IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 1, 2020, accepted July 13, 2020, date of publication July 16, 2020, date of current version July 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3009852

Vision-Based Fault Diagnostics Using Explainable Deep Learning With Class Activation Maps

KYUNG HO SUN[1], HYUNSUK HUH[2], BAYU ADHI TAMA[2], SOO YOUNG LEE[2], JOON HA JUNG[1], AND SEUNGCHUL LEE[2,3,4]

[1]Korea Institute of Machinery and Materials, Daejeon 34103, South Korea
[2]Department of Mechanical Engineering, Pohang University of Science and Technology, Pohang 37673, South Korea
[3]Graduate School of Artificial Intelligence, Pohang University of Science and Technology, Pohang 37673, South Korea
[4]Institute for Convergence Research and Education in Advanced Technology, Yonsei University, Seoul 03722, South Korea

Corresponding author: Seungchul Lee (seunglee@postech.ac.kr)

This work was supported in part by the Korea Institute of Machinery and Materials, in part by the National Research Foundation (NRF) under Grant 2020M2A8A4025171 and Grant 2020R1A2C1009744, and in part by the High-Potential Individuals Global Training Program of Institute for Information and Communications Technology Planning and Evaluation (IITP) under Grant 2019-0-01589.

ABSTRACT In the era of the fourth industrial revolution (Industry 4.0) and the Internet of Things (IoT), real-time data is enormously collected and analyzed from mechanical equipment. By classifying and characterizing the measured signals, the fault condition of mechanical components could be identified. However, most current health monitoring techniques utilize time-consuming and labor-intensive feature engineering, i.e., feature extraction and selection, that are carried out by experts. This paper, on the contrary, deals with an automatic diagnosis method of machine monitoring using a convolutional neural network (CNN) with class activation maps (CAM). A class activation map enables us to discriminate the fault region in the images, thus allowing us to localize the fault precisely. The goal of the paper is to demonstrate how CNN and CAM could be employed to real-world vibration video to characterize the machine's status, representing normal or fault conditions. The performance of the proposed model is validated with a base-excited cantilever beam dataset and a water pump dataset. This paper presents a novel industrial application by developing a promising method for automatic machine condition-based monitoring.

INDEX TERMS Convolutional neural network, class activation maps, discriminative region, fault detection, mechanical component, explainable AI.

I. INTRODUCTION

Fault detection in mechanical equipment has been a serious concern in many industries. Most of the faults are likely to occur during operating conditions, hence making them a significant impact in increasing operational cost. Moreover, unpredicted faults would make machine failure or even put a person in safety jeopardy. Once a mechanical component possesses a fault, the fault would shortly trigger a chain reaction and cause the damage of other components [1]. Understanding the condition of a machine and its components is indispensable. It would keep us away from termination and unexpected costs, gain the machine's lifespan, and enhance safety by identifying irregular behavior of a machine and its components.

Both normal and abnormal operations of mechanical systems that are made up of rotating components could be examined using vibration signal analysis. Hence a vibration analysis plays a significant function in monitoring the condition of the machine and diagnosing its failure. The analysis is able to be conducted to examine various sorts of systems comprised of multiple mechanical components, e.g., gears and bearings, by measuring vibration levels from sensors such as accelerometers. In addition, analyzing vibration signals may refer to acquiring raw data from the device and characterizing substantial properties that are highly dependent on defects. It could be done in the time, frequency, and time-frequency domain depending on occasion and purpose. For instance, Hong and Dhupia [2] proposed a time domain approach to diagnose gearbox fault based on measured vibration signals. At the same time, Gonçalves et al. [3] provided a comparative analysis of vibration signals to diagnose bearing faults.

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

However, there exist several limitations in prior researches on fault diagnosis of vibration signals for mechanical systems. First of all, issues on the acquisition of raw data from sensors have been arisen for applying vibration analysis to today's highly complicated systems. Such a complex system that requires more precise and accurate analysis needs a number of sensors. Therefore, there are excessive costs to investigate all amounts of data. Placing sensors to every region of the system is not merely unrealistic, but also impractical. Besides, with the signals obtained by sensors deployed anywhere in the system, appropriate analysis is not possible to be put into practice. In addition, considering the fact that most of the mechanical equipment includes signals that are not essential for diagnosis, i.e., internal interference and external disturbance, it is not straightforward for engineers to interpret the results. While some researchers, for this reason, provide methodologies for optimizing location and amount of the sensors in various mechanical systems [4], [5], there are still uncertainties of additional efforts and costs for practical applications as diverse types of mechanical equipment are continuously emerging.

Secondly, issues on features from raw signals and those diagnosis techniques have emerged for outstanding performance of fault diagnosis. In traditional fault diagnosis techniques, lots of attempts are performed based on the manual design of fault features. The main drawback of the manual design of features is that it highly depends upon an expert with specific knowledge about mechanical engineering or their domains, which is able to define the features clearly for appropriate classification tasks. Consequently, the techniques are less automatic, going through a series of procedures stated as followed. A feature set is typically extracted through processed measurements, whether collected from hand-operated devices or built-in sensors. This feature set is then pre-processed by some experts that need to be elaborated to represent the machine's behavior. Next, the extracted feature set is given to a machine learning algorithm, along with the labels [6], [7]. The machine learning algorithm will learn a model that is able to classify machine conditions. Hence, as the features are extracted manually with specific kinds of expertise in engineering background, then fed up to the classification algorithm, well-designed features could dominate the performance of the classifier, and it leads the most challenging task for fault diagnosis techniques.

In order to overcome the aforesaid issues, this paper proposes an image-based technique for enhancing discriminative efficiency between two types of fault conditions (normal and fault) using a convolutional neural network (CNN). Feature learning, as well as its localization ability via class activation maps (CAMs) algorithm, is also proposed to detect the most significant part for diagnosing mechanical equipment. We utilize the interpretability of a CAM-embedded deep learning structure to indirectly localize the most vibrating regime without any vibration sensors. Furthermore, we validate the performance of the proposed model by comparing it with several machine learning algorithms, i.e., artificial neural network (ANN), random forest (RF) [8], and support vector machine (SVM) [9]. Lastly, a vision-based significant region detection that is able to be emphasized by the CAM algorithm is also verified.

The rest of the paper is broken down into the following parts. Section II-A briefly explains feature engineering approaches in image recognition, while Section II-B presents a brief review of current existing fault diagnosis techniques. Section III conveys the data set collection and preparation, as well as classification techniques and validation measures used in our experiment of vision-based fault diagnostics. Section IV is dedicated to the experimental results, followed by discussions and remarks. Lastly, Section V concludes the paper.

II. RELATED WORK

Before introducing the image-based fault diagnosis technique, we briefly review feature engineering of image recognition and deep learning for machine fault diagnosis in this section.

