



# Hybrid deep transfer learning architecture for industrial fault diagnosis using Hilbert transform and DCNN–LSTM

Mahe Zabin<sup>1</sup> · Ho-Jin Choi<sup>1</sup>  · Jia Uddin<sup>2</sup>

Accepted: 11 September 2022 / Published online: 12 October 2022  
© The Author(s) 2022

## Abstract

Early-stage fault detection has become an indispensable part of modern industry to prevent potential hazards or sudden hindrances to the production process. With the advent of deep learning (DL) applications in several fields, DL models have been used to classify faults in specific environments. Uniform texture extraction has been performed using transformed-signal processing techniques and deep transfer learning (DTL) architectures in a few studies. Traditional signal processing techniques encounter difficulties in extracting distinct fault features due to the nonlinear and non-stationary nature of the time-series fault data. In this paper, a hybrid DTL architecture comprising a deep convolutional neural network and long short-term memory layers for extracting both temporal and spatial features enhanced by Hilbert transform 2D images is presented. Three standard audio sound fault datasets comprising the malfunctioning industrial machine investigation and inspection dataset, toy anomaly detection in machine operating sounds dataset, and machinery failure prevention technology bearing vibration fault dataset with various loads and noisy environments were utilized in the experimental evaluation. The proposed model with an input size of  $32 \times 32$  achieved an average *F1* score of 0.998 on the tested datasets. The implementation of transfer learning using the three benchmark datasets resulted in the highest accuracy of the proposed model and over fivefold reduction in the training epochs. In addition, the proposed model outperformed the state-of-art models in accuracy in various environments.

**Keywords** Fault diagnosis · Deep CNN-LSTM · Transfer learning · Classification accuracy

---

✉ Ho-Jin Choi  
hojinc@kaist.ac.kr

Extended author information available on the last page of the article

## 1 Introduction

In Industry 4.0 environments, the automation of industrial equipment supervision is required to maintain the desired level of performance. Through fault diagnosis, the current status of the equipment can be determined and abnormal conditions can be identified. Slow, inaccurate, or ineffective fault diagnosis may degrade the performance of the entire system and cause unexpected losses. Although real-time diagnosis models can allow abnormal conditions (faults) in equipment to be identified in the early stages, automatic fault diagnosis is always a challenging task because of the types of equipment used in industrial environments [1].

In general, various Internet of Things sensors, such as acoustic emission, vibration, current, voltage, thermal, and pressure sensors, can be used to monitor equipment. All the sensor data are one-dimensional time-series signals in which information is represented through the signal amplitude, phase, and frequency (as depicted in Fig. 2 in Sect. 2.1). These data are used to identify the type, size, and location of the fault. There are three basic steps in fault diagnosis models based on sensor data. Step 1 comprises data collection from the equipment, step 2 involves the application of signal processing techniques to extract features from the sensor data, and step 3 is the use of classifiers in the final stage [2].

Various researchers have studied transformed-signal processing methods for extracting uniform texture information from fault signals under various conditions. In [3], the discrete orthonormal Stockwell transform was applied to extract identical 2D patterns for each fault signal. Its high computation cost is infeasible for most applications while the time–frequency space partitioning leads to symmetry loss. In [4], bi-spectrum-based higher-order analysis was utilized to extract distinct signal patterns under inconsistent working conditions. Although the traditional bisector representation permits the phase information to be included and eliminates Gaussian noise, the results are unstable because of the randomly changing phase components of the signals. 2D acoustic spectral imaging [5] and local binary pattern (LBP) [6] techniques were applied for extracting uniform patterns from signals. The limitation of LBP is that identical LBP codes are generated for various structural patterns. Short-time Fourier transform (STFT) [7] and a Gabor filter with singular value decomposition (SVD) [8] were applied to acoustic emission signals for extracting the uniform texture patterns of the signals. However, the STFT and Gabor filter took too long to complete.

These state-of-the-art machine learning methods have certain drawbacks. For example, the development of a feature extractor requires domain expertise and skill in signal processing techniques, and the extraction function does not follow the same protocol in every application.

Deep learning (DL) models are used in fault diagnosis, where complex deep features are extracted from the raw data using several hidden layers without human intervention. Xia et al. [9] proposed a convolutional neural network (CNN) architecture for fault detection and evaluated its performance on the Case Western Reserve University (CWRU) bearing dataset and a gearbox dataset. In [10], deep neural network, deep belief network, and CNN were utilized for fault

diagnosis and prognosis. The CNN did not encode the position and orientation of the objects and was invariant with respect to the input data. In [11], a novel fault diagnosis framework based on an end-to-end LSTM model was proposed to learn the features directly from multivariate time-series data and capture the long-term dependencies through the recurrent behavior and gates mechanisms of the LSTM. The RNN and LSTM were difficult to train because they required memory-bandwidth-bound computation in which the LSTM required four linear MLP layers per cell. Peng et al. proposed a model using 1D CNN and 1D residual blocks. The model was experimentally shown to work well despite the presence of strong noise and variable loads [12].

