IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 4, 2021, accepted August 8, 2021, date of publication August 11, 2021, date of current version August 26, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3104189

Supervised-Learning-Based Intelligent Fault Diagnosis for Mechanical Equipment

GEONKYO HONG AND DONGJUN SUH, (Member, IEEE)
Department of Convergence and Fusion System Engineering, Kyungpook National University, Sangju 37224, South Korea

Corresponding author: Dongjun Suh (dongjunsuh@knu.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF 2021R1I1A3049503), (No. NRF-2021R1A5A8033165) and results of a study on the High Performance Computing (HPC) Support Project, supported by the MSIT and the National IT Industry Promotion Agency (NIPA).

ABSTRACT Recently, anomaly detection for improving the productivity of machinery in industrial environments has drawn considerable attention. As large-scale data collection and processing are becoming easier owing to technological developments, data-based deep-learning technology is being developed to detect anomalies in mechanical equipment operation. This study proposes an ensemble model that combines stacked two-dimensional and one-dimensional convolutional neural networks (CNNs), residual long short-term memory (LSTM), and LSTM based on supervised learning. The model, which is called the SCRLSTM model, can detect abnormal data generated by mechanical equipment. The proposed model can extract the spatial features of data using a CNN model and detect anomalous states in the time-series-based vibration datasets of machinery under various environments through residual LSTM. To verify this model, data augmentation was applied to the original time-series-based mechanical vibration dataset, which had unbalanced samples that lowered the performance of the abnormal anomaly detection model. In addition, an image-based analysis was performed by converting time-series-based raw-signal data to Mel-spectrogram images, thereby achieving better performance in the fault diagnosis system to which data augmentation was applied. The proposed SCRLSTM model shows better performance than other supervised-learning-based models on datasets having different lengths under various conditions. This indicates that the proposed anomaly detection model can be expected to improve the productivity of mechanical equipment in industrial settings.

INDEX TERMS Supervised learning, anomaly detection, fault diagnosis, time series, data augmentation, mel-spectrogram.

I. INTRODUCTION

As the technology used with mechanical equipment advances, the complexity of industrial environments and uncertainties pertaining to productivity also increase. If aged mechanical equipment is neglected and the damage it suffers is not adequately rectified, the equipment will become defective and will suffer from productivity reduction. Moreover, the damaged equipment can cause damage to other equipment. Hence, there is a need to develop advanced technology to improve equipment safety [1].

Industrial machines (e.g., valves, pumps, fans, slide rails) and rolling bearings are core components of industrial systems and play a critical role in determining both the performance and life of these systems [2]. Faults can develop in these mechanical components because of various causes, and the most serious cause is the defects in the bearings of electromagnetic drive systems [3]. Various condition-monitoring methods have been used to detect faults in industrial machines and bearings.

Traditional methods include the K-nearest neighbor algorithm, which is based on distance; the local outlier factor (LOF), which is used to detect local anomalies; and the connectivity-based outlier factor, which is an improved version of the LOF and detects anomalies using the radius. Traditional methods based on distance and density have the disadvantage of requiring considerable time for anomaly detection as the number of data points increases [4]. Deep-learning approaches that can overcome this limitation have recently emerged and have shown higher performance than

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

traditional methods [5]. Furthermore, the development of the Internet of things industry has facilitated large-scale data collection. Thus, the importance of anomaly detection based on supervised learning is increasing. According to recent trends, machine-learning-based anomaly detection approaches are composed of three major steps:

1) Data preprocessing – The important features of normal and abnormal data are extracted from time-series-based raw signals.
2) Selection of a deep-learning-based model – A model is selected for use in the fault diagnosis system.
3) Anomaly detection – The deep-learning-based model uses the extracted features to detect anomalies through a learning process.

Machine learning has been used to study and develop time-series-based methods for anomaly detection of industrial machines and bearing faults using data from mechanical equipment. Purohit et al performed anomaly detection and classification by converting the measured MIMII (valve, pump, fan, and slide rails) dataset into Mel spectrogram images through preprocessing [6]. Suefusa et al performed interpolation-based anomaly detection using Mel spectrogram images [7]. Yuan et al converted the bearing vibration data set into time-frequency images to diagnose bearing failures using a model ensemble of convolutional neural network and support vector machine [8]. Yang et al trained the DNN using the bearing dataset to obtain the initial classification results where the classified results are evaluated by testing the subsignals extracted from the raw data, and the sample labels are modified according to the evaluation results. Datasets with modified labels are finally classified through DNN [9]. Zhang et al performed anomaly diagnosis using a few-shot learning-based model on a bearing dataset [10]. Wen et al diagnosed anomalies with a snapshot ensemble learning method that constructs an ensemble by combining local minima using a cyclic learning rate scheduler (CLR) with a bearing dataset [11]. Li et al perform bearing diagnosis through a soft-voting-method-based deep learning model that aggregates prediction results of signals sliced by sliding windows to increase accuracy and stability using bearing data [12]. Liu et al proposed a domain adaptive approach by developing a deep-learning-based JDDA model for fault diagnosis of an electromechanical drive system [13]. These studies established deep learning for mechanical equipment anomaly detection using the time and frequency domains. However, the limitation of these studies is that data from both bearings and industrial machines were not considered. Additionally, the intelligent data-based fault diagnosis model was not tested across a range of loads, durations, and noises; therefore, the data used was insufficient for a comprehensive testing of the fault diagnosis model. Thus, the previous studies lack generalization ability compared to our proposed anomaly detection framework for various environmental adaptations. Therefore, to construct the fault diagnosis system with a robust and generalized performance, it is necessary to conduct an experiment considering the actual environment [14]. In this study, data preprocessing overcomes this limitation by converting raw signals into a Mel-spectrogram with augmented data that can reflect various conditions. In addition, we propose fault diagnosis model that operates on deep learning and can detect time-series based anomalies in machine vibration data, thereby overcoming much limitation posed by existing intelligent-data-based defects.

This study proposes an ensemble model that combines stacked supervised-learning-based two-dimensional (2D) and one-dimensional (1D) CNNs, residual LSTM, and LSTM. This model, which is referred to as the SCRLSTM model, can provide anomaly detection and diagnosis of time-series-based data. The main contributions of this study are as follows:

1) Anomalies were determined using Mel-spectrogram images, which were used to extract the normal and abnormal features of a time-series-based mechanical equipment vibration dataset.
2) Datasets were augmented to solve the data imbalance problem, and this improved the accuracy of the model.
3) Furthermore, experiments were performed on the robustness of various models to noise in consideration of the signal-to-noise ratio (SNR) prevailing in actual industrial sites.
4) The supervised-learning-based model was verified using a time-series-based vibration dataset having various lengths. In addition, the dataset was divided according to the motor load, and various environments were considered by carrying out transfer learning from one load to another.
5) Through 2D and 1D convolutional neural networks, it is possible to extract a feature map representing the spatial relationship of Mel spectrogram images of bearings and industrial machines (valve, pump, fan, slide rails). In addition, temporal feature maps for time-series analysis can be extracted through LSTM.
6) The proposed SCRLSTM model using various load, noise, and time series-based mechanical equipment datasets demonstrates superior generalization ability as it shows better performance in average accuracy and confusion plot compared to the other five network models. This generalization capability effectively handles the task of diagnosing machine faults.

The remainder of this paper is organized as follows. Section 2 briefly describes the core deep-learning concept used in the backgrounds and reviews related studies. Section 3 describes the Mel-spectrogram images obtained by applying a signal processing technique to the time-series-based raw-signal data and presents the new deep-learning-based model framework. Section 4 briefly describes the experimental design and presents the experimental results. Finally, Section 5 presents the conclusions and proposes ideas for additional work.

II. BACKGROUND

The frequency and time domains of vibration sensors have been used to detect anomalies in acoustic scene classification

and event-detection technologies [7], [15], [16]. The use of deep learning to classify acoustic scenes and detect acoustic events has seen considerable development, with several promising studies conducted into these aspects [17], [18]. Anomaly detection systems for mechanical equipment have largely been studied using convolutional neural networks (CNNs) and long short-term memory (LSTM).

A. CONVOLUTIONAL NEURAL NETWORK

Over the past few years, CNN have been widely used in the image and signal processing fields, and the scope of further utilizing CNN in these fields has been ascertained [19]. Recently, a combination of 1D and 2D CNN has been used for image classification and pattern recognition, and the combination has shown promising possibilities in processing structured data. It has been confirmed that the combined use of 1D and 2D CNN provides higher performance than the use of only one of the CNN [20].

CNN is a deep-learning architecture created by imitating the human optic nerve. It can learn the unique features of images regardless of the location and direction of objects, unlike the multilayer perceptron. 1D and 2D CNN are realized using the same principle. The CNN performs by combining convolution (realized using an existing image processing filter) with a neural network. The convolution operation maintains the spatial information of images, dramatically reduces the amount of computation required compared to that required by a fully connected neural network, and shows good performance in image classification.

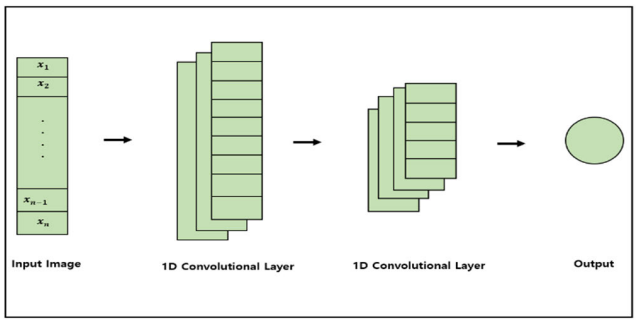


FIGURE 1. 1D convolutional neural network architecture.

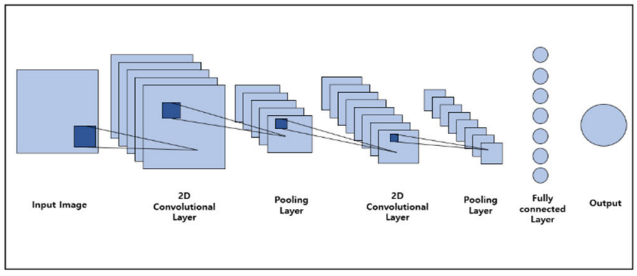


FIGURE 2. 2D convolutional neural network architecture.

