing attention of model, and assumption of explanation are discussed as follows. Classification performance of CNN using STFT time-frequency spectra is discussed in part A firstly. Then, the attention of model is analyzed and explanation of CNN is carried out in part B. The verifications are introduced in Section IV.

A. CLASSIFICATION OF BEARING FAULTS USING TIME-FREQUENCY SPECTRA OF VIBRATION SIGNALS

Time-frequency spectra of one-second signals in CWRU dataset mentioned in the end of part D in Section II are generated using STFT. FIGURE 4(a), FIGURE 4(b), FIGURE 4(c), and FIGURE 4(d) are time-frequency spectra of a normal bearing, a bearing with inner ring fault, a bearing with outer ring fault, and a bearing with ball fault, respectively. By observation, the time-frequency spectra are different between each status of bearings. The structure of CNN using time-frequency spectra to classify bearing faults is shown in TABLE 3. The initial learning rate is 0.001, the optimizer is Adam. Categorical cross-entropy is adopted as loss function. The training and testing accuracy of model are 100%. The confusion matrix of model predicting

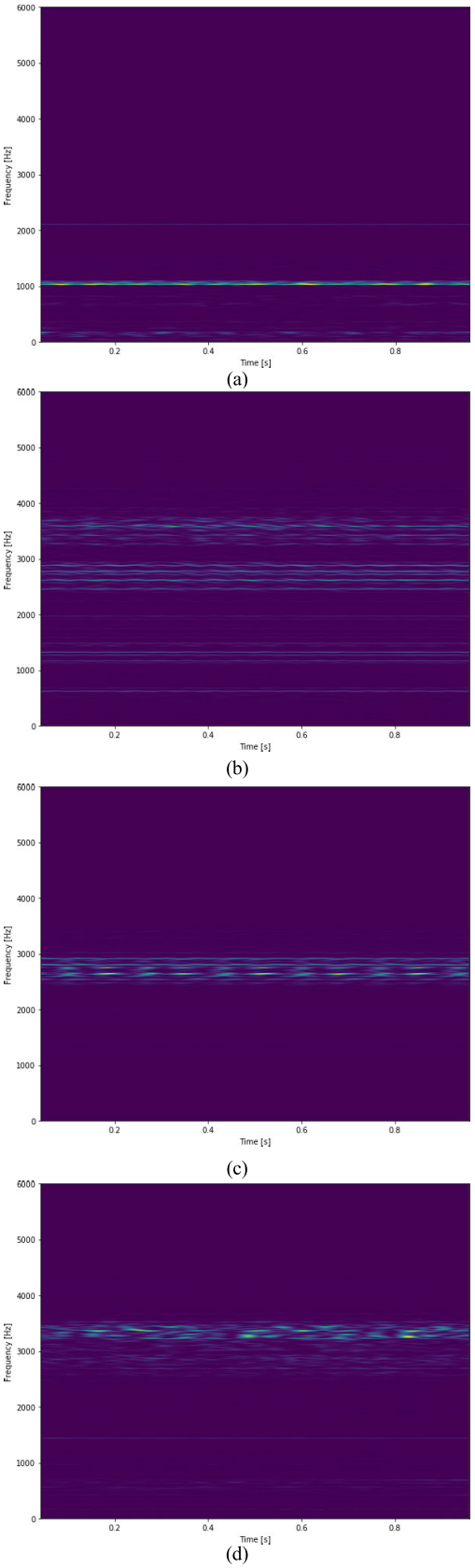

FIGURE 4. Time-frequency spectra after STFT of (a) a normal bearing, (b) a bearing with inner ring fault, (c) a bearing with outer ring fault, and (d) a bearing with ball fault.

TABLE 3. Structure of 2DCNN for bearing classification.

Layer	Filter size	Stride	Number of filters or nodes	Activation function
Conv. 1	9×9	2×2	4	ReLU
Conv. 2	9×9	2×2	8	ReLU
Pool. 2	4×4			
Conv. 3	4×4	2×2	16	ReLU
Conv. 4	4×4	2×2	32	ReLU
Pool. 4	2×2			
Flatten				
Fully-Conn. 1			128	ReLU
Fully-Conn. 2			32	ReLU
Output			4	Softmax
Total parameters	28424			

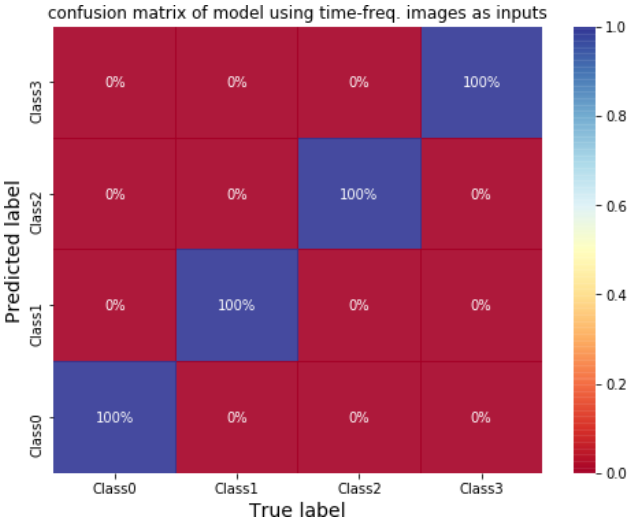

FIGURE 5. Confusion matrix of CNN model for classifying CWRU bearing data.

testing data is shown in FIGURE 5. The accuracy shows that time-frequency spectra can be adopted for classification.