A. AN OVERVIEW OF FEATURE ENGINEERING FOR IMAGE RECOGNITION

With the rapid development of high-speed detection techniques, automatic detectors have enabled efficient and accurate image detection, taking from tens up to thousands of images per second. The development of detection tool for machine health diagnosis could lower processing time to almost zero, offer precise, accurate, and unbiased detection. It provides reliable results, thus making it more reproducible, scalable, and robust using available state-of-the-art computing resources.

Early machine learning algorithms for image recognition and visual object detection mostly drew upon a boosted cascade of simple features [10], [11], and a feature descriptor called Histograms of Oriented Gradients (HOG) [12] to extract the image features, prior to being used as inputs of a classification algorithm. The two above-mentioned approaches had been the most remarkable demonstration of computer vision at that time. Their techniques significantly outperformed existing algorithms for pedestrian detection. Nevertheless, most approaches are considered in collaboration with manually extracted features by experts [13], [14]. Hitherto, deep learning algorithms have become the mainstream of computer vision, though they had been everywhere for a long time. In particular, convolution neural networks (CNNs) [15] have metamorphosed into today's promising approach of image recognition tasks [16]–[18].

A CNN is a type of artificial neural network that carries out convolution directly on a data that has a grid pattern, e.g., images, and it is aimed to be able to automatically and adaptively learn spatial information of image features, from low to high-level patterns [15]. It is typically made up of multiple duplicating layers, i.e., convolution, pooling, and followed by fully connected layers [19]. CNN is an actively growing field, and new CNN architectures are continually being developed and applied in many applications, i.e., smart manufacturing [20], radiology [21], system health management [22], and to name a few.

Imbued by the aforementioned results, herein, a CNN model with class activation maps (CAMs) is developed to determine which parts of the image the model is focusing on. A traditional CNN commonly behaves like an image detector, yet the localizability of the discriminative region is not possible to be provided [23]. On the other hands, CAMs are the techniques to obtain the discriminative image regions used by CNN to recognize a specific class in the image. Roughly speaking, CAMs help us to see which regions in the images are relevant to particular classes. It is capable of interpreting the results of the classification. A detailed discussion about CAMs is presented in Section III. To the best of our knowledge, this is the first attempt to apply CAMs in fault detection research, which is currently lacking in the existing literature.

B. DEEP LEARNING FOR MACHINE FAULT DIAGNOSIS USING VIBRATION DATA

For many decades, researchers have been fascinated by neural networks (NNs). However, conventional NNs have been used by blending with manual pre-defined feature extraction. For instance, Rafiee *et al.* [24] used a multi-layer perceptron NN to predict gears and bearings faults in a gearbox system. However, the high performance of NN enormously relied upon a wavelet-based feature extraction that adopts the standard deviation of the wavelet packet coefficient [25]. A feature representation of multiple frequency resolutions for faulty modes was obtained using wavelet packet decomposition. In order to avoid very large-scale NNs and to decrease the training time, instead of directly using raw input data, feature vectors were specified in fixed dimensions prior to being used as input sets for the NNs.

Bin *et al.* [26] proposed wavelet packet decomposition to extract fault features and NNs for rotating machinery early fault diagnosis. Extracted features were taken as the target input on NN, whilst the 10 types of representative rotor failures label were taken as the output of NN. Lastly, a comparative study between NN and support vector machine (SVM) for fault diagnosis of the rotor-bearing system was carried out by Kankar *et al.* [27]. A statistical method was used to extract features and the dimensionality of original vibration features. From the experiment, it was evidenced that the classification accuracy of NN slightly outperforms SVM.

From the above-listed review, it can be observed that NNs are one of the most frequently employed classification algorithms in the purview of fault diagnosis. The NNs-based techniques presented in current literature typically possess two-fold strides, i.e., a less automatic feature extraction that comprises transforming of measured signals using signal processing techniques and then fault classification using NNs. Besides having manual design features, the NNs architecture commonly used has a shallow model, thus restricting the capability of NNs to learn more complex problems in fault diagnosis.

Deep learning, unlike classical NNs, uses an unrefined representation of the input and allows the algorithm to construct and learn the corresponding representation of the data, e.g., features. Such a technique hereinafter referred to as feature learning (FL). Following this, the work of Amar *et al.* [28] used an automatic feature extraction from vibration spectrum images for bearing fault classification based on NN. A new infrared thermal image-based machine health monitoring using deep learning model was also provided in [29]. In the paper, deep learning was applied to infrared thermal video to detect the condition of the machine. By employing the method, a rotating machine condition was able to be detected accurately.

In the past few years, there has been increasing research interest in fault detection techniques using vibration signals and deep learning [30]–[36]. For instance, Li *et al.* [37] proposed augmented deep sparse autoencoder to diagnose the gear pitting condition using raw signal of vibration. Classification of bearing faults using CNN and vibration spectrum imaging was studied in [38], where temporal vibration signals were extracted using a time-moving segmentation window. In addition, a review of condition monitoring and fault diagnosis of wind turbine planetary gearbox is presented in [39]. The paper reported a pertinent state-of-the-art review, pointed out valuable open research problems, and suggested potential research directions. Wang *et al.* [40] considered a new fault detection technique for rotating machinery based on fusion of multi-vibration-signals and layer optimized CNN. The proposed model was validated on two practical examples, i.e. wind power and centrifugal pump test rigs. A condition monitoring of cantilever vibrating beams was introduced in [41]. The study employed deep learning classifier to recognize a damaged and undamaged beam via time-frequency extended signatures.

However, most current existing methods are 'black box' models which do not reveal their internal mechanism. Consequently, it is harder for human to comprehend why particular predictions have been made. In order to make a distinction with existing approaches, this paper extends a new application of fault diagnosis using deep learning model with CAMs. The proposed method is able to localize the fault instantly, enabling fast and robust fault detection in a real-world industrial setting. More specifically, this study emphasizes a 'white box' model [42] for fault diagnosis that make the predictions understandable to human. The prediction model is explainable and understandable since it helps human deal with the opacity of deep learning models.

III. MATERIAL AND METHODS

In this section, we briefly explain the data set, experimental equipment, and model structure used in this paper. The explanation of validating our proposed method with two different kinds of a mechanical system is generally described in this section.

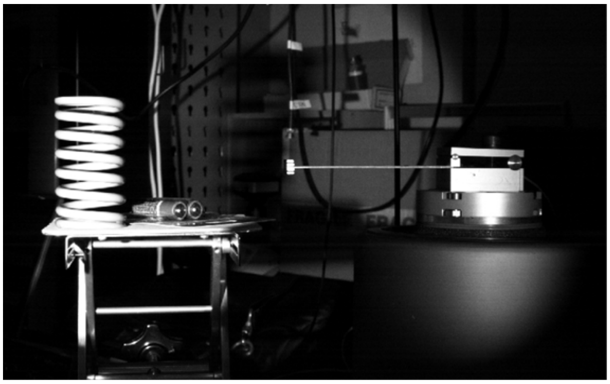

FIGURE 1. Acquired image data of a base-excited cantilever beam and a spring.