The convolution stages shown in Figure 1 and Figure 2 represent a learnable convolution filter set, including pooling tasks. The CNN uses such a convolution filter instead of a pixel to solve the problem encountered by a multilayer perceptron. The convolution filter extracts the features (corners and curves of an image) of the input data by calculating weights and applying an activation function while moving at regular intervals from the top left to the bottom right of the image through a window. If the parts that the operating filter passes through coincide with the features extracted by the convolution filter, a high output can be obtained, thus improving the possibility of achieving good image classification. The major features extracted by learning the image using such an operating filter are input to the pooling layer, which reduces the image size by lowering the dimensions. The pixels in a specific region in the image are grouped and reduced to a representative value. This pooling layer helps to reduce the amount of computation and prevent overfitting. As the convolution operation progresses from the left in each layer, the network learns by extracting more features.

B. RECURRENT NEURAL NETWORKS

A recurrent neural network (RNN) is a deep-learning-based model that processes the input and output in a sequence unit. The RNN performs learning by memorizing and using information about past events. Theoretically, the RNN can process sequence data well by considering the sequence, but successful learning becomes difficult for the RNN as the temporal range of the sequence increases [21].

Figure 3 shows the architecture of an LSTM repetitive neural network architecture. The LSTM can overcome the limitations of RNN learning by introducing two memories and three gates. The LSTM has the ability to learn from a long input sequence. The gradient loss problem can be solved using the memory cell of the LSTM, and multiple LSTM layers can be stacked.

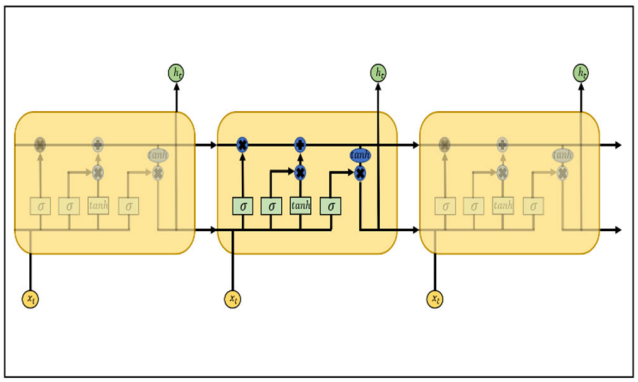


FIGURE 3. Long short-term memory architecture.

The equation of the LSTM state can be expressed as [22]

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (1)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (2)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (3)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (4)$$

$$h_t = o_t \circ tanh(c_t) \quad (5)$$

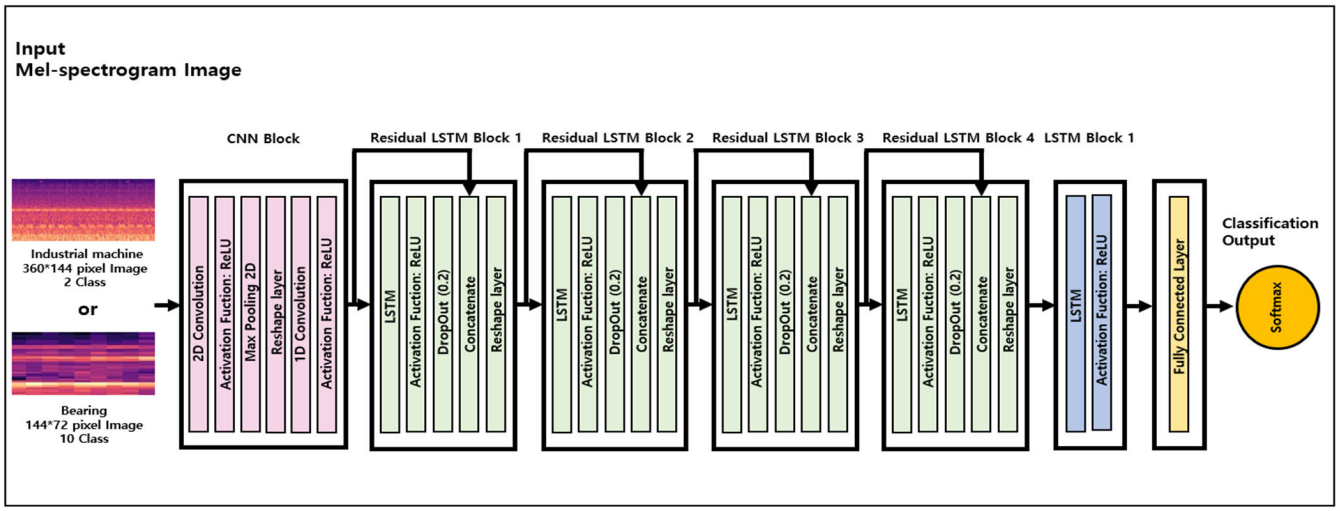


FIGURE 4. Proposed SCRLSTM architecture consisting of stacked 2D and 1D CNNs, residual LSTM, and LSTM.

where x_t is the input vector, f_t is the forget gate at time t, i_t is the input gate at time t, o_t is the output gate at time t, c_t is the memory cell activation vector at time t, and h_t is the hidden state at time t. W_f, W_i, and W_o are the input weights for the forget, input, and output gates, respectively. U_f, U_i, and U_O are the recurrent weights for the forget, input, and output gates, respectively. b_f, b_i, b_o, and b_c are the biases. σ is the sigmoid activation function, $\circ$ is the matrix product, and *tanh* is the hyperbolic tangent activation function. The three gates select the amount of information that is included in the LSTM:

1) Forget gate: It determines the amount of past information included in the LSTM. After h_{t-1} and x_t are input, the value that passes through the sigmoid σ function is the output value of the forget gate. The output range of the sigmoid σ function is between 0 and 1. A value of 0 means that the information of the previous state is forgotten, whereas a value of 1 means that the information of the previous states is memorized.
2) Input gate: It determines the amount of new information that is stored. The sigmoid is taken after receiving h_{t-1} and x_t as inputs, and the hyperbolic tangent is taken with the same inputs. Then, the product operation value becomes the output value of the input gate.
3) Output gate: It determines which value will finally be output. The h_t value in the present moment is operated with c_t and can be used as the input of h_{t+1} of the next time point as soon as h_t is output.

III. SCRLSTM MODEL FOR FAULT DIAGNOSIS

This study proposes a new model that combines 2D and 1D CNNs and residual LSTM and LSTM to detect anomalies in a time-series-based mechanical equipment vibration dataset. Figure 4 shows the proposed SCRLSTM model. In the input layer of the proposed model, the spatial information of Mel-spectrogram image is maintained using the multifilter of the 2D CNN layer, and the features of the adjacent low-frequency image are extracted and learned. Then, the extracted features pass through the pooling layer, where they are gathered and strengthened and are then used as input for the 1D CNN layer. The 1D CNN amplifies the efficiency of operations by lowering the dimensions of the previously extracted important information pertaining to the mechanical equipment. The feature values extracted through the CNN layers are used as the input values of the residual LSTM layer. Figure 5 shows the residual LSTM layer of the proposed model. As mentioned above, the LSTM provides higher accuracy than the RNN because the former has better learning ability. However, even if the number of layers of the LSTM is increased, only a specific number of layers operate, and the network becomes too slow and finds it difficult to learn. Thus, a deep LSTM network can cause gradient loss problems as in the case of an RNN. To solve this problem, a residual LSTM layer that models the difference between the outputs of the intermediate layer and next layer is used in the ensemble [23].The advantages of Residual LSTM is that it can increase the accuracy and improve the computational speed during learning by using the residual connection [24], [25]. Moreover, it helps to solve the problem of gradient loss during backpropagation calculations and can greatly improve the gradient flow [23]. Furthermore, it can learn the feature maps from the spatial feature relationship of the time-series-based Mel image data obtained through the CNN. In the last layer, temporal feature maps between channels are extracted through the LSTM layer and fully connected layer from the output values of the residual LSTM layer, which has the time-series features of the industrial machines and bearing elements [26]. These features are then supplied to the output classification layer, which converts them into a probability distribution between 0 and 1 corresponding to the class using the softmax activation function [27], [28] [29], [30]. Thus, anomalies can be detected by classifying normal and abnormal data into two or

more classes. The softmax function is expressed as follows:

$$Softmax\,(x_i) = \frac{exp(x_i)}{\sum_{j=1}^{k}(x_j)} \quad for\ i = 1, 2 \ldots k \tag{6}$$

where x_t^i is the input, t is the time step, h_t^i is the hidden state, and c_t^i is the cell state. W^i is the weight, and W^{i+1} is the weight of the $t-1$. Residual connections between $LSTM_i$ and $LSTM_{i+1}$ Accordingly, the following equation is obtained:

$$\begin{aligned} c_t^i, h_t^i &= LSTM_i\left(c_{t-1}^i, h_{t-1}^i, x_t^{i-1}; W^i\right) \\ x_t^i &= h_t^i + x_t^{i-1} \\ c_t^i, h_t^i &= LSTM_{i+1}\left(c_{t-1}^{i+1}, h_{t-1}^{i+1}, x_t^i; W^{i+1}\right) \end{aligned} \tag{7}$$

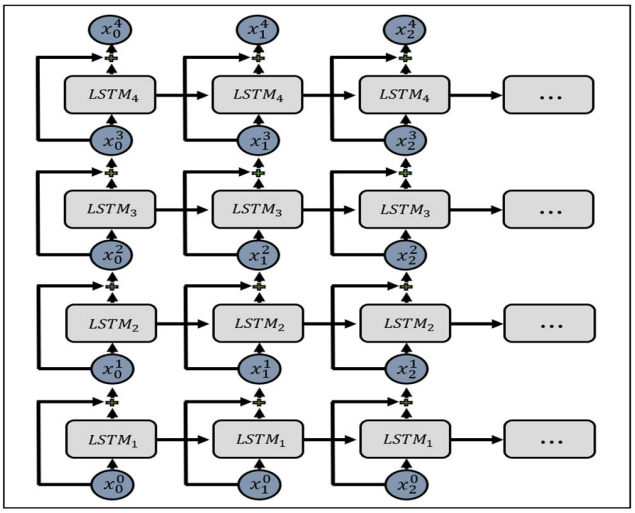


FIGURE 5. Residual LSTM architecture.

The reason for each hyperparameter configuration is as follows. Selected (2,2) as. The kernel size in the 2D convolutional and the Maxpooling layers produced improved performance so the (2,2) kernel size was selected instead of the commonly used (3,3) kernel [31]. The kernel size in the 1D convolutional layer is larger than that of the 2D convolutional layer to capture the basic components with the general characteristics of the previously extracted image. We also used heuristic approach to specify the number of convolutional layer nodes with good performance. In this method, the number of nodes in the 2D convolutional layer was 10, 8 nodes were used in the 1D convolutional layer for the bearing dataset, and 54 nodes were used in the industrial machine dataset. The LSTM layer is configured differently for each data. The layer 4 and layer 5 have the same number of nodes because they are residual connections, and the layers 6 and 7 are also configured with the same number of nodes for the same reason. The LSTM layer increases the number of nodes when connected from layer 5 to layer 6 to learn features in more time domains according to the scalability of node capacity. Moreover, the LSTM layer, which is layer 8, obtained the best results when an experiment was conducted with the same number of nodes as the output nodes.