B. EXPLANATION UAING TIME-FREQUENCY SPECTRA

Firstly, the attentions of model using Grad-CAM are shown in FIGURE 6. FIGURE 6(a), FIGURE 6(b), FIGURE 6(c), and FIGURE 6(d) are the results of a normal bearing, a bearing with inner ring fault, a bearing with outer ring fault, and a bearing with ball fault, respectively. The rotating speed of bearings are 1797 rpm. The left ones are the input images with axes and the right ones are attention maps. By observing the attention map of a normal bearing in FIGURE 6(a), the model focuses at low-frequency band since there is no obvious structure resonance for a normal bearing. As shown in FIGURE 6(b), the attention of model using a bearing with inner ring fault is focusing at high-frequency bands from about 1000 Hz to 4000 Hz, which is caused by structure resonance [43], [44] but not where the characteristic frequency locates. The time-frequency spectra of bearing with outer ring fault and ball fault are shown in

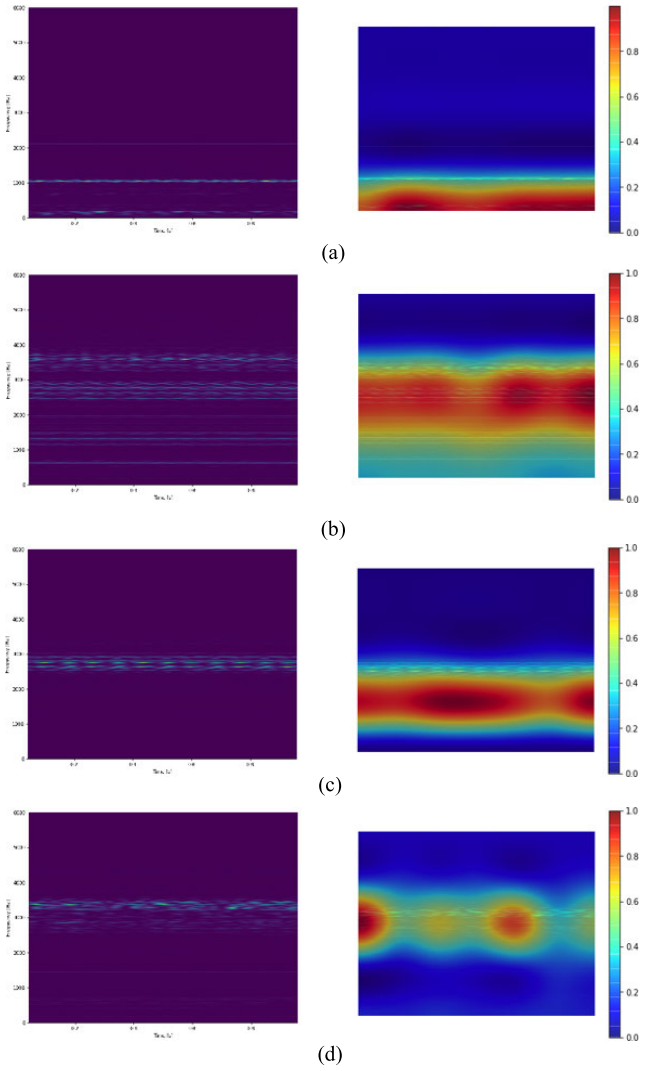

FIGURE 6. Input image with axis (left) and input image overlapped with attention map (right) of (a) a normal bearing, (b) a bearing with inner ring fault, (c) a bearing with outer ring fault, and (d) a bearing with ball fault under 1797 rpm.

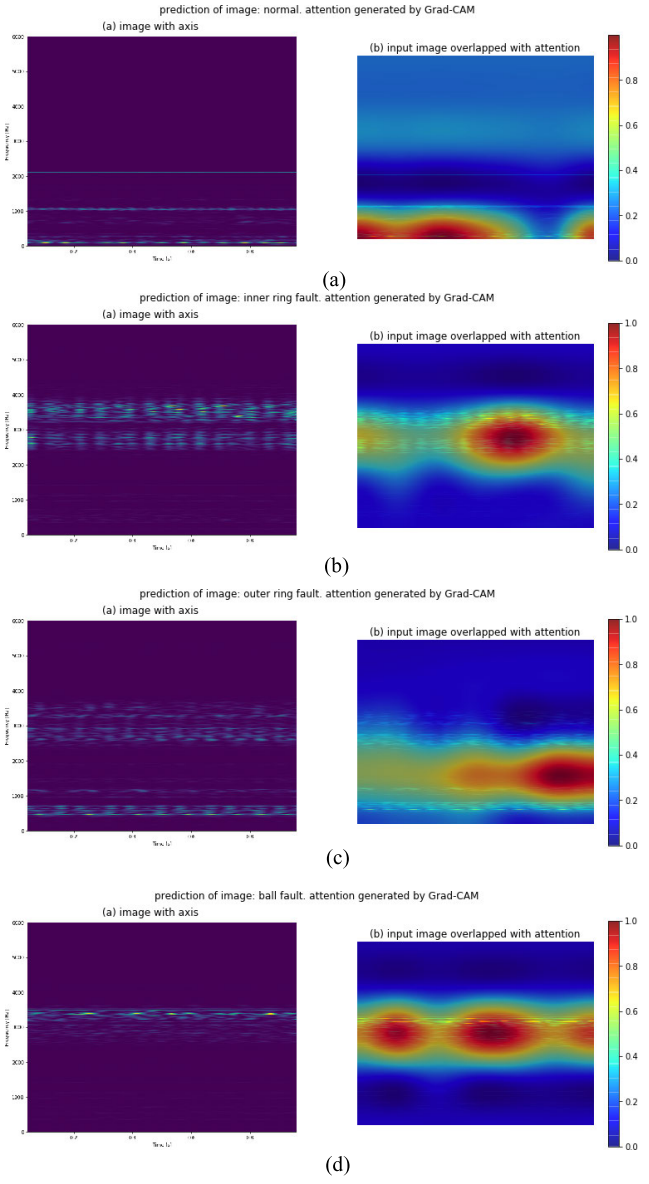

FIGURE 7. Input image with axis (left) and input image overlapped with attention map (right) of (a) a normal bearing, (b) a bearing with inner ring fault, (c) a bearing with outer ring fault, and (d) a bearing with ball fault under 1772 rpm.