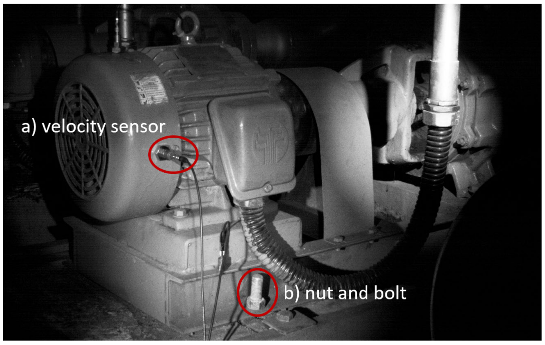

FIGURE 2. Water pump used in the experiment: a) velocity sensor, b) nut and bolt.

A. DATA COLLECTION AND VALIDATION TECHNIQUE

An image data set is extracted from the video, representing normal and fault behavior of a vibrating mechanical system, as shown in Figure 1. As a video is typically made up of multiple frames (e.g., images), we are able to treat it as a collection of images used as inputs of a deep learning model. For this experiment, two videos were recorded for 10 seconds by the high-speed camera (e.g., Phantom Miro C110) that has a resolution of 1280 × 800, a frame rate of 500 fps each, and in grayscale mode. We have collected two recordings about the machine's condition containing 5000 images at each normal and faulty operations, respectively. The distribution of each class in the samples is equally balanced, in which the ratio between normal and fault is 1:1. Besides, the experiment circumstances such as flow rate in pump, lighting, camera location, etc. were fixed, except the nut for excitation and base-excited cantilever acceleration.

Furthermore, the scores are specified during k-fold cross validation (k-fcv). This implies that the original samples are randomly split into k equal sized subsamples, in which the individual 1 subsample is used for testing, and the rest $k-1$ is used to train the CNN model. This is done k times so that every individual subsample is used as a test set once. In the experiment, we choose k=10, which is also known as 10-fold cross validation (10-fcv). The performance result reported in this paper is the average value of over 10 elements.

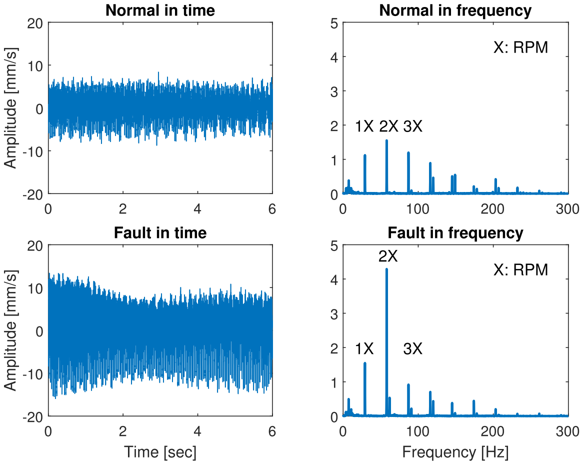

FIGURE 3. Measured signals.

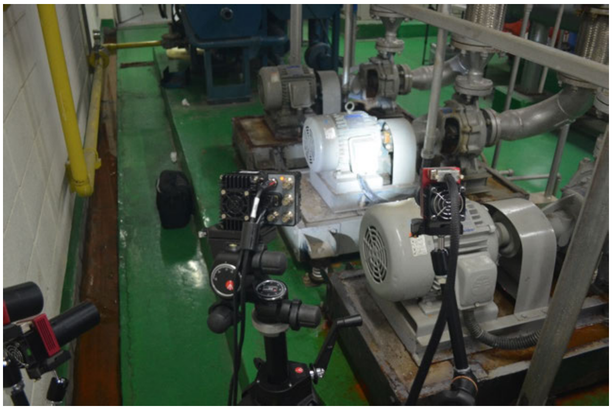

FIGURE 4. Water pump and camera setting.

B. BASE-EXCITED CANTILEVER BEAM

In the first case of the experiment, a base-excited cantilever beam is used to be validated by our proposed method as a way of imitating the actual mechanical system. Also, a spring next to the cantilever beam which has a little dependency on vibration, is placed on purpose, and it is indirectly affected by excitation. Image data set is acquired with two classes; one is excited by 0.3g of gravitational acceleration while the other is gravitationally accelerated with 0.03g. Figure 1 shows the cantilever beam and the exciter used in this experiment.

C. WATER PUMP

The experiment in the case of a water pump is also conducted to verify the feasibility of our proposed method for real mechanical equipment that is comprised of a large number of its components. The image data set is collected during the operation in a real industrial environment. Figure 2 shows the pump vibrated by rotating machinery. Data of the operating water pump are separated into two classes, i.e., normal and fault, under the looseness by manipulating the nut of the region marked in red, as shown in Figure 2. The velocities are measured both at normal and faulty conditions, and plotted in Figure 3. It clearly shows that looseness causes the misalignment of the rotating shaft, resulting in a large

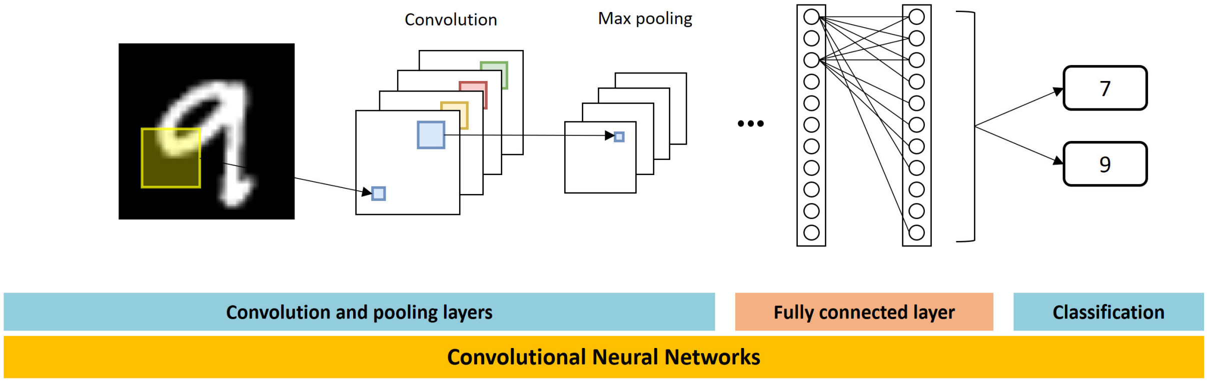

FIGURE 5. An example architecture of conventional CNN.

amplitude in time and an increased peak at the 2X frequency component [43]. Figure 4 shows the pump and camera setting.