TABLE 1. Architecture parameters of model.

Data	Bearing	Industrial machines
Input Layer	144*72*3	360*144*3
Layer1	2D CNN (2 × 2, 10)	2D CNN (2 × 2, 10)
Layer2	MaxPooling (2 × 2)	MaxPooling (2 × 2)
Layer3	1D CNN (9 × 1, 8)	1D CNN (9 × 1, 54)
Layer4	LSTM (8)	LSTM (54)
Layer5	LSTM (8)	LSTM (54)
Layer6	LSTM (16)	LSTM (108)
Layer7	LSTM (16)	LSTM (108)
Layer8	LSTM (10)	LSTM (2)
Output Layer	Dense (10)	Dense (2)

IV. EXPERIMENTAL ANALYSIS

A. DESCRIPTION OF CASE WESTERN RESERVE UNIVERSITY BEARING DATASET

To verify the accuracy of the SCRLSTM model, which is the proposed deep-learning-based model, for the fault diagnosis system of an industrial plant, the time-series-based bearing dataset provided by the Case Western Reserve University (CWRU), which is the benchmark dataset for bearing data, was used [32]. The data used in this study were collected from the drive accelerometer of the device (Figure 6) and the accelerometer at the end of the fan. To reflect the various cases encountered in industrial sites, the data were processed and collected from various points for various motor loads (0, 1, 2, 3 hp) and motor speeds (1720–1797 rpm) [33]. However, because the data in the 0 hp load had missing values, experiments were performed only for 1, 2, and 3 hp; i.e., the 0 hp load was excluded in this study. Table 2 lists information pertaining to the fault labels of the dataset. Sampling frequencies of 12 and 48 kHz were used for data collection. Many studies have performed experiments considering sampling vibration data for both 12 and 48 kHz or for only one of these frequencies [10], [34]. In this experiment, 48 kHz sampling vibration signals were used. In the case of the dataset, the normal data contain only 48 kHz sampling data. Because many bearing faults appear at high frequencies, effective defect diagnosis research should use high-frequency sampling data [35]. The number of open CWRU bearing datasets is insufficient for training the deep-learning-based model. Therefore, in this experiment, the amount of data available for training the model was augmented. This method can improve the accuracy of the model and reduce overfitting [36]. To increase the amount of data, the overlapping samples in the time-series-based data were extracted using the sliding window method [37]–[39]. As shown in Figure 7, the learning dataset is composed of a window length of 4096 data points and a step size of 64 [40]. To secure two or more rotation data points of the rotating motor shaft, the bearing defect area was extracted by selecting a window of 4096 data points. A window length of 4096 data points ensures the capture of the effects of at least two bearing faults. The number of data points in the full rotation of the load axis is expressed

as follows:

$$N = F_S \times \frac{60}{\omega} \tag{8}$$

where N is the number of data points, F_S is the sampling frequency, and ω is the rotating shaft speed. As listed in Table 3, rotating shaft speeds of 1772, 1750, and 1730 rpm were used for the operation conditions. This approach is similar to the approach described in the literature that used 48 kHz sampling data [40]. The test data were generated using the same window length, but the dataset was configured such that the data samples would not overlap. The number of data points according to the labels of the training and test datasets listed in Table 4 was configured identically for each load condition.

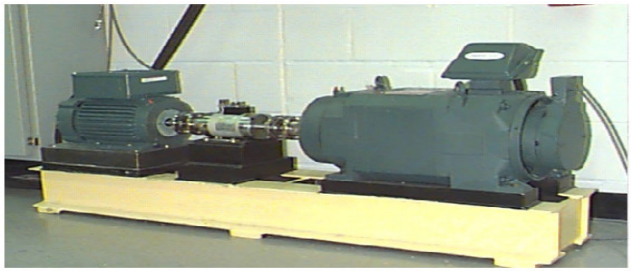


FIGURE 6. CWRU center test apparatus.

TABLE 2. Faults labels of CWRU bearing dataset.

Fault label	Fault type	Severity
0	Normal	None
1	Ball fault	0.007 inch
2	Ball fault	0.014 inch
3	Ball fault	0.021 inch
4	Inner race fault	0.007 inch
5	Inner race fault	0.014 inch
6	Inner race fault	0.021 inch
7	Outer race fault	0.007 inch
8	Outer race fault	0.014 inch
9	Outer race fault	0.021 inch

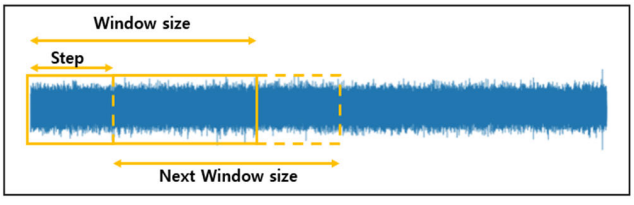


FIGURE 7. Sliding window method for data augmentation.

B. DESCRIPTION OF MALFUNCTIONING INDUSTRIAL MACHINE INVESTIGATION AND INSPECTION DATASET

The malfunctioning industrial machine investigation and inspection (MIMII) dataset contains time-series-based vibration sound data [6]. The dataset is composed of data for four types of machines: valves, pumps, fans, and slide rails. Furthermore, the dataset is composed of 70% training and 30% test datasets. The abnormal data of the valve dataset were collected repeatedly according to the operation state of the solenoid valve (open and closed). In addition, the pump dataset was generated by continuously draining water from a pool and continuously filling water into the pool. Furthermore, the fan dataset represents the gas flow or airflow in the plant. The slide rails represent a linear slide system consisting of a mobile platform and fixed stage base. The entire data were collected using microphones arranged at 10 cm intervals. Each dataset was recorded as audio signals containing +6, 0, and −6 dB noise. The signals include 16-bit audio data sampled at 16 kHz. The abnormal faults in each dataset were as follows:

1) Valve: The abnormal data generated when the valve is opened or closed contain a cracking sound.
2) Pump: The abnormal pump data contain sounds related to leakage, contamination, and clogging.
3) Fan: An abnormal dataset of fans contains data generated by unbalanced voltage changes.
4) Slide rails: The abnormal slide rail data contain sounds related to rail damage, a loose belt, and a lack of grease.

As mentioned above, the MIMII datasets have different SNR values. The different SNRs of each dataset indicate the difference between the data categories. The SNR equation is as follows:

$$SNR_{dB} = 10log_{10}(\frac{P_{signal}}{P_{noise}}) \tag{9}$$

where P_{Signal} denotes the average signal power and P_{noise} denotes the average noise power. For each SNR of the datasets, the average power P_{Signal} was calculated while considering P_{noise} for every segment of the valve data.

TABLE 3. Dataset information.

Label	Motor load	Shaft speed
A	1 hp	1772 rpm
B	2 hp	1750 rpm
C	3 hp	1730 rpm

TABLE 4. Total dataset samples.

Fault label	Number of training samples	Number of test samples
0	11250	176
1	11250	176
2	11250	176
3	11250	176
4	11250	176
5	11250	176
6	11250	176
7	11250	176
8	11250	176
9	11250	176

C. MEL-SPECTROGRAM METHODS

To ensure that the CNN developed for image processing can be applied to audio, preprocessing is required by which the audio, which is in the form of 1D data, can be converted to an appropriate input. In the preprocessing stage, the audio feature data in the time–frequency domain based on sound (e.g., represented by a Mel-spectrogram) are extracted. Then, the CNN is trained to consider these data as one image. This method can be effectively applied to audio [41], [42], and the Mel-spectrogram image in the time–frequency domain can be used to detect anomalies [43]–[45]. The Mel- spectrogram can be generated through a three-step signal processing method:

1) Audio signals are expressed in a digital format using the samples of the mechanical vibration dataset together with the time series.
2) The audio signals are mapped from the time domain to the frequency domain using fast Fourier transform, and the mapping is performed in the overlapping window segment of the audio signals.
3) Finally, the spectrogram is configured by converting the y-axis (frequency) to the log scale and the color dimension (amplitude) to decibels, and the Mel spectrogram is formed by mapping the y-axis (frequency) to the Mel scale.

The sampling rate of signal processing, Mel band, frame length, and frame stride were used as the parameters for the conversion to the Mel-spectrogram image. The sampling rate refers to the number of samplings per unit time, and the Mel band reduces the frequency axis to a specific size. The window size indicates the frequency resolution of signal processing. The hop length for the Fourier transform of the specified time domain was set and compressed according to the Mel curve. The parameter values used in this experiment were 0.025 s for the frame length; 0.010 s for the frame stride; 40 and 48 Hz for the Mel band and sampling rate of the bearing data, respectively; and 80 and 16 kHz for the Mel band and sampling rate of the industrial machine data, respectively.

The transformed Mel-spectrogram image can better represent the special function of the sound (signal) state related to the difference between the normal and abnormal states than the original dataset. Figure 8 shows the raw signals and transformed Mel-spectrogram images of the bearing dataset, and Figure 9 shows the raw signals and transformed Mel-spectrogram images of the industrial machine dataset. The variations in the amplitude of each set of time-series-based data are reflected in the Mel-spectrogram image. The x-axis of the transformed bearing image is expressed in terms of a time period of 0.085 sec with a window length of 4096 data points; these values are the same as those for the raw signals. The x-axis of the industrial machine image has a time period of 10 sec. The y-axis of the bearing and industrial machine images represents the Mel-scale-mapped value of the frequency. When the actual experiment was performed, the x-axis and y-axis of the Mel-spectrogram image were removed and used as the input value of the actual model.