FIGURE 6(c) and FIGURE 6(d). The model is still focusing at high-frequency bands which is identical to the result using a bearing with inner ring fault. The attention and analysis of attention using bearings under 1772, 1750, 1730 rpm are shown in FIGURE 7, FIGURE 8, and FIGURE 9. The same phenomena can be observed.

After analyzing the attentions, an assumption of explanation for the model can be carried out: the features in high-frequency band can be applied for classification more easily for machine learning models than the characteristics shown in TABLE 1. The verifications of explanation are introduced in next section.

IV. VERIFICATIONS OF EXPLANATIONS

In this section, several examples are introduced to verify using simple neural network, decision trees, and adaptive network-based fuzzy inference system (ANFIS) that the features of high-frequency bands can be used to classify the faults. The comparison between decision trees and ANFIS is also discussed. Finally, the summary of verifications is brought out in E.

A. OBSERVATION AND ANALYSIS FOR HIGH-FREQUENCY BANDS

At first, the frequency distribution of bearings with different conditions in CWRU bearing data are observed. The average frequency spectra of bearings under different conditions are computed and shown in FIGURE 10. FIGURE 10(a), FIGURE 10(b), FIGURE 10(c), and FIGURE 10(d) are spectra of normal bearings, bearings with inner ring faults, outer ring faults, and ball faults, respectively. By observation, the distributions of frequency in 1000∼4000 Hz are different and can be applied for classification preliminarily.

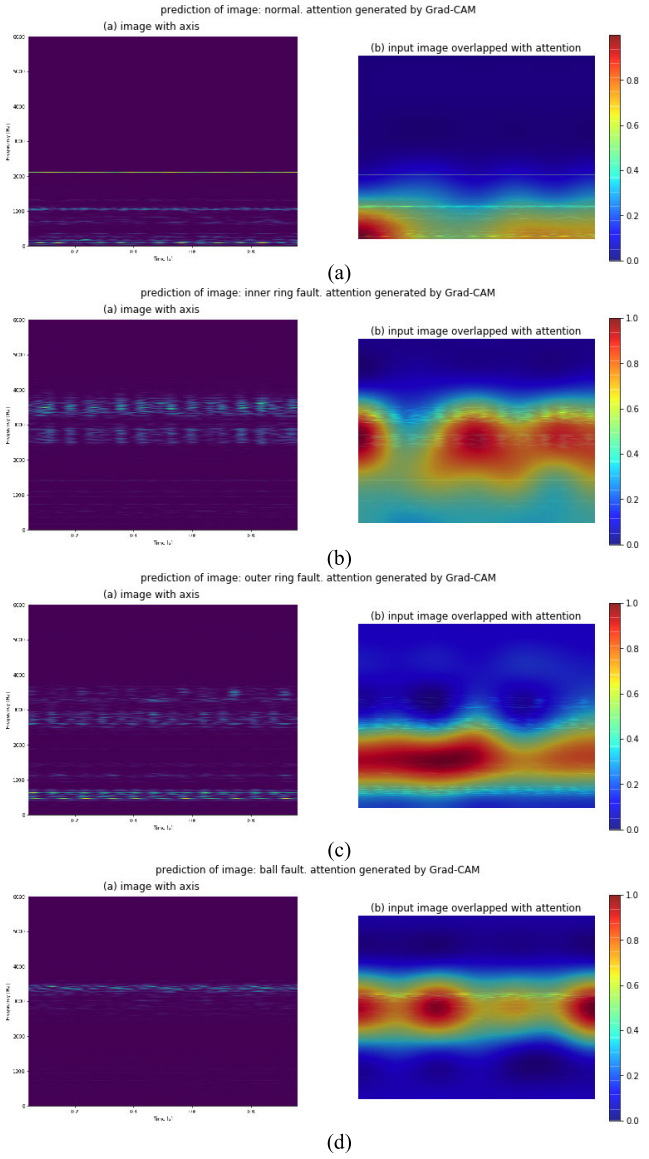

FIGURE 8. Input image with axis (left) and input image overlapped with attention map (right) of (a) a normal bearing, (b) a bearing with inner ring fault, (c) a bearing with outer ring fault, and (d) a bearing with ball fault under 1750 rpm.