The preceding studies demonstrate that DL models can perform automated deep feature extraction from the raw fault data. However, the models have some limitations such as the increase in the number of parameters with the increasing number of layers and the computational burden of training large networks with huge parameters from scratch requires a massive amount of label data. The performance of these DL models is also affected by parameter optimization and hyperparameter tuning.

To overcome the limitations of DL models, transfer learning (TL) was incorporated into DL models so that knowledge from one problem can be used to solve a related different problem [4, 13–20]. For example, pre-trained ImageNet weights have been used to initialize the parameters in deep networks for classifying industrial faults instead of random initialization. Although the deep transfer learning (DTL) model parameters are initialized using pre-trained weights, a large training target dataset is still required for parameter optimization in complex architectures to improve their accuracy. Optimizing the parameter initialization is a vital consideration in ensuring the accuracy of DL-based fault diagnosis.

## 1.1 Literature review

TL models have recently been applied in machine fault diagnosis. In [13], a new temporal CNN with a depth of 51 convolutional layers was applied, with ResNet-50 trained using ImageNet as a feature extractor. Wen et al. used the VGG-19 architecture in [14] and Inception V3 and TrAdaBoost as feature extractors in [15]. Grover et al. [4] utilized bi-spectrum contour maps of the vibration signals in four pre-trained networks comprising Alexnet, VGG-19, GoogleNet, and Resnet-50. These architectures are limited by their large number of layers and high computational complexity. In addition, since TrAdaBoost depends on only a single source, its learning effects degrade when the source and target domains are weakly correlated.

Transfer component analysis (TCA) [16] and weighted transfer component analysis (WTCA) [17] have been used for fault diagnosis in rolling bearings under variable operating conditions. The large number of super-parameters in TCA/WTCA led to difficulties during model training.

In [18], a sparse auto-encoder with three layers was used to scrutinize the raw data and extract features. The maximum mean discrepancy was applied as the discrepancy penalty to be minimized between the source and target data. In [19], a VGG16 pre-trained network was used to extract the lower-level features and label

wavelet transform images. Fan et al. [20] implemented TL in a CNN by generating texture images using empirical mode decomposition with the pseudo-Wigner–Ville distribution. The state-of-art DTL models used pre-trained ImageNet weights for implementing TL. Although there are 1000 object classes in the ImageNet dataset, the classes are indirectly related to the target fault domain. In addition, the pre-trained DTL architectures consist of significantly more layers and trainable parameters than conventional DL architectures.

## 1.2 Contributions

Motivated by the earlier work, Peng et al. [12], in this study, a texture-based 2D hybrid deep-CNN–LSTM architecture was investigated because deep LSTM layers with a deep CNN architecture can learn the features of fault signals adaptively. Similar to the approach in Hasan et al. [21], in this study, DL was implemented by splitting the datasets into two subsets comprising the source and target task datasets for various environments. The source task-dataset is used for training and validating the model to save the weights of the deep architecture and then the target task dataset is used for testing the classification accuracy using the weights of the source task dataset. Unlike state-of-the-art TL models, pre-trained weights were not used to initialize the architectures in this study.

The contributions of this study are summarized as follows:

- First, we explain how Hilbert transform images based on analytical signals demonstrate invariant image patterns for fault signals using the benchmark fault datasets. In addition, Hilbert transform analytical imaging is compared with state-of-art methods comprising discrete wavelet transform (DWT), fast Fourier transform (FFT), and gammatone spectrogram-based texture approaches.
- Second, a 2D DTL-based CNN–LSTM hybrid architecture for fault classification is presented. The proposed architecture is compared with conventional deep architectures such as the DCNN and deep LSTM. In our hybrid architecture, the additional LSTM blocks and DCNN can adaptively learn fault features more accurately and the complexity of the hybrid architecture is significantly lower than those of the state-of-art architectures used in TL models. For example, the ratio of trainable parameters in VGG16 to that in the proposed architecture is 12.65:1.
- Finally, the effectiveness of the proposed hybrid TL model is evaluated in various environments with different noise, loads, and machines using audio records of machine anomalies in three public benchmark datasets of different sizes comprising the malfunctioning industrial machine investigation and inspection (MFPT) dataset for bearing vibration signals and the malfunctioning industrial machine investigation and inspection (MIMII) and toy anomaly detection in machine operating sounds (ToyADAMOS) datasets. The proposed hybrid architecture with TL demonstrated better performance for the vibration and audio fault datasets by achieving higher *F1* scores with approximately five times less epochs.