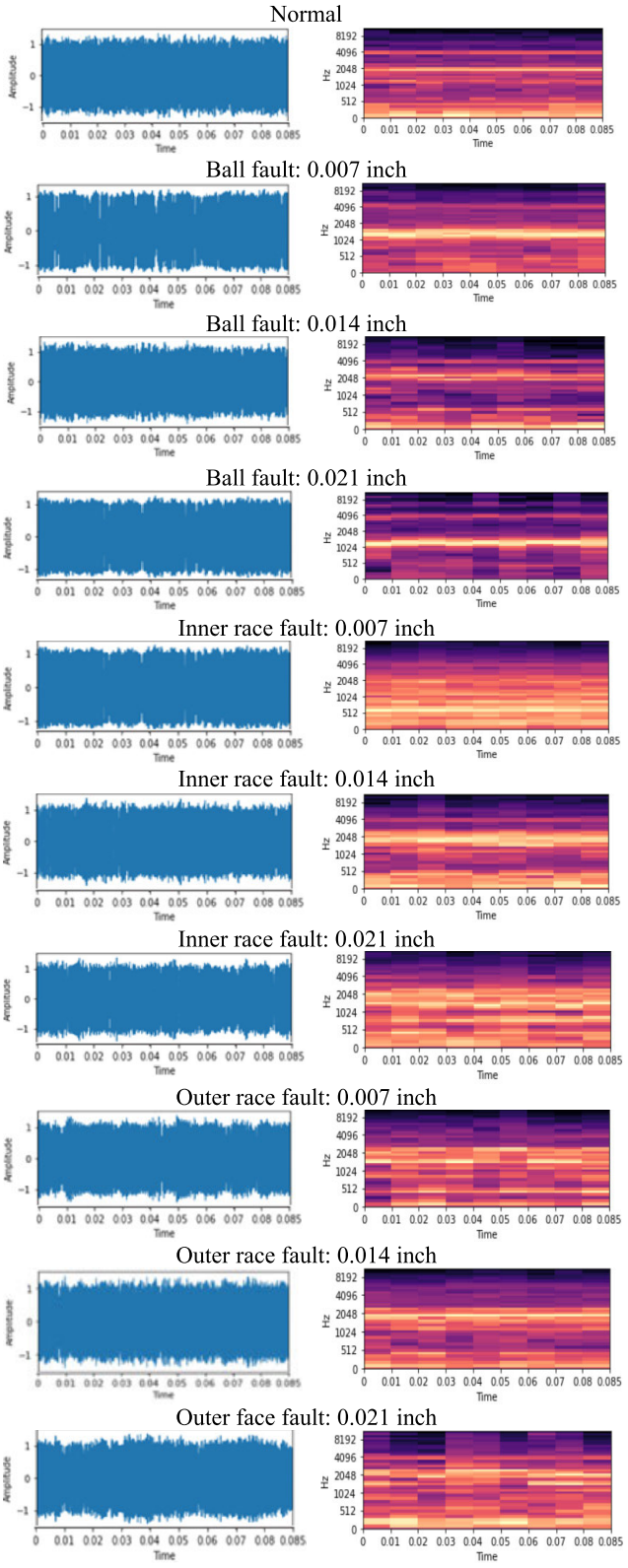


FIGURE 8. Raw signals and Mel-spectrogram (144 × 72 pixel) images of bearing dataset.

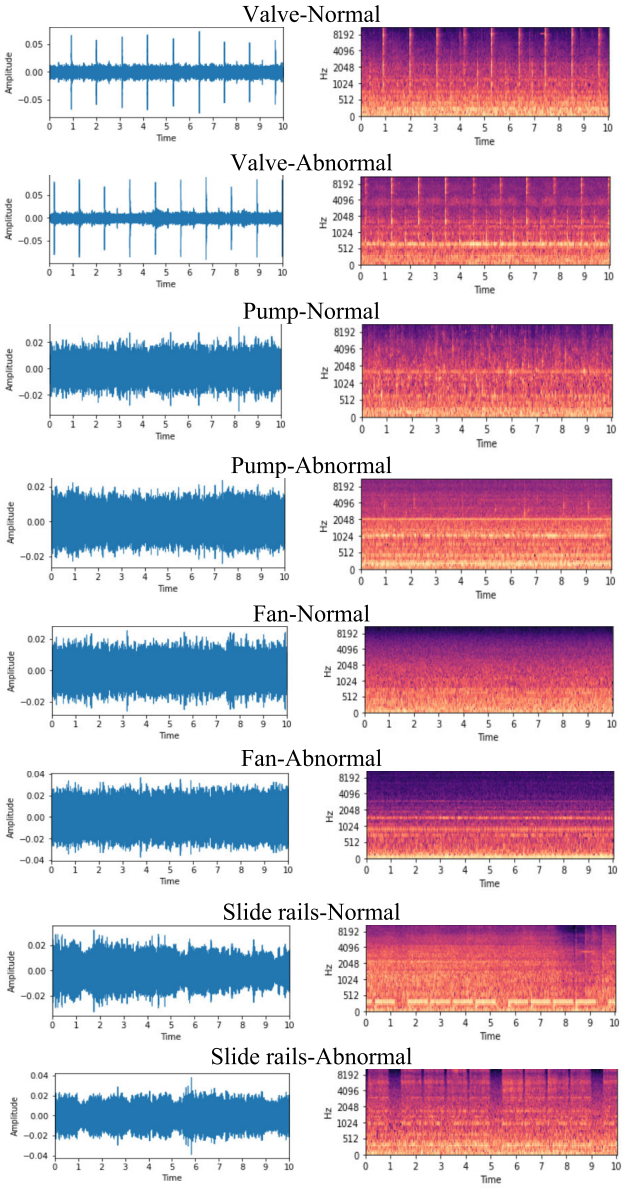


FIGURE 9. Raw signals and mel-spectrogram (360 × 144 pixel) images of industrial machine dataset.

D. DATA AUGMENTATION AND STRUCTURAL SIMILARITY INDEX MEASURE METHODS

The dataset acquired from the real world has differences in the number of data points for each class, and in a critical case, all the data belong to one class only. Machine learning algorithms assume that each class has an equal ratio of data. In the case of a dataset with unbalanced classes, machine learning algorithms cannot perform accurate learning and are biased toward a class that makes up a large proportion [46]. Consequently, class imbalance occurs, wherein classes cannot be accurately classified although the overall accuracy is high. Therefore, data augmentation aimed at increasing deep-learning performance is being actively researched, and the accuracy of deep-learning-based models has increased with the data augmentation ratio [47].

The MIMII dataset used in this experiment has an imbalanced dataset. The dataset (valve, pump, fan, and slide rail data) contains considerably less abnormal data than normal data. When the model was tested with a biased number of normal and abnormal data points in the given dataset, the accuracy was high, but detection was not performed properly because most abnormal data were classified as normal data. Therefore, we performed audio dataset augmentation in the time domain. Data augmentation was performed by moving the time, increasing the time, reducing the volume, and adding noise to all abnormal audio datasets.

The structural similarity index measure (SSIM) scale was used to compare the differences between the original and augmented Mel-spectrogram image data [48]. The SSIM equation is as follows:

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{10}$$

where μ_x is the average of the x image, μ_y is the average of the y image, and C_1 is the normalization constant of luminance. The following equation represents the average brightness based on the above parameters:

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{11}$$

where σ_x is the standard deviation of the x image, σ_y is the standard deviation of the y image, and C_2 is the constant of the contrast term. The following equation represents the contrast of the image:

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{12}$$

where σ_{xy} represents the correlation coefficient between x and y. The structures of the original and augmented images were compared by calculating the correlation coefficients of x and y.

$$SSIM(x, y) = \mathrm{l}(\mathrm{x}, \mathrm{y})\,\mathrm{c}(\mathrm{x}, \mathrm{y})\,\mathrm{s}(\mathrm{x}, \mathrm{y}) \tag{13}$$

After comparing the original and augmented image data using the SSIM scale, only the image data with an SSIM value of 85% or higher were used as the input values of the deep-learning-based model. Table 5 lists the values before and after the data augmentation for the industrial machines with different SNRs.

E. COMPARISON BETWEEN RAW SIGNALS AND MEL SPECTROGRAMS

In this experiment, the performance differences between the signal data of the time-series-based vibration dataset were compared using the supervised-learning-based SCRL-STM model and the Mel-spectrogram image data in the time–frequency domain obtained by transforming the data. As mentioned above, the dataset obtained by applying data augmentation to the original dataset was used for the raw-signal dataset and Mel-spectrogram image dataset. The model using the 1D raw-signal data as input was constructed using

TABLE 5. Industrial machine dataset samples.

SNR	Original valve dataset		Data augmentation
-6 dB	Normal: 3691	Abnormal: 479	Abnormal: 2395
0 dB	Normal: 3691	Abnormal: 479	Abnormal: 2395
6 dB	Normal: 2699	Abnormal: 479	Abnormal: 2395

SNR	Original pump dataset		Data augmentation
-6 dB	Normal: 3749	Abnormal: 456	Abnormal: 2280
0 dB	Normal: 3749	Abnormal: 456	Abnormal: 2280
6 dB	Normal: 3749	Abnormal: 456	Abnormal: 2280

SNR	Original fan dataset		Data augmentation
-6 dB	Normal: 4075	Abnormal: 1475	Abnormal: 4075
0 dB	Normal: 4075	Abnormal: 1475	Abnormal: 4075
6 dB	Normal: 4075	Abnormal: 1475	Abnormal: 4075

SNR	Original slide rails dataset		Data augmentation
-6 dB	Normal: 3204	Abnormal: 890	Abnormal: 3204
0 dB	Normal: 3204	Abnormal: 890	Abnormal: 3204
6 dB	Normal: 3204	Abnormal: 890	Abnormal: 3204

the 1D CNN layer instead of the 2D CNN layer. The parameters of the SCRLSTM (1D CNN+1D CNN+ Residual LSTM+LSTM) model for using raw signal data as input values are configured differently according to bearing data and industrial datasets. The first 1D convolutional layer for the bearing dataset used 4 kernel sizes and 356 kernels, and the second 1D CNN layer used 10 kernel sizes and 54 kernels. Moreover, the first 1D convolutional layer for the industrial dataset used a kernel size of 362 and the number of kernels of 1032. The second 1D convolutional layer used a kernel size of 10 and 386 kernels. As listed in Table 6, anomaly detection using the data was carried out using the precision, recall, and F1 score, which are the metrics for classification performance. Precision is the ratio of the actual abnormal data to the data classified as abnormal by the model. Recall is the ratio of the actual abnormal data to the data predicted by the model to be abnormal. Furthermore, the F1 score is the harmonic average of precision and recall. The Mel-spectrogram image data of the bearing exhibited excellent performance with higher values of precision, recall, and F1 score than the raw signals of the bearing. Both datasets showed similar results for each load. The difference in the performance between the raw signals and Mel-spectrogram images in the industrial machine dataset was larger than that in the bearing dataset. Furthermore, as the noise signal power ratio of the raw signals and image dataset of the industrial machines considering the SNR decibel values increased, the precision, recall, and F1 score values of the pump and fan data decreased Moreover, the same result was obtained for the slide rails data, indicating a lower performance. In contrast, as the data signal power ratio of the SNR decibel values increased, a better performance result was obtained for the industrial machine dataset. Overall, the performance of the deep-learning-based model using the Mel-spectrogram image was higher with higher precision, recall, and F1 score values than the model using the raw-signal data. It can be concluded that Mel-spectrogram data are less affected by the recognition rates of CNN and LSTM according to noise compared to raw data. In particular, the raw data showed a tendency to show a large change in classification accuracy according to the SNR ratio. Therefore, in other experiments, the accuracy of the fault diagnosis system between the proposed model and other models was compared using the Mel-spectrogram image obtained through the signal processing of the bearing and industrial machine datasets.