D. CLASS ACTIVATION MAPS

The conventional CNNs can be conceptually divided into two sections. One section is feature extraction, and the other is classification. In the feature extraction process, a convolution layer is used to extract the features from the rough input data, prior to being used for classification tasks [15]. The classification task classifies which class each input data belongs to by taking into consideration the extracted features from the raw input data. When we visually identify the images, we do not look at the whole image; instead, we intuitively focus on the essential parts of the image. CNN's learning is almost identical to the way a human does. When its weights are optimized, the more critical parts are given higher weights. However, generally, we are not able to recognize this because the CNN goes through a fully connected layer and makes the features extracted by the convolution layer more abstract (see Figure 5). As CNNs are specified to deal with images, several properties have been attached to give not only a faster training, but also less training parameters. These properties are briefly discussed as follows.

- *Local connectivity*. When dealing with an image as input, each pixel value is not connected to every neuron in the first layer. Instead, each neuron obtains input from a small local group of the pixels in the input image.
- *Weight sharing*. It is also known as filter or kernel. Weights are a grid structure and serve on a particular small section of the image. Weights in CNN extract features from the input and form a feature map. Through weight sharing, the same features from the input are extracted in different locations of the input. Furthermore, weight sharing improves learning efficiency by dramatically lowering the number of the parameter being learned.
- *Pooling*. After implementing a convolutional layer, pooling is done. The goal is to reduce the spatial size of the feature maps, thus reducing the number of parameters, as well as controlling the over-fitting.

A class activation map (CAM) is a method employed to identify the most noticeable regions that help CNN to predict a particular class [23]. In CAM, a global average pooling (GAP) layer is placed in the network right after the final convolutional layer [44]. It possesses two main characteristics: (i) shed light on how it explicitly enables the convolutional neural network to have remarkable localization ability, and (ii) the heat-map is the class activation map, highlighting the importance of the image region for the prediction.

The deep learning model is a black-box model. When input data is received, a classification result of 1 or 0 is simply returned for the binary classification problem, without knowing how the classification results are derived. It is also possible for CAM to perform multi-class classification problem. As opposed to conventional deep learning models, a CAM is able to interpret the classification results. It is possible to estimate the localization of the object within an image. Through an analysis of which part of the image the model is focusing on, we are able to interpret which part of the image is considerably important. In addition, a CAM is a modified convolution layer. It directly highlights the salient parts of the spatial grid of an image. Therefore, it offers information about where the emphasized regions of the model [23]. Figure 6 illustrates the procedure of how class activation mappings are generated.

The feature maps of the last convolution layer can be interpreted as a collection of spatial locations focused on by the model. The CAM can be obtained by taking a linear sum of the features. They all have different weights and thus can obtain spatial locations according to various input images through a linear combination. More formally, let a given image, $f_k(x, y)$ denotes the feature map of unit k in the last convolution layer at a spatial location (x, y). For a given class c, the class score $\mathcal{S}_c$ is expressed as the following equation.

$$\mathcal{S}_c = \sum_k \omega_k^c \sum_{x,y} f_k(x, y) = \sum_{x,y} \sum_k \omega_k^c f_k(x, y) \tag{1}$$

where ω_k^c is the weight corresponding to class c for unit k. The class activation map for class c is depicted as $\mathcal{M}_c$.

$$\mathcal{M}_c(x, y) = \sum_k \omega_k^c f_k(x, y) \tag{2}$$

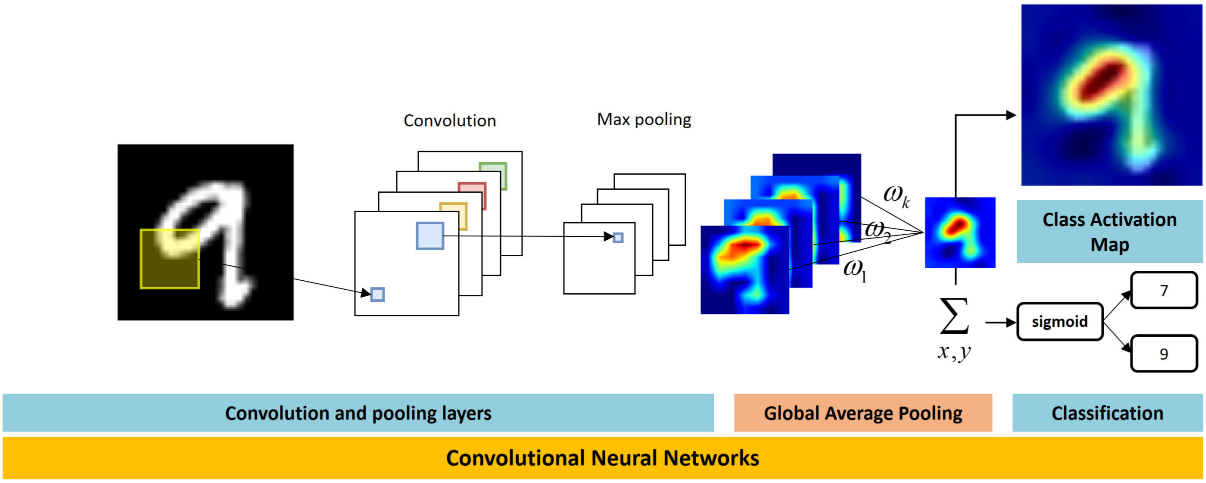

FIGURE 6. An example of CNN architecture with CAM.