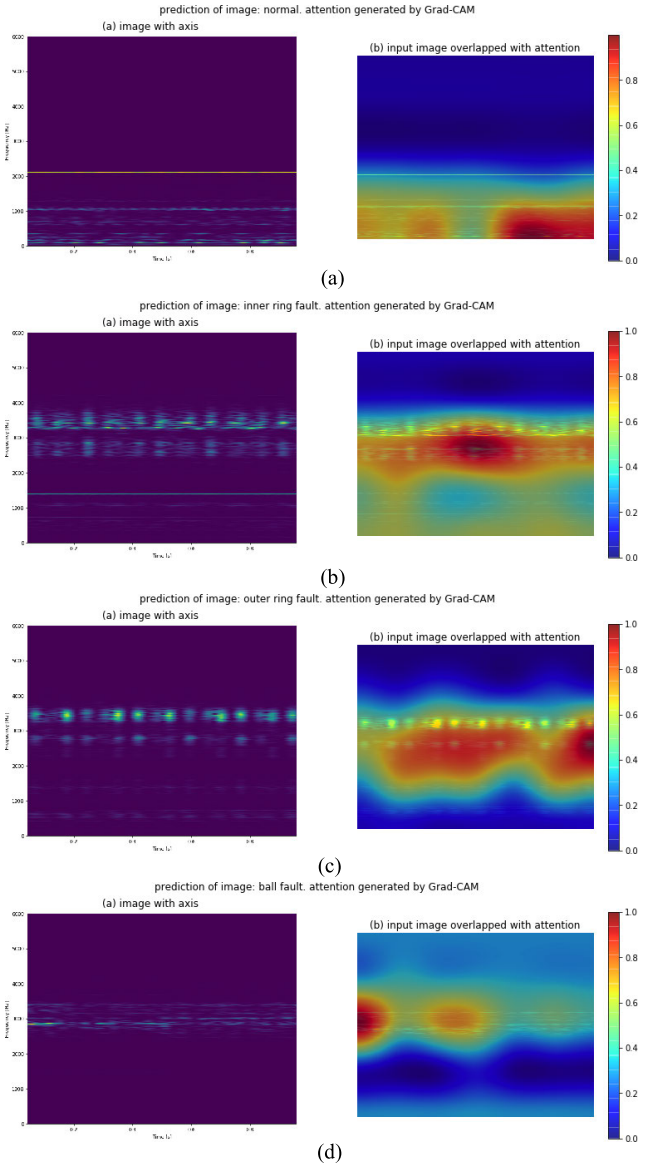

FIGURE 9. Input image with axis (left) and input image overlapped with attention map (right) of (a) a normal bearing, (b) a bearing with inner ring fault, (c) a bearing with outer ring fault, and (d) a bearing with ball fault under 1730 rpm.

Next, the features of high-frequency bands are sorted out and applied in classification. The high-frequency band is separated into 1000∼2000 Hz, 2000∼3000 Hz, and 3000∼4000 Hz. The eight statistical features are average magnitude and kurtosis in 1000∼2000 Hz, average magnitude, kurtosis, and skewness in 2000∼3000 Hz and 3000∼4000 Hz. The computation of kurtosis and skewness can be represented as

$$\text{kurtosis} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i-\overline{x})^4}{\sigma^4}, \tag{18}$$

$$\text{skewness} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i-\overline{x})^3}{\sigma^3}, \tag{19}$$

B. VERIFICATION USING NEURAL NETWORK

A simple neural network (NN) is set up for classification. The structure of NN is shown in TABLE 4. The learning rate is 0.008. The optimizer is Adam. Categorical cross-entropy is applied as loss function. The training and testing accuracy are both 100%. The confusion matrix of model is shown in FIGURE 11. The result shows that high-frequency bands also contain key features for classification. Though high-frequency bands are excited by structure resonance, different types of faults can generated different frequency distributions which can be used for classification and diagnosis. The differences in high-frequency bands are more obvious than low-frequency bands. Therefore, the model tends to focus on high-frequency bands instead of low-frequency bands.

C. VERIFICATION USING DECISION TREES

Decision tree is a simple algorithm mostly applied for classification (classification and regression tree, CART) [45],

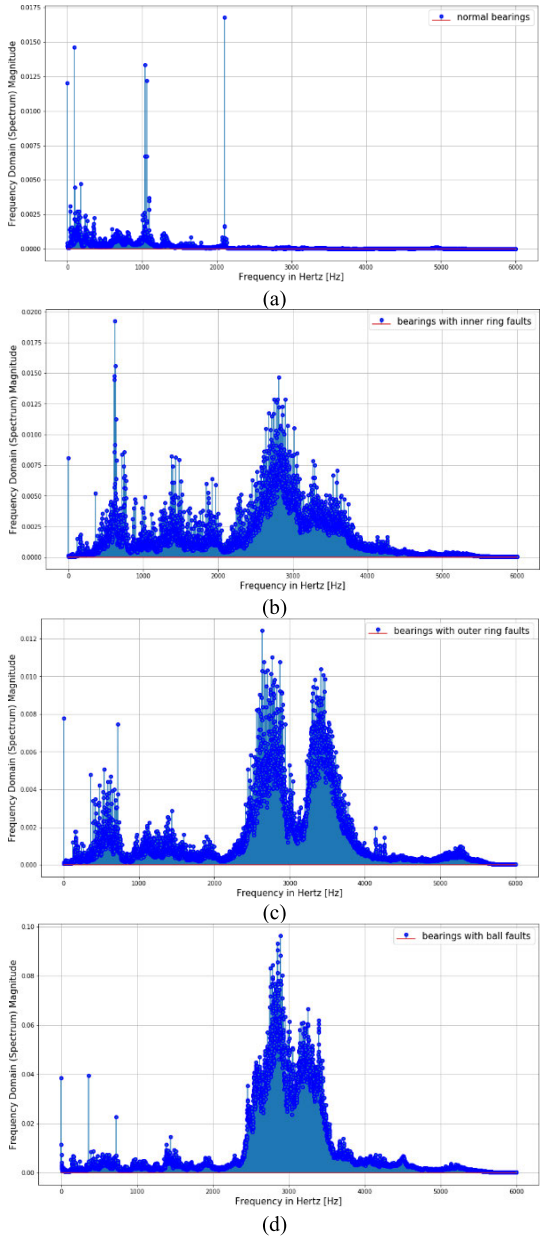

FIGURE 10. Average frequency spectra of (a) normal bearings, (b) bearings with inner ring faults, (c) bearings with outer ring faults, and (d) bearings with ball faults.

TABLE 4. Structure of NN for classifying bearing faults using features of high-frequency bands.

Layer	nodes	Activation function
Input layer	8	None
Hidden layer 1	10	Sigmoid
Hidden layer 2	10	Sigmoid
Output layer	4	Softmax

the basic structure of a tree contains nodes and branches. The nodes are divided into root node, internal nodes, and leaf nodes. The root node represents the start of the tree and contains entire dataset. The internal nodes, also known as decision nodes, are the conditions that can separate the dataset or subset into two subsets.