Figure 10 shows the receiver operating characteristic (ROC) curve of the SCRLSTM model using the Mel-spectrogram image as the input. The x-axis of the ROC curve indicates the false positive rate, which is the rate at which the model predicts actual normal data as abnormal data, whereas the y-axis indicates the true positive rate, which is the rate at which the model predicts actual abnormal data as normal data. A larger area of the ROC curve indicates higher anomaly detection and classification performance.

F. PERFORMANCE UNDER DIFFERENT LOADS

In this experimental set, the domain adaptation performance of the SCRLSTM model was tested. The experiment was performed to examine the amount of data that was generalized under various load conditions. A dataset with three loads was used in this study (Table 3). Two scenarios were considered. In the first scenario, the model was trained using data collected under one set of load conditions, and the data for different load conditions were tested. In the second scenario, the model was trained using data with two different loads, and the test was performed using a different single load.

1) SCENARIO 1: SINGLE LOAD TO SINGLE LOAD

Table 7 presents information on the accuracy of the model for the dataset used in this scenario. In this scenario, when the dataset was configured, the training and test datasets for each load were combined and trained. Likewise, the verification load dataset for testing the model validation was obtained by combining the training and test datasets. Figure 11 compares the accuracy between the proposed SCRLSTM model and other models including CNN-LSTM based on 2D convolutional layer and LSTM layer, [49], ResNet-SVM model [8], RESNET-SVM model, an ensemble model of SVM model and 2D CNN-based RESNET-18, Snapshot CNN [11] based on LENET-5 using cyclical learning rate scheduler, WDCNN [40] based on 1DCNN and LSTM parallel architecture, SRDCNN [50] model based on residual 1D dilated convolution using the input gate architecture of LSTM. As listed in Table 7 and shown in Figure 11, there were only small differences in the adaptation for each load. The models that contained the LSTM layer, which can reflect the time-series characteristics, showed excellent performance. The model that used only the CNN generally showed lower performance than the model that contained the LSTM layer, which suggests that the model faces difficulty in adapting to other load areas. The model showed the best

TABLE 6. Results of raw signals and mel-spectrogram images.

Data	Bearing								
Load	1 hp			2 hp			3 hp		
Performance Measures	P	R	F1	P	R	F1	P	R	F1
Raw Signal	0.82	0.83	0.82	0.81	0.81	0.81	0.80	0.80	0.80
Mel Image	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Data	Valve								
SNR	-6 dB			0 dB			6 dB		
Performance Measures	P	R	F1	P	R	F1	P	R	F1
Raw Signal	0.65	0.72	0.68	0.75	0.71	0.72	0.72	0.69	0.70
Mel Image	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Data	Pump								
SNR	-6 dB			0 dB			6 dB		
Performance Measures	P	R	F1	P	R	F1	P	R	F1
Raw Signal	0.63	0.68	0.65	0.72	0.73	0.72	0.74	0.74	0.74
Mel Image	0.86	0.88	0.87	0.91	0.93	0.92	0.99	0.99	0.99

Data	Fan								
SNR	-6 dB			0 dB			6 dB		
Performance Measures	P	R	F1	P	R	F1	P	R	F1
Raw Signal	0.62	0.62	0.62	0.72	0.74	0.73	0.74	0.74	0.74
Mel Image	0.81	0.81	0.81	0.94	0.94	0.94	0.99	0.99	0.99

Data	Slide rails								
SNR	-6 dB			0 dB			6 dB		
Performance Measures	P	R	F1	P	R	F1	P	R	F1
Raw Signal	0.65	0.65	0.65	0.65	0.68	0.66	0.74	0.75	0.74
Mel Image	0.92	0.92	0.92	0.95	0.95	0.95	0.98	0.98	0.98

performance in training the 1 hp load, whereas its performance generally dropped when training the 3 hp load.

Moreover, it shows the best performance at 1hp to 2hp and the lowest accuracy at 2hp to 3hp. It can be assumed

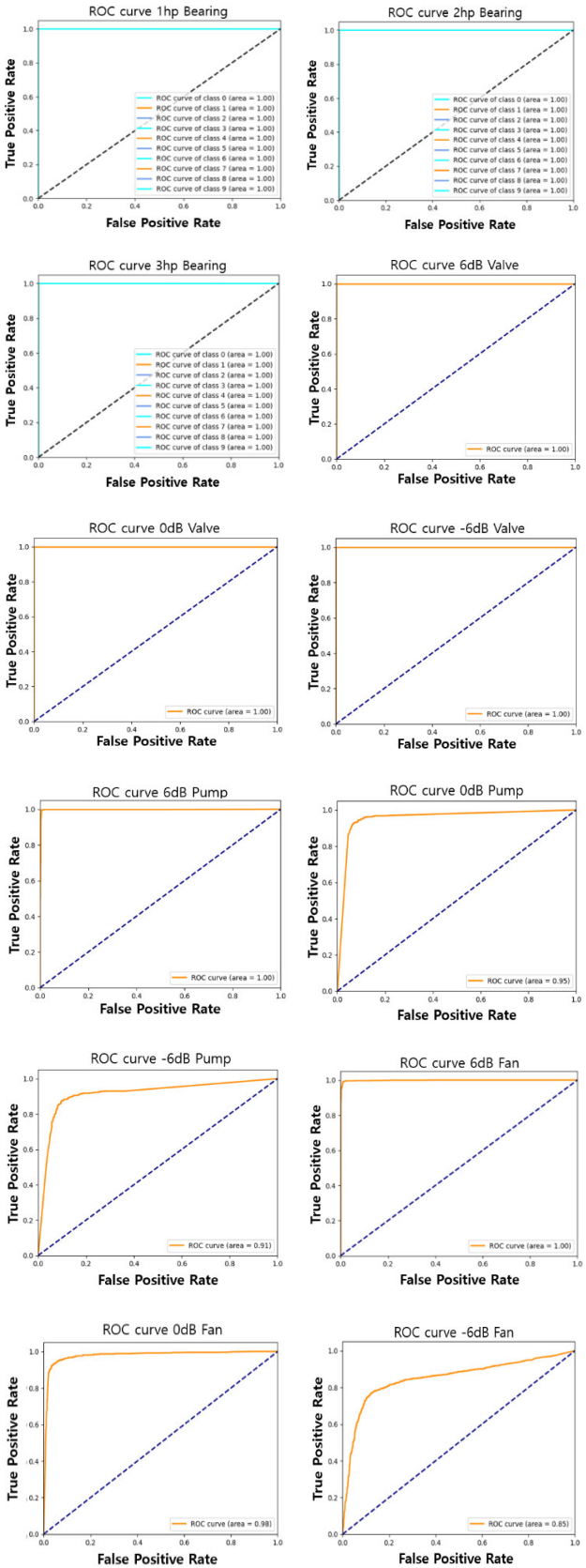


FIGURE 10. ROC curve of mel-spectrogram image datasets.

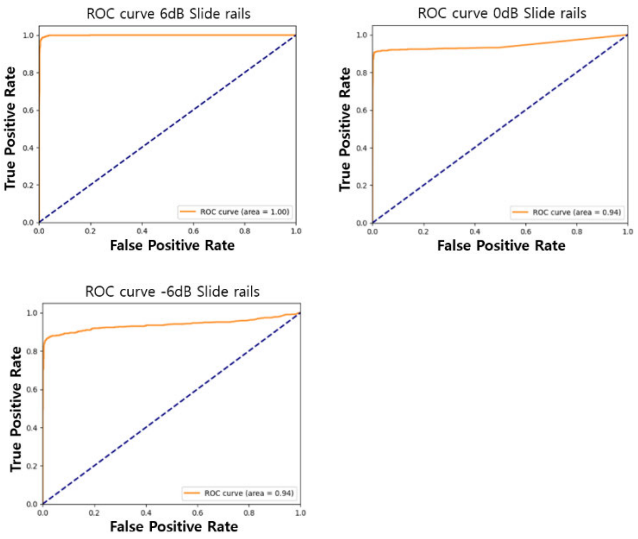


FIGURE 10. *(continued.)* **ROC curve of mel-spectrogram image datasets.**

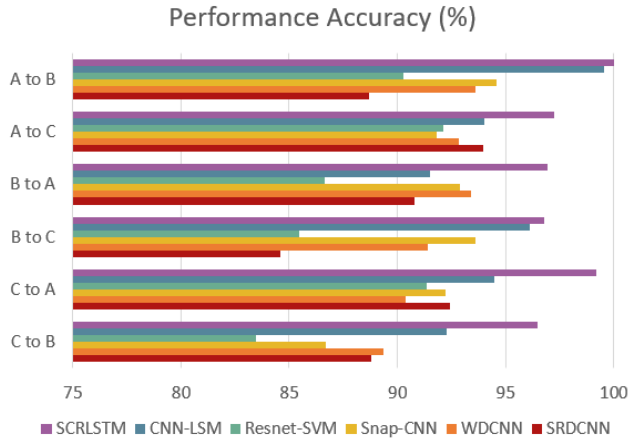


FIGURE 11. **Results for scenario 1 (single load to single load).**

TABLE 7. **Accuracy (%) of models in scenario 1 (single load to single load).**

	SCR LSTM	CNN-LSTM	ResNet-SVM	Snap-CNN	WD CNN	SRD CNN
A to B	100	99.55	90.29	94.58	93.60	88.69
A to C	97.26	94.02	92.13	91.80	92.85	93.96
B to A	96.94	91.52	86.65	92.87	93.39	90.78
B to C	96.80	96.10	85.46	93.59	91.40	84.60
C to A	99.21	94.50	91.35	92.23	90.38	92.45
C to B	96.48	92.63	83.45	86.69	89.36	88.80

that the adaptation experiment of the domain according to each load is not related to the distance for each load. The proposed model generally achieved excellent accuracy in transfer learning for each load compared to the other models. In addition, the proposed SCRLSTM model generally showed excellent results regardless of the load conditions and performance. Figure 12 shows the confusion plots obtained when the proposed model was tested for load characteristics different from the trained load characteristics.