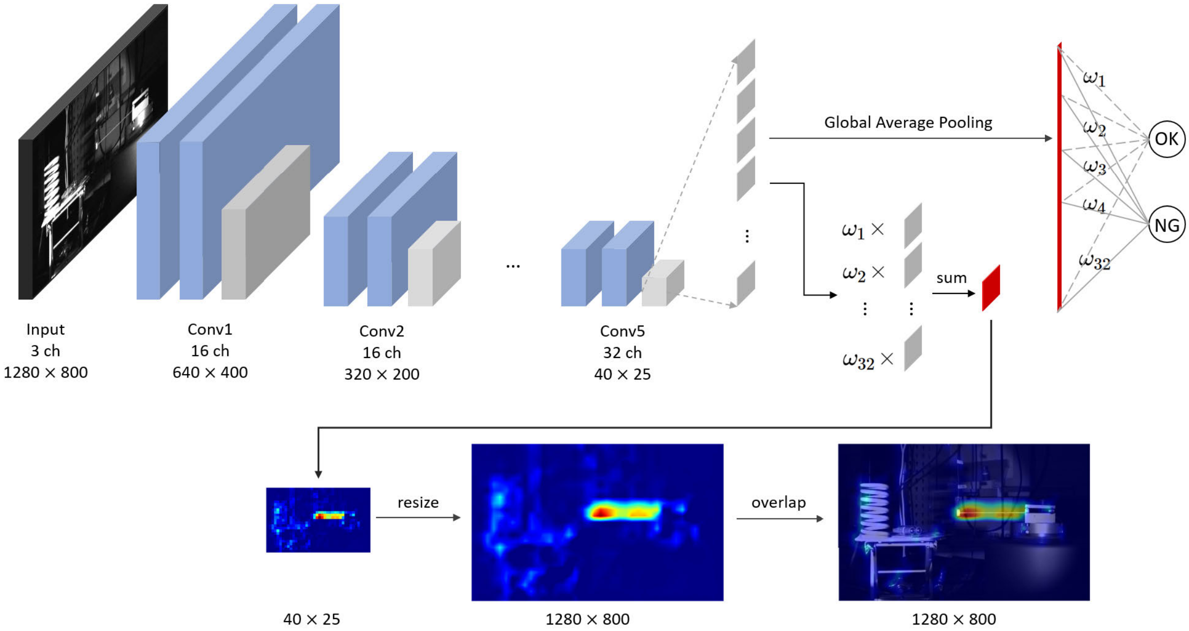

FIGURE 7. Architecture of CAM for machine health diagnosis.

The $\mathcal{M}_c$ directly indicates the importance of the feature map at a spatial grid (x, y) of class c. Finally, the output of the softmax for class c is defined as follows.

$$\mathcal{P}_c = \frac{exp(\mathcal{S}_c)}{\sum_c exp(\mathcal{S}_c)} \tag{3}$$

In the case of CNN, the size of the feature map is reduced by the pooling layer. By simple up-sampling, it is possible to identify the region of interests (ROIs) that most relevant for the particular class.

E. NETWORK ARCHITECTURE

The details of CAM architecture used in this work are visualized in Figure 7 and summarized in Table 1. The network is made up of 10 convolutional layers, followed by rectified linear (ReLU) as a nonlinear function at each layer. In addition, a pooling layer is appended between consecutive layers in order to improve the detection performance. The output layer consists of 2 neurons, representing the two different condition classes, i.e., normal and fault. Zero-padding is utilized in order to maintain the spatial dimensions of the feature maps unaltered throughout the model. The training is initialized with a learning rate of 0.0001, whilst an optimization function, e.g., *AdamOptimizer* in *TensorFlow* is also applied to optimize the objective functions. The model is trained for 5000 epochs and 25 mini-batches.

F. PERFORMANCE METRICS

Some performance measures, i.e., accuracy, precision, and recall, are used for machine health detection performance.

TABLE 1. Proposed architecture of CNN with CAMs.

Layer	Type	Channels	Kernel size	Stride	Layer	Type	Channels	Kernel size	Stride
0	Input	3	-	-	9	2D Maxpooling	-	2×2	2
1	2D Convolution	16	3×3	1	10	2D Convolution	32	3×3	1
2	2D Convolution	16	3×3	1	11	2D Convolution	32	3×3	1
3	2D Maxpooling	-	2×2	2	12	2D Maxpooling	-	2×2	2
4	2D Convolution	16	3×3	1	13	2D Convolution	32	3×3	1
5	2D Convolution	16	3×3	1	14	2D Convolution	32	3×3	1
6	2D Maxpooling	-	2×2	2	15	2D Maxpooling	-	2×2	2
7	2D Convolution	16	3×3	1	16	GAP[a]	32	-	-
8	2D Convolution	16	3×3	1	17	Linear	2	-	-

[a] Global average pooling

TABLE 2. Contingency table.

		Actual condition	
		Normal	Fault
Predicted condition	Normal	TP	FP
	Fault	FN	TN

TABLE 3. Training parameters of the three machine learning algorithms.

Multilayer perceptron	Random forest	Linear SVM
#hidden nodes: 300	Split criterion: gini	C: 1.0
#input nodes: 788	Min_samples_leaf: 1	Max_iterations: 1000
#output nodes: 2	#trees: 100	Optimization: dual
#iterations: 2500	Min_samples_split: 2	Tolerance: 0.0001

Accuracy denotes the ratio between the number of image samples that are correctly classified and all the image samples in total. The precision metric indicates the ratio between normal image samples that are correctly classified and all the normal image samples in total. It is also known as the percentage of correct classification predictions in all the machine-label conditions. Furthermore, we also calculate the performance of classifier algorithms in terms of recall metric, which represents the percentage of correct classification predictions in all the human-label conditions. By referring to a contingency matrix in Table 2, the aforementioned metrics can be calculated as follows.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

IV. RESULT AND DISCUSSION

This paper proposes that CNN, an image-based deep learning algorithm, can be used for fault diagnosis as well as image recognition. The feasibility of CNN & CAM algorithm for image-based fault diagnosis is described in the experiment of cantilever beam case. Also, we demonstrate the possibility of applying the proposed method technique with a more complicated mechanical system, herein water pump case, for fault diagnosis in the actual industrial field. Dataset is divided into training and test set at a ratio of 7:3 in both experiments. The model is trained within the deep learning model package from *TensorFlow* on a single TITAN X Pascal GPU. The architecture of the network is carefully tuned by modifying the number of layers, channels, sizes of kernels and strides in each layer in order to obtain the best classification model for fault diagnosis tasks, as listed in Table 1.

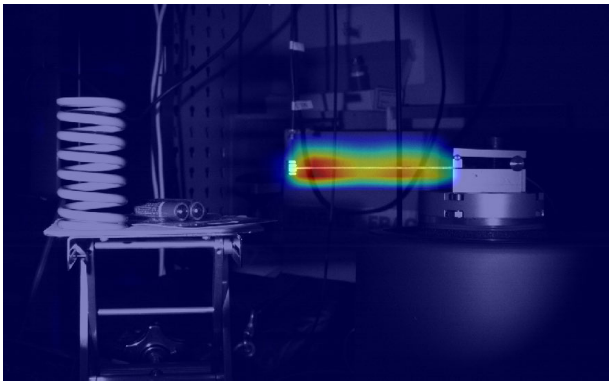

FIGURE 8. The CAM visualizes the most vibrating regime without any vibration sensors.