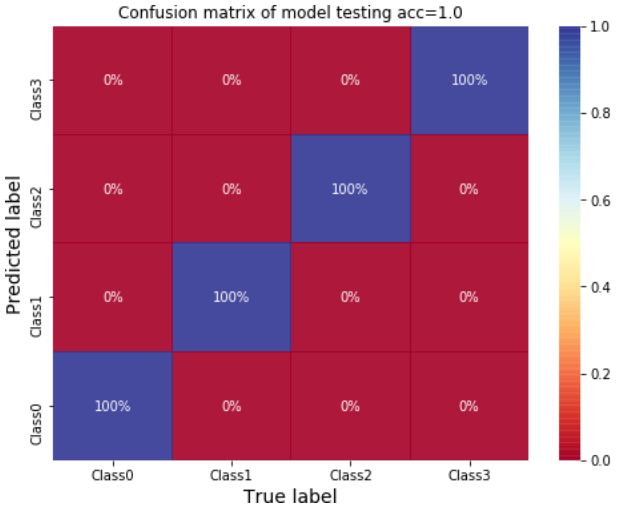

FIGURE 11. Confusion matrix of NN for classifying bearing faults using features of high-frequency bands.

In order to assess and choose the best decision, information gain is applied, including information entropy and Gini impurity. The information entropy can be represented as

$$\text{entropy} = \sum\nolimits_c p_c \log_2 p_c \tag{20}$$

while Gini impurity can be represented as

$$\text{Gini Impurity} = \sum\nolimits_c p_c(1-p_c) = \sum (p_c - p_c^2) = 1 - \sum p_c^2 \tag{21}$$

where p_c is the percentage of class c in the dataset. The target of decision is to maximize the separated information. Therefore, the information gain must be minimized. If summation of information gain in the subsets after the decision is smaller than other decisions, the decision is selected. The process will stop until all categories in data are separated completely.

Herein, decision tree is adopted for classification. The tree using entropy and Gini impurity as information gain are shown in FIGURE 12(a) and FIGURE 12(b), respectively. It takes 11 and 10 layers to complete the classification for each tree. By observing the decision tree, skewness of 3000∼4000 Hz is not applied in both decision trees. Therefore, a NN is used to check if the feature is not necessary for classification. The structure of NN is shown in TABLE 5. The learning rate, optimizer, and loss function are the same as the NN in B. The testing accuracy of model is 100%. The result shows that the skewness of 3000∼4000 Hz is not essential for classification using CWRU bearing dataset.

D. VERIFICATION USING ADAPTIVE NETWORK-BASED FUZZY INFERENCE SYSTEM (ANFIS)

Fuzzy inference system is based on the way human beings making decisions with imprecise and non-numerical information. It uses If-Then rules that are defined by dataset to predict or reach control purposes. ANFIS is a combination of fuzzy inference system and neural network. It has layer structure similar to neural networks but with the operations

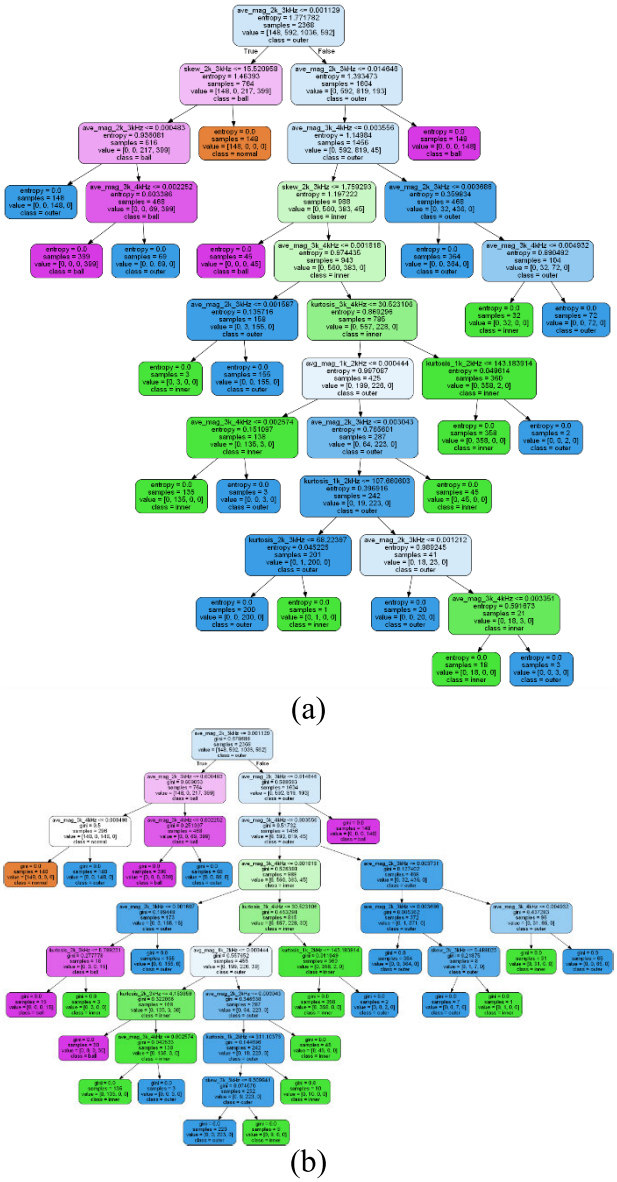

FIGURE 12. Decision tree for classification using features in high-frequency band. (Information gain: (a) entropy, (b) Gini impurity).

TABLE 5. Structure of NN for classifying bearing faults using features of high-frequency bands (without skewness of 3000∼4000 Hz).