2) SCENARIO 2: MULTIPLE LOADS TO SINGLE LOAD

We considered the case in which the specified data of a different single load is used after training the bearing data for scenario 2 under various load conditions according to an approach used in a previous study[34]. In this case, a powerful fault diagnosis system can be obtained because the model can be trained using data collected under various load conditions. The data in scenario 2 were configured by combining the training and test datasets. Figure 12 shows the confusion plots obtained when the proposed model was tested for a load whose characteristics were different from those of the trained load. Table 8 and Figure 13 shows the results of scenario 2. The data corresponding to 1 and 2 hp after training the model with different loads showed good classification performance. However, the performance for detecting anomalies in the 3 hp data after training the model using the 1hp and 2 hp data was considerably lower. A shift in this domain to 3 hp could presumably impede classification if there are potential additional undiagnosed fault conditions. The results for scenarios 1 and 2 showed that the characteristics of the 3 hp data were different from those of the 1hp and 2 hp data. Furthermore, all the algorithms that were trained using data with various load conditions achieved an accuracy of 80% or higher. The models having the LSTM layer showed a considerably high accuracy, but the ResNet-SVM model showed the lowest accuracy. Overall, the results of the SCRLSTM model were the best for all the loads.

TABLE 8. **Accuracy (%) of models in scenario 2 (multiple loads to single load).**

	SCR LSTM	CNN-LSTM	ResNet-SVM	Snap-CNN	WD CNN	SRD CNN
A&B to C	91.41	90.89	85.20	82.60	90.60	90.88
A&C to B	99.18	96.32	93.53	95.12	94.86	94.86
B&C to A	99.97	98.60	94.76	96.62	97.52	96.20

G. PERFORMANCE UNDER DIFFERENT SNR SIGNALS

Next, the robustness of the proposed model to a noisy environment was investigated. In actual industrial sites, there can be various causes of noise (e.g., accidents in the power grid, abnormal functioning of the inverters and motors of power sources, operation of internal facilities such as inverters and motors, and electric processing). Therefore, an effective fault diagnosis system must be robust to noise generated in industrial sites.

Therefore, in this experiment, based on the results obtained from transfer learning, a number of samples were configured for a dataset that was 100% augmented from the total number of samples by adding various levels of white Gaussian noise in consideration of different SNR signals (−8, −6, −4, −2, 0, 2, 4, 6, and 8 dB) in the 1 hp load learning dataset and the test

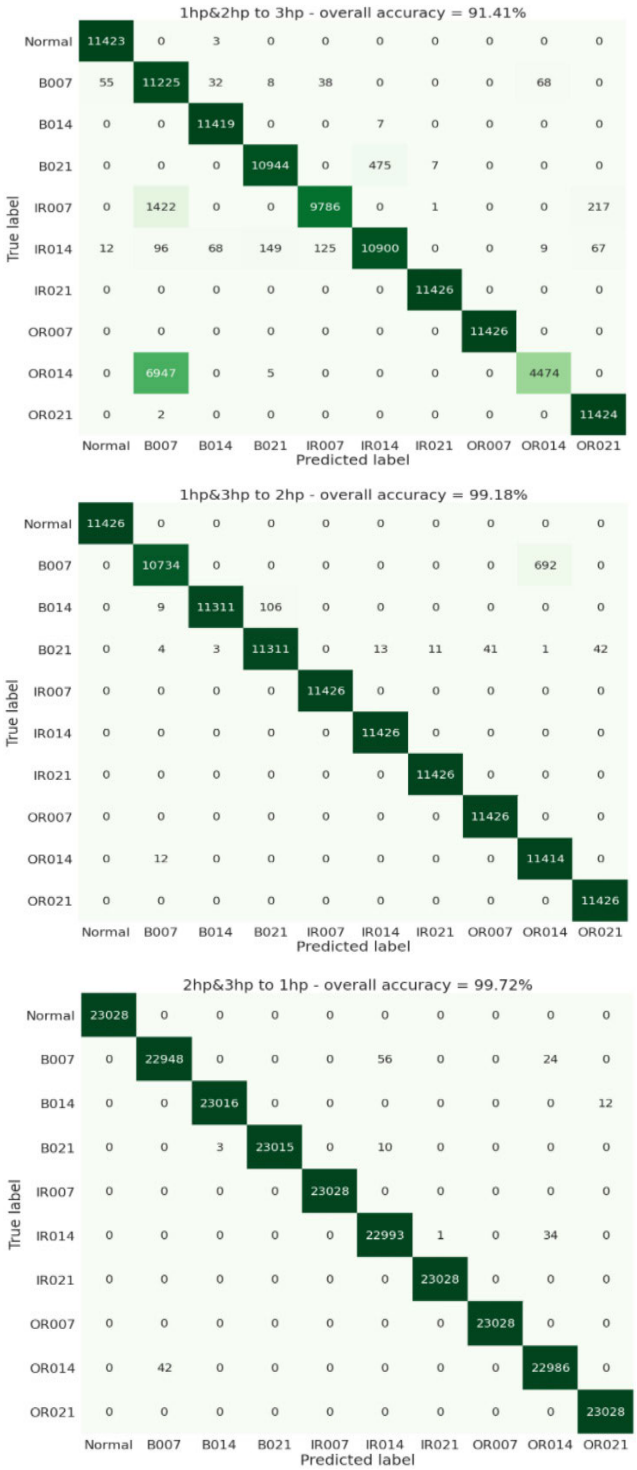


FIGURE 12. Confusion plots obtained using results for scenario 2 (multiple loads to single load).

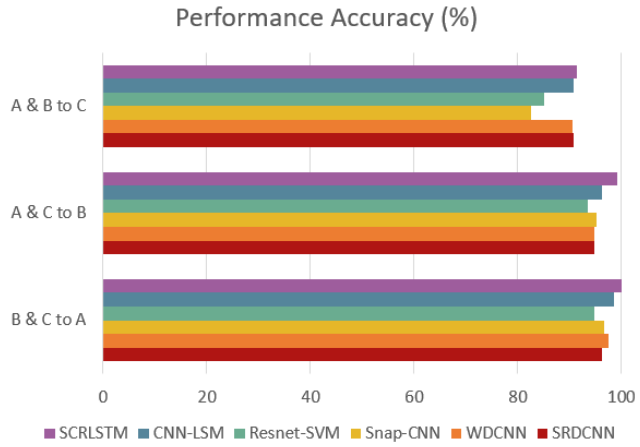


FIGURE 13. Results for scenario 2 (multiple loads to single load).

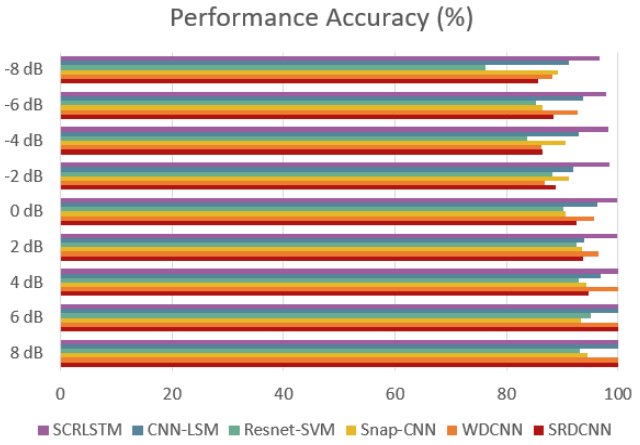


FIGURE 14. Results of different models for noisy dataset.

TABLE 9. Accuracy (%) of models for noisy dataset.

	SCR LSTM	CNN-LSTM	ResNet-SVM	Snap-CNN	WD CNN	SRD CNN
-8 dB	96.78	91.29	76.19	89.17	88.19	85.69
-6 dB	97.95	93.73	85.33	86.48	92.87	88.43
-4 dB	98.30	93.04	83.69	90.59	86.29	86.46
-2 dB	98.55	92.08	88.19	91.23	86.89	88.85
0 dB	99.89	96.32	90.22	90.59	95.64	92.51
2 dB	99.77	93.89	92.64	93.66	96.48	93.76
4 dB	100	96.87	93.05	94.26	100	94.75
6 dB	100	100	95.06	93.27	100	100
8 dB	100	100	93.19	94.59	100	100

dataset representing the load. The experiment was performed by converting the raw-signal data obtained by considering the SNR to a Mel-spectrogram image.

The performance of all the models that diagnose signals with different SNRs is shown in Figure 14 and Table 9. All of these models were trained on 225,000 learning samples which had noise-free and noisy datasets; the accuracies of each model corresponding to the nine SNRs were compared against those of the other models. The classification accuracy of models for a high noise power ratio of −8 dB to −2 dB decreased considerably compared to the SNR value, which generally reduced. Furthermore, the accuracy of most

TABLE 10. Performance (%) of models for industrial machine dataset.

Data	Valve								
SNR	-6dB			0dB			6dB		
Performance Measures	P	R	F1	P	R	F1	P	R	F1
SCR LSTM	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
CNN-LSTM	0.92	0.90	0.93	0.93	0.93	0.93	0.92	0.93	0.92
ResNet-SVM	0.86	0.83	0.84	0.86	0.88	0.86	0.90	0.88	0.88
Snap-CNN	0.89	0.90	0.89	0.90	0.92	0.90	0.89	0.92	0.90
WD CNN	0.93	0.90	0.93	0.93	0.93	0.93	0.93	0.95	0.95
SRD CNN	0.93	0.93	0.93	0.94	0.93	0.93	0.94	0.95	0.94

Data	Pump								
SNR	-6dB			0dB			6dB		
Performance Measures	P	R	F1	P	R	F1	P	R	F1
SCR LSTM	0.86	0.88	0.87	0.91	0.93	0.92	0.99	0.99	0.99
CNN-LSTM	0.82	0.84	0.83	0.88	0.88	0.88	0.89	0.90	0.90
ResNet-SVM	0.85	0.86	0.85	0.86	0.87	0.86	0.88	0.88	0.88
Snap-CNN	0.84	0.86	0.85	0.84	0.86	0.85	0.87	0.88	0.87
WD CNN	0.83	0.85	0.84	0.90	0.91	0.90	0.91	0.92	0.91
SRD CNN	0.88	0.89	0.88	0.91	0.92	0.91	0.86	0.87	0.86

Data	Fan								
SNR	-6dB			0dB			6dB		
Performance Measures	P	R	F1	P	R	F1	P	R	F1
SCR LSTM	0.81	0.81	0.81	0.94	0.94	0.94	0.99	0.99	0.99
CNN-LSTM	0.80	0.80	0.80	0.88	0.88	0.88	0.92	0.93	0.92
ResNet-SVM	0.76	0.78	0.77	0.84	0.86	0.85	0.84	0.84	0.84
Snap-CNN	0.80	0.80	0.80	0.84	0.86	0.85	0.89	0.90	0.89
WD CNN	0.76	0.78	0.77	0.92	0.94	0.93	0.95	0.96	0.95
SRD CNN	0.72	0.74	0.73	0.85	0.87	0.86	0.96	0.98	0.97

models decreased when the SNR value decreased to −2dB to −8 dB, where the power ratio of the signal was higher than the noise power ratio. The model accuracy improved

TABLE 10. *(Continued.)* Performance (%) of models for industrial machine dataset.