A. BASE-EXCITED CANTILEVER BEAM

Table 4 conveys the effectiveness of our proposed deep learning model, compared to other shallow machine learning algorithms, i.e., multilayer perceptron (MLP), random forest (RF) and linear support vector machine (SVM). As most of machine learning algorithms require pre-designed feature representation, the features are firstly extracted and then fed into the algorithms. Particularly, the extraction techniques for image processing could be used with global feature extractor (FE) approaches, i.e., Hu moments [45], Haralick texture [46], and color histogram [47] which have been proved to be successful in a wide variety of computer vision tasks such as object detection and image classification. 788 features are totally gathered and concatenated by considering the combination of 7 features for Hu moments, 13 features for Haralick textures, and 768 features for a color histogram that are able to quantify and emphasize shape, texture, and color of the images, respectively.

After extracting, concatenating, and saving the features and labels from the training data set, the above-mentioned shallow machine learning algorithms are taken to create the classi-

TABLE 4. Performance of classifiers for base-excited cantilever beam experiment.

Classifier	Feature extraction method	Accuracy (%)	Precision (%)	Recall (%)
CNN with CAM	FL	95.85	100	92.33
MLP	FE	49.35	100	50.00
RF	FE	65.20	60.00	66.29
SVM_{linear}	FE	49.35	100	49.35

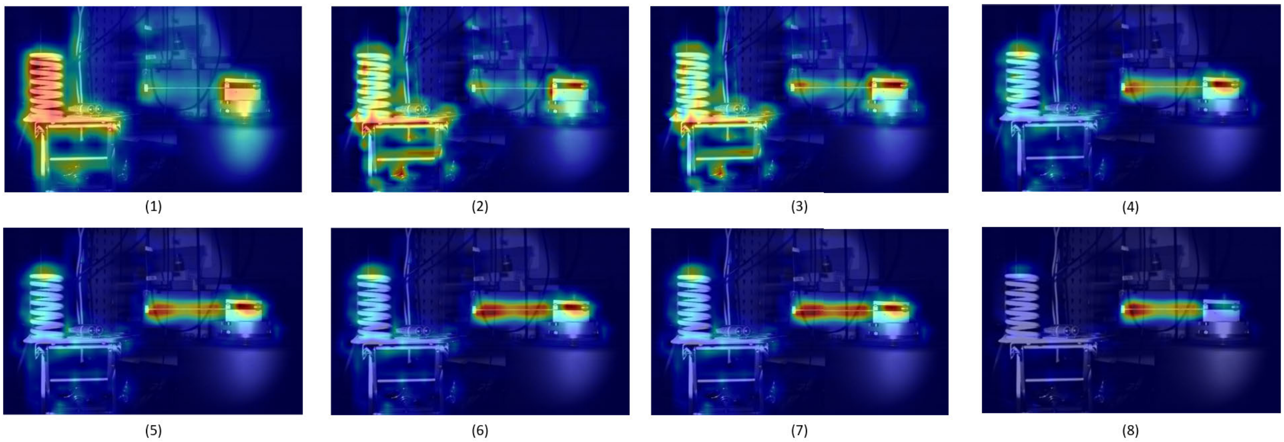

FIGURE 9. CAM of cantilever beam during the training process.

TABLE 5. Performance of classifiers for water pump experiment.

Classifier	Feature extraction method	Accuracy (%)	Precision (%)	Recall (%)
CNN with CAM	FL	100	100	100
MLP	FE	50.44	100	50.00
RF	FE	98.50	97.00	98.20
SVM_{linear}	FE	49.65	50.35	49.65

fication models. The learning parameters of the algorithms are listed in Table 3. It can be seen from Table 4 that feature learning (FL) offers better and outstanding performance for this task. It is notable that the CNN approach yields an exceptional result at the highest performance, 95.85% accuracy, 100% precision, and 92.33% recall, without experts' prior knowledge and manual design of the features.

CAM algorithm was also applied to demonstrate the attention where the CNN model distinguishing between two classes at a high accuracy notice, as visualized in Figure 8. The figure exhibits that the marked region of an image is highly important for the model to discriminate one class from the other, overlapping the class activation map with an original image. Likewise, the region activated in red depicts the most significant part, herein a cantilever beam, for the proposed algorithm to classify two different types of conditions. This explainable deep learning enables us to identify the most vibrating regime from images, not any vibration sensors.

Figure 9 shows CAM results of a base-excited cantilever beam during the model training process, representing three dominantly activated regions in these images; a cantilever beam, an exciter, and a spring on the table. It is clear to say that there actually exists vibration to those activated parts in the experiment. The most vibrating part is from the cantilever beam, which is amplified by the exciter, while the exciter has the second-highest vibration as a source of the acceleration. There is also a little movement in the spring on the table since it is not completely able to be isolated from the excitation. At the early stage of the training, it is observed that the CAM algorithm focuses not only on the beam and the exciter but the spring on the table that is the relatively meaningless region for classifying two different classes. The focusing part moves from the spring on the table to the beam and the exciter as the training goes halfway and finally converges towards the cantilever beam in the end. Hence, the highlighted region of the CAM algorithm is gradually meeting at a point of the beam as training of the model progresses.

B. WATER PUMP

Table 5 shows the comparative performance of CNN with CAM and shallow machine learning algorithms for the water pump experiment. Configurations of the models and features are identically used as in the cantilever beam case when classifying with those machine learning algorithms in this experiment, whereas the structure of CNN with CAM is slightly different from one. Additional layers are stacked

TABLE 6. Architecture of CNN with CAMs for water pump experiment.

Layer	Type	Channels	Kernel size	Stride	Layer	Type	Channels	Kernel size	Stride
0	Input	3	-	-	12	2D Maxpooling	-	2×2	2
1	2D Convolution	16	3×3	1	13	2D Convolution	32	3×3	1
2	2D Convolution	16	3×3	1	14	2D Convolution	32	3×3	1
3	2D Maxpooling	-	2×2	2	15	2D Maxpooling	-	2×2	2
4	2D Convolution	16	3×3	1	16	2D Convolution	32	3×3	1
5	2D Convolution	16	3×3	1	17	2D Convolution	32	3×3	1
6	2D Maxpooling	-	2×2	2	18	2D Maxpooling	-	2×2	2
7	2D Convolution	16	3×3	1	19	2D Convolution	64	3×3	1
8	2D Convolution	16	3×3	1	20	2D Convolution	64	3×3	1
9	2D Maxpooling	-	2×2	2	18	2D Maxpooling	-	2×2	2
10	2D Convolution	32	3×3	1	16	GAP[a]	64	-	-
11	2D Convolution	32	3×3	1	17	Linear	2	-	-