Layer	nodes	Activation function
Input layer	7	None
Hidden layer 1	10	Sigmoid
Output layer	4	Softmax

of fuzzy inference system, e.g. rules layer, defuzzification layer. A first-order Sugeno-type ANFIS (TSK) is applied in this study and shown in FIGURE 13. In the training process of ANFIS, the membership functions and parameters in defuzzification layer are adjusted by training algorithm (backpropagation, least mean square (LMS)) according to If-Then rules.

The previous features in part A are inputs of ANFIS, the membership function utilized here is triangular

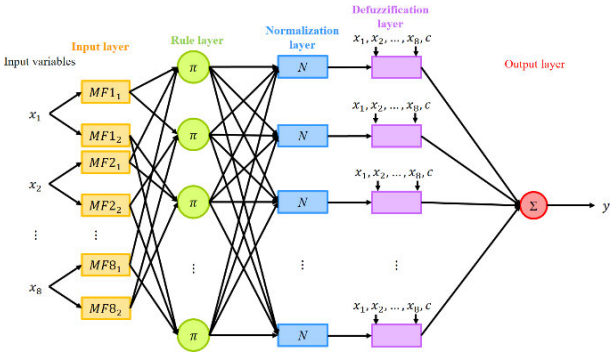

FIGURE 13. Structure of ANFIS.

TABLE 6. Confusion matrix of ANFIS using testing data.

		Actual class			
		Normal	Inner ring	Outer ring	Ball
Predicted class	Ball	0	0	2	174
	Outer ring	0	5	322	15
	Inner ring	0	157	0	0
	Normal	35	0	0	0

TABLE 7. Rules of ANFIS.

Index of Rules	[Avg. 1k∼2k Hz, Kur. 1k∼2k Hz, Avg. 2k∼3k Hz, Kur. 2k∼3k Hz, Skew. 2k∼3k Hz, Avg. 3k∼4k Hz, Kur. 3k∼4k Hz, Skew. 3k∼4k Hz]	Class Criteria
1	[Low, Low, Low, Low, Low, Low, Low, Low]	[122.8173, 6.2942, -92.7101, 3.7964, 272.4220, 55.0627, -157.9949, -1.7471]×**X** +4.0015
2	[Low, Low, Low, Low, Low, Low, Low, High]	[306.3599, -31.4746, 122.6128, -180.6154, -210.1118, -8.8727, 39.2566, -30.6]×**X** +17.6153
…		
256	[High, High, High, High, High, High, High, High]	[-0.0226, -0.0019, 0.0013, -0.0082, -0.0011, -0.0077, -0.0073, -0.0113]×**X** -0.0284

membership function. There are two membership functions for each input. After training, the classification accuracy using ANFIS is 96.9%. The confusion matrix of ANFIS using testing data is shown in TABLE 6. As above, there are total 256 rules shown in TABLE 7. The result also shows that the features in high-frequency band can be applied for classification even in a fuzzy system.

Herein, we try to find the relation between ANFIS rules and decision tree. Since all of the features need to be considered in ANFIS rules, a decision with more complete features is chosen. The selected decision is shown in FIGURE 14. Noting that the normalization operation is done for inputs and outputs. The prediction result using ANFIS is 3.08 which belongs to outer ring fault. The firing strength of partial rules are shown in FIGURE 15. As FIGURE 15 shows that the first rule has a larger firing strength. In other words, the first rule is more important while predicting the selected data. By observing the details of the first rule in TABLE 7, the features of chosen data match the membership functions, in which the features are all belong to low, of the first rule. The comparison shows that some of decisions match the rules

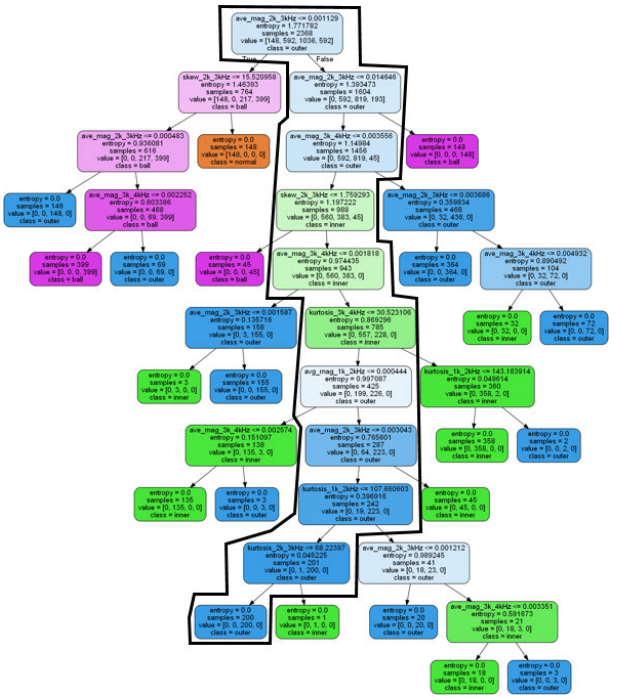

FIGURE 14. The chosen decision of the decision tree.

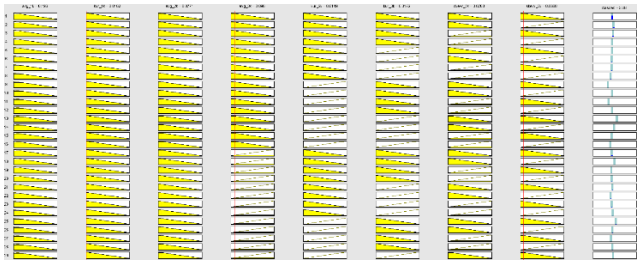

FIGURE 15. Firing strength of ANFIS rules using chosen data.

in ANFIS. However, since the decision tree does not need full features to complete the classification, not all of decisions have corresponding rules.