Data	Slide rails								
SNR	-6dB			0dB			6dB		
Performance Measures	P	R	F1	P	R	F1	P	R	F1
SCR LSTM	0.92	0.92	0.92	0.95	0.95	0.95	0.98	0.98	0.98
CNN-LSTM	0.91	0.92	0.91	0.90	0.90	0.90	0.93	0.94	0.93
ResNet-SVM	0.83	0.83	0.83	0.86	0.88	0.87	0.90	0.90	0.90
Snap-CNN	0.84	0.86	0.85	0.92	0.92	0.92	0.93	0.93	0.93
WD CNN	0.86	0.87	0.86	0.91	0.93	0.92	0.93	0.94	0.93
SRD CNN	0.88	0.90	0.89	0.92	0.93	0.92	0.94	0.94	0.94

when the SNR value increased with the power ratio of the signals.

In summary, CWRU bearing data presents indications of internal and external raceway faults, and noise is applied to this dataset to obscure the actual fault condition, which can interfere with the anomaly detection classification performance. However, the SCRLSTM model performed well in an environment with many types of noise and had the best performance compared to other models when the noise power ratio was high. Even in an environment with noise, the model having the LSTM layer showed good efficiency for noise removal. The accuracy of different models for each SNR ratio is shown in Figure 14.

H. ANOMALY DETECTION FOR INDUSTRIAL MACHINE

The anomaly detection model was investigated for a time-series-based vibration industrial machine dataset. The implementation of the deep-learning-based model using the given dataset as the input considerably decreased the accuracy because of data imbalance, and most abnormal data were detected as normal data during the verification. Thus, as listed in Table 5, the data imbalance problem was solved through data augmentation. The overall accuracy improved when the number of normal and abnormal data points was similar. Table 10 compares the five models to demonstrate the performance of the SCRLSTM model. The proposed model showed good performance under various types of noise when the SNR was −6, 0, and +6 dB for the valve, pump, fan, and slide rails data. The classification performance of the proposed model is shown in Figure 15. As in the case of the noise experiment for the bearing, a higher performance was obtained when the signal power was higher. At −6 dB with the highest noise power, most models showed low accuracy. Furthermore, the algorithm using the LSTM model generally achieved high performance, but the ResNet-SVM model

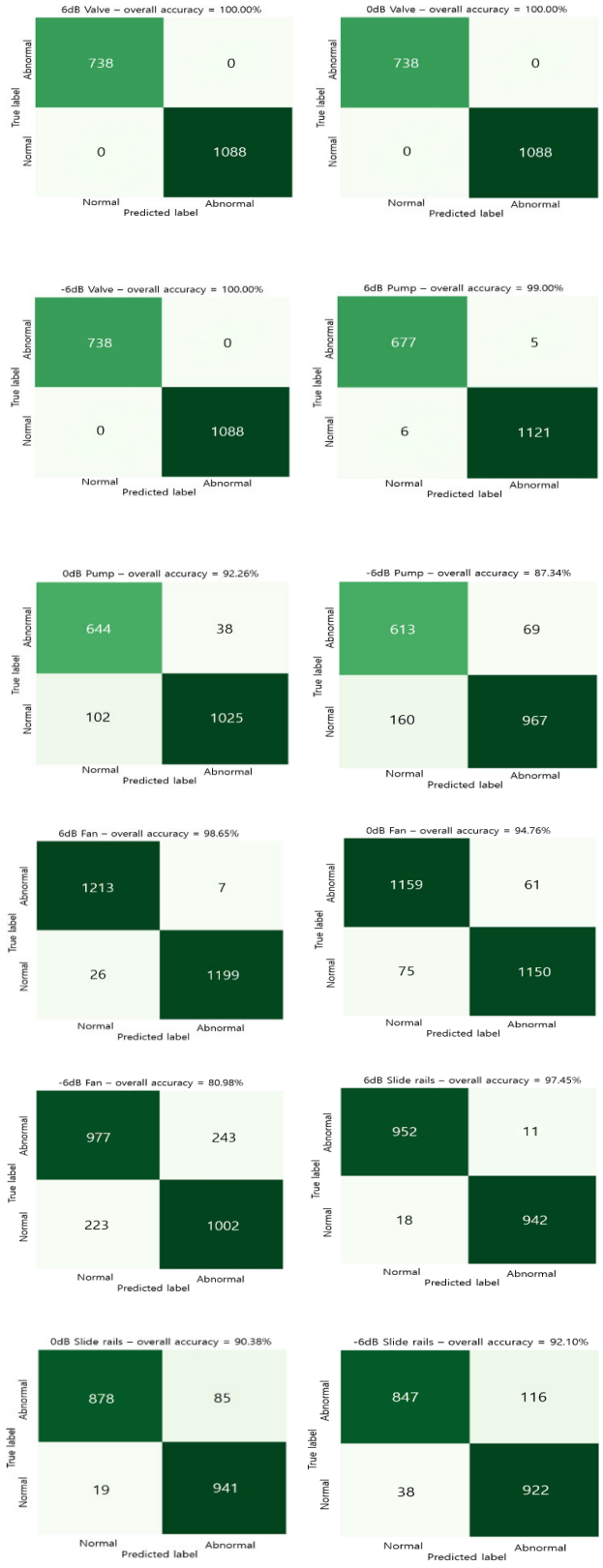


FIGURE 15. Results of SCRLSTM model for industrial machine dataset.

showed the lowest accuracy. In the industrial machine dataset, there was a difference in classification performance according to the noise effect. The smallest performance difference was found in the Valve dataset, and a large difference existed in the Fan dataset. Compared to the Valve dataset, the Fan dataset is estimated to be the most vulnerable to noise.

V. CONCLUSION

The main contribution of this study is that it overcomes the data imbalance problem in the time-series-based data for bearings and industrial machines through data augmentation. Also, faults can be detected early by considering various types of noise in actual industrial environments. Furthermore, it lends itself to the development of a data-based, intelligent fault diagnosis system to improve the productivity of the aforementioned equipment by using Mel-spectrogram that contain various features obtained from raw signals. The proposed SCRLSTM ensemble model, which combines the CNN and LSTM, not only provides state-of-the-art classification performance but also solves many problems encountered in existing intelligent data-based fault diagnosis techniques, such as variations in time-series-based data and load and robustness to noise. The SCRLSTM ensemble model proposed in this study was validated using datasets for bearings and industrial machines (valves, pumps, fans, and slide rails). The model outperforms existing supervised-learning-based fault diagnosis methods in various actual environments. The proposed model can extract the spatial features of mel-spectrogram images that show the characteristics of conspicuous noise through 2D and 1D CNNs. Moreover, it can efficiently detect normal and abnormal data by extracting the features of time-series-based vibration datasets through residual LSTM and LSTM layers.

To verify the performance of our proposed SCRLSTM model, we plan to conduct future experiments by adding noises under various conditions to the datasets of other types of mechanical equipment such as motors and gearboxes.

REFERENCES

[1] Y. Li, Y. Yang, G. Li, M. Xu, and W. Huang, "A fault diagnosis scheme for planetary gearboxes using modified multi-scale symbolic dynamic entropy and mRMR feature selection," *Mech. Syst. Signal Process.*, vol. 91, pp. 295–312, Jul. 2017, doi: 10.1016/j.ymssp.2016.12.040.

[2] T. Lu, F. Yu, B. Han, and J. Wang, "A generic intelligent bearing fault diagnosis system using convolutional neural networks with transfer learning," *IEEE Access*, vol. 8, pp. 164807–164814, 2020, doi: 10.1109/ACCESS.2020.3022840.

[3] A. Varga, "Fault diagnosis," *Stud. Syst. Decis. Control*, vol. 84, no. 3, pp. 27–56, 2017, doi: 10.1007/978-3-319-51559-5_3.

[4] M. Munir, M. A. Chattha, A. Dengel, and S. Ahmed, "A comparative analysis of traditional and deep learning-based anomaly detection methods for streaming data," in *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2019, pp. 561–566, doi: 10.1109/ICMLA.2019.00105.

[5] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM Comput. Surveys*, vol. 54, no. 2, pp. 1–38, Apr. 2021, doi: 10.1145/3439950.

[6] H. Purohit, R. Tanabe, K. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," in *Proc. Detection Classification Acoustic Scenes Events Workshop*, 2019, pp. 209–213, doi: 10.33682/m76f-d618.

[7] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 271–275, doi: 10.1109/ICASSP40776.2020.9054344.

[8] L. Yuan, D. Lian, X. Kang, Y. Chen, and K. Zhai, "Rolling bearing fault diagnosis based on convolutional neural network and support vector machine," *IEEE Access*, vol. 8, pp. 137395–137406, 2020, doi: 10.1109/ACCESS.2020.3012053.

[9] Y. Yang, P. Fu, and Y. He, "Bearing fault automatic classification based on deep learning," *IEEE Access*, vol. 6, pp. 71540–71554, 2018, doi: 10.1109/ACCESS.2018.2880990.

[10] A. Zhang, S. Li, Y. Cui, W. Yang, R. Dong, and J. Hu, "Limited data rolling bearing fault diagnosis with few-shot learning," *IEEE Access*, vol. 7, pp. 110895–110904, 2019, doi: 10.1109/ACCESS.2019.2934233.

[11] L. Wen, L. Gao, and X. Li, "A new snapshot ensemble convolutional neural network for fault diagnosis," *IEEE Access*, vol. 7, pp. 32037–32047, 2019, doi: 10.1109/ACCESS.2019.2903295.

[12] C. Li, W. Zhang, G. Peng, and S. Liu, "Bearing fault diagnosis using fully-connected winner-take-all autoencoder," *IEEE Access*, vol. 6, pp. 6103–6115, 2017, doi: 10.1109/ACCESS.2017.2717492.