[a] Global average pooling

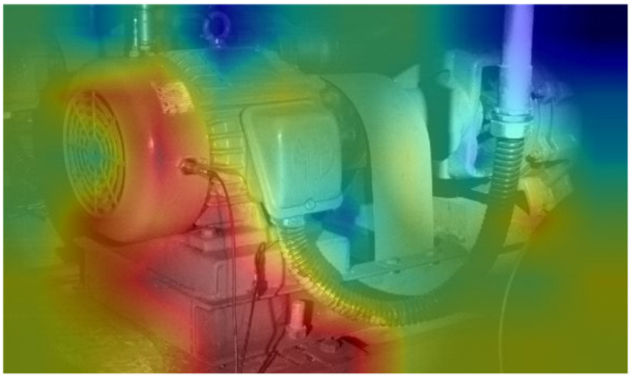

FIGURE 10. CAM result of water pump dataset.

for CNN with the CAM model to earn better performance and activate more extensive areas owing to the fact that the pump occupies more space than the cantilever beam in the image data set. The description of the architecture is detailed in Table 6. As expected, the deep learning method shows the best performance of all classifiers compared in this experiment, 100% accuracy, precision, and recall. The CAM result for the water pump case is visualized in Figure 10, as the same result as anticipated because of the looseness on the bottom side nut of the system. The experimental result can be explained that the most vibrating region is the most significant one in fault diagnosis. From this case study, we can conclude that explainable deep learning can assist to point out the faulty components to repair from images.

To sum up, based on two above-mentioned experiments, distinguishing normal and fault images are not that easy by using naked-eye. We have also performed an experiment to see such differences by comparing the sharpness of the vibrating area. However, the sharpness of the two images were not significantly different. Rather than simply classifying the images into fault or normal class, our study proposes new approach of fault detection using CAM, where the important feature near the vibrating part could be understandable to human or engineers.

V. CONCLUSION

This paper presented image-based fault detection and diagnosis method of vibrating mechanical system using a convolutional neural network (CNN) and class activation map (CAM), giving some examples of base-excited cantilever beam and water pump system. It is evidenced that the presented approach, an feature learning (FL)-based technique, is not only able to detect the machine's condition but also it is able to localize the fault instantly, enabling a real-time and accurate fault detection. Unlike traditional shallow machine learning algorithms that require feature extraction, the advantage of FL is that there is no expert knowledge is required to extract the features representation of the problem. The FL-approach gave at least 2% accuracy improvement over the FE-approaches, as indicated in our aforementioned experimental result. Furthermore, since it is not straightforward to understand where in the image to observe in to localize the fault condition, we have shown that by employing CAM, a significant understanding of the faulty regions of the image could be discovered. Concerning future research directions, it is interesting to perform an experiment using different equipment settings (e.g., frame per seconds and resolution) in order to understand the minimum settings that should be met. Moreover, as this study only considered mechanical looseness as a fault, other faults such as bearing defect, impeller defect, misalignment, and cavitation are valuable to be further investigated.

ACKNOWLEDGMENT

(Kyung Ho Sun and Hyunsuk Huh contributed equally to this work.)

REFERENCES

[1] P. K. Wong, J. Zhong, Z. Yang, and C. M. Vong, "Sparse Bayesian extreme learning committee machine for engine simultaneous fault diagnosis," *Neurocomputing*, vol. 174, pp. 331–343, Jan. 2016.

[2] L. Hong and J. S. Dhupia, "A time domain approach to diagnose gearbox fault based on measured vibration signals," *J. Sound Vibrat.*, vol. 333, no. 7, pp. 2164–2180, Mar. 2014.

[3] M. J. M. Goncalves, R. C. Creppe, E. G. Marques, and S. M. A. Cruz, "Diagnosis of bearing faults in induction motors by vibration signals—Comparison of multiple signal processing approaches," in *Proc. IEEE 24th Int. Symp. Ind. Electron. (ISIE)*, Jun. 2015, pp. 488–493.

[4] K. Worden and A. P. Burrows, "Optimal sensor placement for fault detection," *Eng. Struct.*, vol. 23, no. 8, pp. 885–901, Aug. 2001.

[5] X. Liu, S. Weerakkody, and B. Sinopoli, "Sensor placement for reliable observability: A structured systems approach," in *Proc. IEEE 55th Conf. Decis. Control (CDC)*, Dec. 2016, pp. 5414–5421.

[6] B. A. Tama and K.-H. Rhee, "An in-depth experimental study of anomaly detection using gradient boosted machine," *Neural Comput. Appl.*, vol. 31, no. 4, pp. 955–965, Apr. 2019.

[7] W. Choi, H. Huh, B. A. Tama, G. Park, and S. Lee, "A neural network model for material degradation detection and diagnosis using microscopic images," *IEEE Access*, vol. 7, pp. 92151–92160, 2019.

[8] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[9] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[10] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. CVPR*, Dec. 2001, p. 1.

[11] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[13] B. Li, M.-Y. Chow, Y. Tipsuwan, and J. C. Hung, "Neural-network-based motor rolling bearing fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 47, no. 5, pp. 1060–1069, 2000.

[14] Z. Chen, C. Li, and R.-V. Sanchez, "Gearbox fault identification and classification with convolutional neural networks," *Shock Vibrat.*, vol. 2015, pp. 1–10, Apr. 2015.

[15] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," in *The Handbook Brain Theory Neural Netw.*, vol. 3361, no. 10, 1995, pp. 255–258.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[18] S. Hoo-Chang, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, p. 1285, 2016.

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[20] J. Wang, Y. Ma, L. Zhang, R. X. Gao, and D. Wu, "Deep learning for smart manufacturing: Methods and applications," *J. Manuf. Syst.*, vol. 48, pp. 144–156, Jul. 2018.

[21] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," *Insights Imag.*, vol. 9, no. 4, pp. 611–629, 2018.

[22] S. Khan and T. Yairi, "A review on the application of deep learning in system health management," *Mech. Syst. Signal Process.*, vol. 107, pp. 241–265, Jul. 2018.

[23] 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.

[24] J. Rafiee, F. Arvani, A. Harifi, and M. H. Sadeghi, "Intelligent condition monitoring of a gearbox using artificial neural network," *Mech. Syst. Signal Process.*, vol. 21, no. 4, pp. 1746–1754, May 2007.

[25] A. Sadeghian, Z. Ye, and B. Wu, "Online detection of broken rotor bars in induction motors by wavelet packet decomposition and artificial neural networks," *IEEE Trans. Instrum. Meas.*, vol. 58, no. 7, pp. 2253–2263, Jul. 2009.

[26] G. F. Bin, J. J. Gao, X. J. Li, and B. S. Dhillon, "Early fault diagnosis of rotating machinery based on wavelet packets—Empirical mode decomposition feature extraction and neural network," *Mech. Syst. Signal Process.*, vol. 27, pp. 696–711, Feb. 2012.

[27] P. K. Kankar, S. C. Sharma, and S. P. Harsha, "Vibration-based fault diagnosis of a rotor bearing system using artificial neural network and support vector machine," *Int. J. Model., Identificat. Control*, vol. 15, no. 3, pp. 185–198, 2012.