E. SUMMARY OF VERIFICATIONS FOR EXPLANATION OF CNN

From the verification results, the assumption by observing attention maps can become a correct explanation for CNNs in classifying CWRU bearings: the features in high-frequency band can be applied for classification more easily for the model instead of focusing at characteristic frequencies which are applied in most researches. The explanation is verified using simple NN models, ANFIS, and decision trees to increase the persuasive and correctness of explanation.

V. CONCLUSIONS

In this paper, XAI approach of CNNs in using vibration analysis is discussed using bearing faults classification. First, CNN for classifying bearing faults using time-frequency spectra is carried out. The results show that CNN can be applied for vibration analysis and provide great performance. Then, explanation for CNN is discussed. Grad-CAM is applied to help explaining model in vibration analysis. Attentions of CNN are computed and analyzed. The attentions show that the model pays more attention at high-frequency bands which are excited by structure resonance. Finally, an explanation for model is sorted out and verified using NN, ANFIS, and decision trees: The features in high-frequency band can be applied for classification more easily for machine learning than focusing on characteristics computed by traditional signal analysis.

REFERENCES

[1] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Appl. Acoust.*, vol. 158, Jan. 2020, Art. no. 107020.

[2] E. G. Plaza, P. J. N. López, and E. M. B. González, "Efficiency of vibration signal feature extraction for surface finish monitoring in CNC machining," *J. Manuf. Processes*, vol. 44, pp. 145–157, Aug. 2019.

[3] G. He, K. Ding, and H. Lin, "Fault feature extraction of rolling element bearings using sparse representation," *J. Sound Vib.*, vol. 366, pp. 514–527, Mar. 2016.

[4] R. Xiao, Q. Hu, and J. Li, "Leak detection of gas pipelines using acoustic signals based on wavelet transform and support vector machine," *Measurement*, vol. 146, pp. 479–489, Nov. 2019.

[5] Z. Ren, S. Zhou, Chunhui E, M. Gong, B. Li, and B. Wen, "Crack fault diagnosis of rotor systems using wavelet transforms," *Comput. Electr. Eng.*, vol. 45, pp. 33–41, Jul. 2015.

[6] T. Lambrou, P. Kudumakis, R. Speller, M. Sandler, and A. Linney, "Classification of audio signals using statistical features on time and wavelet transform domains," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, vol. 6, May 1998, pp. 3621–3624.

[7] N. Huang, "Ensemble empirical mode decomposition: A noise-assisted data analysis method," Center Ocean Land Atmos. Stud., Calverton, MD, USA, Tech. Rep. 85, Jan. 2006.

[8] A. V. Oppenheim, *Applications of Digital Signal Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 1978.

[9] B. Dolenc, P. Boškoski, and Đ. Juričić, "Distributed bearing fault diagnosis based on vibration analysis," *Mech. Syst. Signal Process.*, vols. 66–67, pp. 521–532, Jan. 2016.

[10] W. AL-Salman, Y. Li, and P. Wen, "K-complexes detection in EEG signals using fractal and frequency features coupled with an ensemble classification model," *Neuroscience*, vol. 422, pp. 119–133, Dec. 2019.

[11] S. Rukhsar, Y. U. Khan, O. Farooq, M. Sarfraz, and A. T. Khan, "Patient-specific epileptic seizure prediction in long-term scalp EEG signal using multivariate statistical process control," *IRBM*, vol. 40, no. 6, pp. 320–331, Dec. 2019.

[12] Y. Yang, W. Yang, and D. Jiang, "Simulation and experimental analysis of rolling element bearing fault in rotor-bearing-casing system," *Eng. Failure Anal.*, vol. 92, pp. 205–221, Oct. 2018.

[13] T. Wang, M. Liang, J. Li, and W. Cheng, "Rolling element bearing fault diagnosis via fault characteristic order (FCO) analysis," *Mech. Syst. Signal Process.*, vol. 45, no. 1, pp. 139–153, Mar. 2014.

[14] D. Zhao, T. Wang, R. X. Gao, and F. Chu, "Signal optimization based generalized demodulation transform for rolling bearing nonstationary fault characteristic extraction," *Mech. Syst. Signal Process.*, vol. 134, Dec. 2019, Art. no. 106297.

[15] Y. Liu, L. Guo, Q. Wang, G. An, M. Guo, and H. Lian, "Application to induction motor faults diagnosis of the amplitude recovery method combined with FFT," *Mech. Syst. Signal Process.*, vol. 24, no. 8, pp. 2961–2971, Nov. 2010.

[16] W. K. Lee, M. M. Ratnam, and Z. A. Ahmad, "Detection of chipping in ceramic cutting inserts from workpiece profile during turning using fast Fourier transform (FFT) and continuous wavelet transform (CWT)," *Precis. Eng.*, vol. 47, pp. 406–423, Jan. 2017.

[17] X. Yan and M. Jia, "A novel optimized SVM classification algorithm with multi-domain feature and its application to fault diagnosis of rolling bearing," *Neurocomputing*, vol. 313, pp. 47–64, Nov. 2018.

[18] C. Abdelkrim, M. S. Meridjet, N. Boutasseta, and L. Boulanouar, "Detection and classification of bearing faults in industrial geared motors using temporal features and adaptive neuro-fuzzy inference system," *Heliyon*, vol. 5, no. 8, Aug. 2019, Art. no. e02046.