[13] Z.-H. Liu, B.-L. Lu, H.-L. Wei, X.-H. Li, and L. Chen, "Fault diagnosis for electromechanical drivetrains using a joint distribution optimal deep domain adaptation approach," *IEEE Sensors J.*, vol. 19, no. 24, pp. 12261–12270, Dec. 2019, doi: 10.1109/JSEN.2019.2939360.

[14] T. Han, Y.-F. Li, and M. Qian, "A hybrid generalization network for intelligent fault diagnosis of rotating machinery under unseen working conditions," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021, doi: 10.1109/tim.2021.3088489.

[15] Y. Koizumi, S. Murata, N. Harada, S. Saito, and H. Uematsu, "SNIPER: Few-shot learning for anomaly detection to minimize false-negative rate with ensured true-positive rate," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 915–919, doi: 10.1109/ICASSP.2019.8683667.

[16] Y. Kawachi, Y. Koizumi, S. Murata, and N. Harada, "A two-class hyper-spherical autoencoder for supervised anomaly detection," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2019, pp. 3047–3051.

[17] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 2, pp. 379–393, Feb. 2018, doi: 10.1109/TASLP.2017.2778423.

[18] S. S. R. Phaye, E. Benetos, and Y. Wang, "SubSpectralNet–using sub-spectrogram based convolutional neural networks for acoustic scene classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2019, pp. 825–829.

[19] M. Cheung, J. Shi, O. Wright, L. Y. Jiang, X. Liu, and J. M. F. Moura, "Graph signal processing and deep learning: Convolution, pooling, and topology," *IEEE Signal Process. Mag.*, vol. 37, no. 6, pp. 139–149, Nov. 2020, doi: 10.1109/MSP.2020.3014594.

[20] Y. M. Galvão, J. Ferreira, V. A. Albuquerque, P. Barros, and B. J. T. Fernandes, "A multimodal approach using deep learning for fall detection," *Expert Syst. Appl.*, vol. 168, Apr. 2021, Art. no. 114226, doi: 10.1016/j.eswa.2020.114226.

[21] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994, doi: 10.1109/72.279181.

[22] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, 2000, doi: 10.1162/089976600300015015.

[23] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, and J. Klingner, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*. [Online]. Available: http://arxiv.org/abs/1609.08144

[24] C. Chen, K. Li, S. G. Teo, X. Zou, K. Wang, J. Wang, and Z. Zeng, "Gated residual recurrent graph neural networks for traffic prediction," in *Proc. 33rd AAAI Conf. Artif. Intell. (AAAI), 31st Innov. Appl. Artif. Intell. Conf. (IAAI), 9th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, 2019, pp. 485–492, doi: 10.1609/aaai.v33i01.3301485.

[25] A. Kusupati, M. Singh, K. Bhatia, A. Kumar, P. Jain, and M. Varma, "FastgRNN: A fast, accurate, stable and tiny kilobyte sized gated recurrent neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2018, pp. 9017–9028.

[26] T. Han, C. Liu, L. Wu, S. Sarkar, and D. Jiang, "An adaptive spatiotemporal feature learning approach for fault diagnosis in complex systems," *Mech. Syst. Signal Process.*, vol. 117, pp. 170–187, Feb. 2019, doi: 10.1016/j.ymssp.2018.07.048.

[27] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," Jan. 2018, *arXiv:1811.03378*. [Online]. Available: http://arxiv.org/abs/1811.03378

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 25, Dec. 2012, pp. 1097–1105.

[29] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Jan. 2017. [Online]. Available: https://arxiv.org/pdf/1511.00561.pdf

[30] M. Lin, Q. Chen, and S. Yan, "Network in network," in *Proc. 2nd Int. Conf. Learn. Represent. Conf. Track (ICLR)*, 2014, pp. 1–10.

[31] G. Shomron and U. Weiser, "Spatial correlation and value prediction in convolutional neural networks," *IEEE Comput. Archit. Lett.*, vol. 18, no. 1, pp. 10–13, Jun. 2019, doi: 10.1109/LCA.2018.2890236.

[32] D. Neupane and J. Seok, "Bearing fault detection and diagnosis using case western reserve university dataset with deep learning approaches: A review," *IEEE Access*, vol. 8, pp. 93155–93178, 2020, doi: 10.1109/ACCESS.2020.2990528.

[33] S. Zhang, S. Zhang, B. Wang, and T. G. Habetler, "Deep learning algorithms for bearing fault diagnostics—A comprehensive review," *IEEE Access*, vol. 8, pp. 29857–29881, 2020, doi: 10.1109/ACCESS.2020.2972859.

[34] Z. Guangquan, W. Kankan, G. Yongcheng, L. Yongmei, and H. Cong, "Bearing fault diagnosis from raw vibration signals using multi-layer extreme learning machine," in *Proc. 14th IEEE Int. Conf. Electron. Meas. Instrum. (ICEMI)*, no. 2, Nov. 2019, pp. 1287–1293, doi: 10.1109/ICEMI46757.2019.9101840.

[35] 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, doi: 10.1016/j.ymssp.2015.04.021.

[36] Q. Wen, L. Sun, F. Yang, X. Song, J. Gao, X. Wang, and H. Xu, "Time series data augmentation for deep learning: A survey," 2020, *arXiv:2002.12478*. [Online]. Available: http://arxiv.org/abs/2002.12478

[37] S. Li, G. Liu, X. Tang, J. Lu, and J. Hu, "An ensemble deep convolutional neural network model with improved D-S evidence fusion for bearing fault diagnosis," *Sensors*, vol. 17, no. 8, p. 1729, Jul. 2017, doi: 10.3390/s17081729.

[38] X. Wang and F. Liu, "Triplet loss guided adversarial domain adaptation for bearing fault diagnosis," *Sensors*, vol. 20, no. 1, p. 320, Jan. 2020, doi: 10.3390/s20010320.

[39] H. Zhu, Z. He, J. Wei, J. Wang, and H. Zhou, "Bearing fault feature extraction and fault diagnosis method based on feature fusion," *Sensors*, vol. 21, no. 7, pp. 1–25, 2021, doi: 10.3390/s21072524.

[40] A. Shenfield and M. Howarth, "A novel deep learning model for the detection and identification of rolling element-bearing faults," *Sensors*, vol. 20, no. 18, pp. 1–24, 2020, doi: 10.3390/s20185112.

[41] G. Kovács, L. Tóth, D. Van Compernolle, and S. Ganapathy, "Increasing the robustness of CNN acoustic models using autoregressive moving average spectrogram features and channel dropout," *Pattern Recognit. Lett.*, vol. 100, pp. 44–50, Dec. 2017, doi: 10.1016/j.patrec.2017.09.023.

[42] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3D log-mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019, doi: 10.1109/ACCESS.2019.2938007.

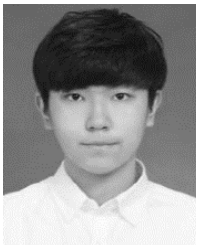
[43] S. Abbasi, M. Famouri, M. J. Shafiee, and A. Wong, "OutlierNets: Highly compact deep autoencoder network architectures for on-device acoustic anomaly detection," 2021, *arXiv:2104.00528*. [Online]. Available: http://arxiv.org/abs/2104.00528

[44] T. Tran and J. Lundgren, "Drill fault diagnosis based on the scalogram and mel spectrogram of sound signals using artificial intelligence," *IEEE Access*, vol. 8, pp. 203655–203666, 2020, doi: 10.1109/ACCESS.2020.3036769.

[45] J. Kim, H. Lee, S. Jeong, and S.-H. Ahn, "Sound-based remote real-time multi-device operational monitoring system using a convolutional neural network (CNN)," *J. Manuf. Syst.*, vol. 58, pp. 431–441, Jan. 2021, doi: 10.1016/j.jmsy.2020.12.020.

[46] R. O'Brien and H. Ishwaran, "A random forests quantile classifier for class imbalanced data," *Pattern Recognit.*, vol. 90, pp. 232–249, Jun. 2019, doi: 10.1016/j.patcog.2019.01.036.

[47] P. Ganesan, S. Rajaraman, R. Long, B. Ghoraani, and S. Antani, "Assessment of data augmentation strategies toward performance improvement of abnormality classification in chest radiographs," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 841–844, doi: 10.1109/EMBC.2019.8857516.
[48] S. S. Channappayya, A. C. Bovik, and R. W. Heath, Jr., "Rate bounds on SSIM index of quantized images," *IEEE Trans. Image Process.*, vol. 17, no. 9, pp. 1624–1639, Sep. 2008, doi: 10.1109/TIP.2008.2001400.
[49] H. Xie, L. Zhang, and C. P. Lim, "Evolving CNN-LSTM models for time series prediction using enhanced grey wolf optimizer," *IEEE Access*, vol. 8, pp. 161519–161541, 2020, doi: 10.1109/ACCESS.2020.3021527.
[50] Z. Zhuang, H. Lv, J. Xu, Z. Huang, and W. Qin, "A deep learning method for bearing fault diagnosis through stacked residual dilated convolutions," *Appl. Sci.*, vol. 9, no. 9, p. 1823, May 2019, doi: 10.3390/app9091823.



GEONKYO HONG received the B.S. degree from the School of Convergence and Fusion System Engineering, Kyungpook National University (KNU), South Korea. He is currently pursuing the M.S. degree with the Department of Convergence and Fusion System Engineering.

His research interests include machine learning and deep learning, anomaly detection, intelligent fault diagnosis, and signal processing.

DONGJUN SUH (Member, IEEE) received the Ph.D. degree from the Department of Civil and Environmental Engineering (construction-ICT convergence track), in 2014. From 2007 to 2010, he worked as a Software Engineer at Humax. From 2014 to 2015, he was a Research Assistant Professor at KAIST Institute for Information Technology Convergence. From 2015 to 2018, he was a Senior Researcher at Korea Institute of Science and Technology Information (KISTI). From 2017 to 2018, he was an Associate Professor at the Department of Science Technology Information Science, University of Science and Technology (UST). He is currently an Assistant Professor at the Department of Convergence and Fusion System Engineering, Kyungpook National University (KNU), South Korea. His current research interests include big data and smart control systems using predictive analytics encompassing a variety of statistical techniques based on modeling, machine learning, data mining, and various theories that analyze current and historical facts to make predictions about future events. He has considerable experience related to ICT-based building science, smart grids/microgrids, energy systems, disaster risk analysis, and ICT convergence technology, including smart systems, and has published many scientific articles in refereed journals and given presentations at many international scientific as well as technical conferences.