[28] M. Amar, I. Gondal, and C. Wilson, "Vibration spectrum imaging: A novel bearing fault classification approach," *IEEE Trans. Ind. Electron.*, vol. 62, no. 1, pp. 494–502, Jan. 2015.

[29] O. Janssens, R. Van de Walle, M. Loccufier, and S. Van Hoecke, "Deep learning for infrared thermal image based machine health monitoring," *IEEE/ASME Trans. Mechatronics*, vol. 23, no. 1, pp. 151–159, Feb. 2018.

[30] P. Potočnik and E. Govekar, "Semi-supervised vibration-based classification and condition monitoring of compressors," *Mech. Syst. Signal Process.*, vol. 93, pp. 51–65, Sep. 2017.

[31] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," *Sensors*, vol. 17, no. 2, p. 425, Feb. 2017.

[32] J. Pan, Y. Zi, J. Chen, Z. Zhou, and B. Wang, "LiftingNet: A novel deep learning network with layerwise feature learning from noisy mechanical data for fault classification," *IEEE Trans. Ind. Electron.*, vol. 65, no. 6, pp. 4973–4982, Jun. 2018.

[33] D.-T. Hoang and H.-J. Kang, "Rolling element bearing fault diagnosis using convolutional neural network and vibration image," *Cognit. Syst. Res.*, vol. 53, pp. 42–50, Jan. 2019.

[34] M. He and D. He, "A new hybrid deep signal processing approach for bearing fault diagnosis using vibration signals," *Neurocomputing*, vol. 396, pp. 542–555, Jul. 2020.

[35] T. Chen, Z. Wang, X. Yang, and K. Jiang, "A deep capsule neural network with stochastic delta rule for bearing fault diagnosis on raw vibration signals," *Measurement*, vol. 148, Dec. 2019, Art. no. 106857.

[36] H. Chen, N. Hu, Z. Cheng, L. Zhang, and Y. Zhang, "A deep convolutional neural network based fusion method of two-direction vibration signal data for health state identification of planetary gearboxes," *Measurement*, vol. 146, pp. 268–278, Nov. 2019.

[37] X. Li, J. Li, Y. Qu, and D. He, "Semi-supervised gear fault diagnosis using raw vibration signal based on deep learning," *Chin. J. Aeronaut.*, vol. 33, no. 2, pp. 418–426, Feb. 2020.

[38] A. Youcef Khodja, N. Guersi, M. N. Saadi, and N. Boutasseta, "Rolling element bearing fault diagnosis for rotating machinery using vibration spectrum imaging and convolutional neural networks," *Int. J. Adv. Manuf. Technol.*, vol. 106, nos. 5–6, pp. 1737–1751, Jan. 2020.

[39] T. Wang, Q. Han, F. Chu, and Z. Feng, "Vibration based condition monitoring and fault diagnosis of wind turbine planetary gearbox: A review," *Mech. Syst. Signal Process.*, vol. 126, pp. 662–685, Jul. 2019.

[40] H. Wang, S. Li, L. Song, and L. Cui, "A novel convolutional neural network based fault recognition method via image fusion of multi-vibration-signals," *Comput. Ind.*, vol. 105, pp. 182–190, Feb. 2019.

[41] H. D. M. Onchis, "A deep learning approach to condition monitoring of cantilever beams via time-frequency extended signatures," *Comput. Ind.*, vol. 105, pp. 177–181, Feb. 2019.

[42] M. Tulio Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," 2016, *arXiv:1606.05386*. [Online]. Available: http://arxiv.org/abs/1606.05386

[43] R. Gulati and R. Smith, *Maintenance and Reliability Best Practices*. New York, NY, USA: Industrial Press, 2009.

[44] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene CNNs," 2014, *arXiv:1412.6856*. [Online]. Available: http://arxiv.org/abs/1412.6856

[45] M.-K. Hu, "Visual pattern recognition by moment invariants," *IEEE Trans. Inf. Theory*, vol. IT-8, no. 2, pp. 179–187, Feb. 1962.

[46] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vols. SMC–3, no. 6, pp. 610–621, Nov. 1973.

[47] X.-Y. Wang, J.-F. Wu, and H.-Y. Yang, "Robust image retrieval based on color histogram of local feature regions," *Multimedia Tools Appl.*, vol. 49, no. 2, pp. 323–345, Aug. 2010.

KYUNG HO SUN received the Ph.D. degree in mechanical engineering from Seoul National University, Seoul, South Korea, in 2010. He is currently a Principal Researcher with the System Dynamics Laboratory, Korea Institute of Machinery and Materials, Daejeon, South Korea. His research interests include prognostics and health management (PHM) for rotating machinery, and machine vision.

HYUNSUK HUH received the B.S. degree from the Ulsan National Institute of Science and Technology, Ulsan, South Korea, in 2018. He is currently pursuing the M.S./Ph.D. degree with POSTECH. His research interests include artificial intelligence in mechanical systems, and robotics.

BAYU ADHI TAMA received the Ph.D. degree from Pukyong National University, South Korea. He has worked as a Postdoctoral Researcher with the School of Management Engineering, Ulsan National Institute of Science and Technology, South Korea. He is currently with the Industrial Artificial Intelligence Laboratory, POSTECH, as a Postdoctoral Researcher. He has published more than 30 papers in academic conferences and journals. His research interests include machine learning and artificial intelligence techniques applied for industrial applications, medical informatics, and cybersecurity.

SOO YOUNG LEE received the B.S. degree from Chung-Ang University, Seoul, South Korea, in 2019. He is currently pursuing the M.S./Ph.D. degree with the Industrial Artificial Intelligence Laboratory, POSTECH, South Korea. His research interests include industrial artificial intelligence with mechanical systems and the AI-based smart manufacturing.

JOON HA JUNG received the Ph.D. degree in mechanical engineering from Seoul National University, Seoul, South Korea, in 2019. He is currently a Senior Researcher with the System Dynamics Laboratory, Korea Institute of Machinery and Materials, Daejeon, South Korea. His research interests include prognostics and health management (PHM) for rotating machinery, and artificial intelligence.

SEUNGCHUL LEE received the M.S. and Ph.D. degrees from the University of Michigan at Ann Arbor, USA, in 2008 and 2010, respectively. He has worked as an Assistant Professor with the Ulsan National Institute of Science and Technology, South Korea, and a Postdoctoral Research Fellow with the University of Michigan at Ann Arbor. He has been an Assistant Professor with the Department of Mechanical Engineering, POSTECH, since 2018. His research interests include industrial artificial intelligence with mechanical systems, deep learning for machine healthcare, and the IoT-based smart manufacturing.