[19] J. Liu, Z. Xu, L. Zhou, W. Yu, and Y. Shao, "A statistical feature investigation of the spalling propagation assessment for a ball bearing," *Mechanism Mach. Theory*, vol. 131, pp. 336–350, Jan. 2019.
[20] D. Cabrera, F. Sancho, C. Li, M. Cerrada, R.-V. Sánchez, F. Pacheco, and J. V. de Oliveira, "Automatic feature extraction of time-series applied to fault severity assessment of helical gearbox in stationary and non-stationary speed operation," *Appl. Soft Comput.*, vol. 58, pp. 53–64, Sep. 2017.
[21] C. Cintas, M. Lucena, J. M. Fuertes, C. Delrieux, P. Navarro, R. González-José, and M. Molinos, "Automatic feature extraction and classification of iberian ceramics based on deep convolutional networks," *J. Cultural Heritage*, vol. 41, pp. 106–112, Jan. 2020.
[22] O. Yildirim, M. Talo, B. Ay, U. B. Baloglu, G. Aydin, and U. R. Acharya, "Automated detection of diabetic subject using pre-trained 2D-CNN models with frequency spectrum images extracted from heart rate signals," *Comput. Biol. Med.*, vol. 113, Oct. 2019, Art. no. 103387.
[23] J. Zhang, Y. Sun, L. Guo, H. Gao, X. Hong, and H. Song, "A new bearing fault diagnosis method based on modified convolutional neural networks," *Chin. J. Aeronaut.*, vol. 33, no. 2, pp. 439–447, Feb. 2020.
[24] M. M. M. Islam and J.-M. Kim, "Motor bearing fault diagnosis using deep convolutional neural networks with 2D analysis of vibration signal," in *Advances in Artificial Intelligence*, Cham, Switzerland: Springer, 2018, pp. 144–155.
[25] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
[26] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Towards medical XAI," 2019, *arXiv:1907.07374*. [Online]. Available: http://arxiv.org/abs/1907.07374
[27] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2018, pp. 80–89.
[28] A. B. Tickle, R. Andrews, M. Golea, and J. Diederich, "The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks," *IEEE Trans. Neural Netw.*, vol. 9, no. 6, pp. 1057–1068, Nov. 1998.
[29] M. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 97–101.
[30] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.
[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[32] B. Kim, J. Gilmer, F. Viegas, U. Erlingsson, and M. Wattenberg, "TCAV: Relative concept importance testing with linear concept activation vectors," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Vancouver, BC, Canada, 2018.
[33] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," 2014, *arXiv:1403.6382*. [Online]. Available: http://arxiv.org/abs/1403.6382
[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," presented at the 31st Int. Conf. Neural Inf. Process. Syst., Long Beach, CA, USA, 2017.
[35] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[37] E. Sejdić, I. Djurović, and J. Jiang, "Time–frequency feature representation using energy concentration: An overview of recent advances," *Digit. Signal Process.*, vol. 19, no. 1, pp. 153–183, Jan. 2009.
[38] S. Wang, W. Huang, and Z. K. Zhu, "Transient modeling and parameter identification based on wavelet and correlation filtering for rotating machine fault diagnosis," *Mech. Syst. Signal Process.*, vol. 25, no. 4, pp. 1299–1320, May 2011.
[39] B. Li, P.-L. Zhang, D.-S. Liu, S.-S. Mi, G.-Q. Ren, and H. Tian, "Feature extraction for rolling element bearing fault diagnosis utilizing generalized s transform and two-dimensional non-negative matrix factorization," *J. Sound Vib.*, vol. 330, no. 10, pp. 2388–2399, May 2011.
[40] X. Li, J. Ma, X. Wang, J. Wu, and Z. Li, "An improved local mean decomposition method based on improved composite interpolation envelope and its application in bearing fault feature extraction," *ISA Trans.*, vol. 97, pp. 365–383, Feb. 2020.
[41] X. Li, Y. Yang, H. Pan, J. Cheng, and J. Cheng, "A novel deep stacking least squares support vector machine for rolling bearing fault diagnosis," *Comput. Ind.*, vol. 110, pp. 36–47, Sep. 2019.
[42] W. A. Smith and R. B. Randall, "Rolling element bearing diagnostics using the case western reserve university data: A benchmark study," *Mech. Syst. Signal Process.*, vols. 64–65, pp. 100–131, Dec. 2015.
[43] G. Zhang, Y. Zhang, T. Zhang, and R. Mdsohel, "Stochastic resonance in an asymmetric bistable system driven by multiplicative and additive Gaussian noise and its application in bearing fault detection," *Chin. J. Phys.*, vol. 56, no. 3, pp. 1173–1186, Jun. 2018.
[44] Q. He, J. Wang, Y. Liu, D. Dai, and F. Kong, "Multiscale noise tuning of stochastic resonance for enhanced fault diagnosis in rotating machines," *Mech. Syst. Signal Process.*, vol. 28, pp. 443–457, Apr. 2012.
[45] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Boca Raton, FL, USA: CRC Press, 1984.

HAN-YUN CHEN was born in Taichung, Taiwan, in 1996. He received the B.S. degree in mechanical engineering, in 2018, and the M.S. degree in mechanical engineering from National Chung Hsing University, Taichung, in 2020.

He has published an article on energy reports. His research interests include the applications of artificial intelligence, convolutional neural networks, fuzzy control, and optimization algorithms.

CHING-HUNG LEE (Senior Member, IEEE) was born in Taiwan, in 1969. He received the B.S. and M.S. degrees from the Department of Control Engineering, National Chiao Tung University, Hsinchu, Taiwan, in 1992 and 1994, respectively, and the Ph.D. degree from the Department of Electrical and Control Engineering, National Chiao Tung University, in 2000. He is currently a Distinguished Professor with the Department of Mechanical Engineering, National Chung Hsing University, Taichung, Taiwan. His main research interests include artificial intelligence, fuzzy neural systems, neural networks, signal processing, nonlinear control systems, robotics control, and CNC motion control and optimization. He received the Wu Ta-Yu Medal (Young Researcher Award) from the Ministry of Science and Technology, Taiwan, in 2008. He also awarded the Youth, Excellent Automatic Control Engineering, and Fellow Awards from the Chinese Automatic Control Society, Taiwan, in 2009, 2016, and 2019, respectively.