### 1.3 Outline

The rest of the paper is organized as follows. In Sect. 2, the detailed architecture of the proposed hybrid DTL architecture is presented. The experimental results are discussed in Sect. 3. The paper is concluded in Sect. 4.

## 2 Proposed fault diagnosis methodology

The detailed steps in the proposed model are shown in Fig. 1. The model consists of three major blocks comprising dataset preparation with various loads, signal-to-noise ratios (SNRs), and revolutions per minute (RPMs); Hilbert transform 2D image generation; and a deep CNN–LSTM hybrid architecture as a classifier. In addition, to reduce the computational complexity, TL was implemented in the proposed model by dividing the datasets into the source and target task datasets. A brief description of each model block is presented in the following subsections.

### 2.1 Test rig and data descriptions

Three standard public fault datasets comprising the MFPT vibration fault dataset and the MIMII and ToyADAMOS machine audio fault datasets were used to evaluate the proposed model. The MFPT dataset was collected from a NICE bearing with 0.235 roller diameter, 1.245 pitch diameter, eight elements, and 0° contact angle [22]. There are two environments in the MFPT dataset corresponding to loads of 50–150 lbs and 200–300 lbs and two classes of faults comprising inner and outer race signals, as summarized in Table 1. There are 429 samples with a sample size of 1024 for each type of signal in each environment.

MIMII is an industrial sound dataset in which sounds corresponding to different anomalies comprising contamination, leakage, unbalanced rotation, and rail damage were collected with background noise from four machines comprising a fan, pump,

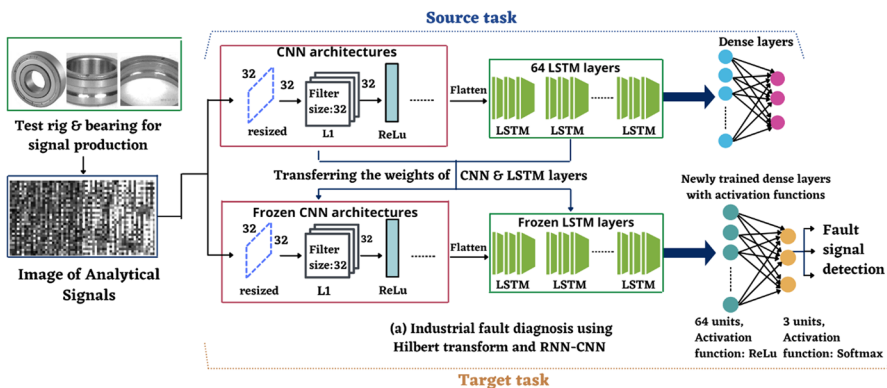


Fig. 1 Step-by-step representation of the proposed model

**Table 1** Details of various load environments in MFPT dataset

Environment	Load	Sample size (1024)	
		Inner	Outer
Environment 1/2	50–150/200–300	429	429

valve, and slide rail in an actual factory [23]. Eight-channel microphone arrays with a sampling rate of 16 kHz placed 45° apart from one another were used in the collection rig. Among the eight microphones, the sound from the microphone nearest to each machine was used for the machine in the dataset. For example, the sound from the microphone at 180° was used for the fan, that from the microphone at 270° for the slide rail, that from the microphone at 0° for the valve, and that from the microphone at 90° for the pump. To construct the dataset for validating the model, the.wav audio files were converted into.mat files, and each.mat file was resized to a length of 1024. Three environments with SNRs of –6 dB, 0 dB, and +6 dB are included in the MIMII dataset, as summarized in Table 2. There are eight classes of faults in the MIMII dataset comprising the normal and abnormal conditions for the fan, pump, slider, and valve. The dataset contains 400 signals for the normal fan, pump, slider, valve, and abnormal fan, and 143, 356, and 119 signals for the abnormal pump, slider, and valve, respectively.

ToyADAMOS is a machine operating sound dataset that was collected from four microphones at a sampling rate of 48,000 Hz [24]. There are two types of sounds comprising normal and anomalous sounds for three different toy machines comprising a toy car (machine-condition inspection), toy conveyor (fixed), and toy train (moving). The dataset contains sounds for three different environments denoted as case1, case2, and case3. Each case contains a total of 72,000 individual samples, which includes normal and anomalous sounds from the toy car, toy conveyor, and toy train, as summarized in Table 3. The anomalous sounds were collected by damaging the machine components or adding additional objects. The three cases in the toy car data were generated by changing the motor and bearing, those in the toy conveyor cases were generated with three different sizes of machines, and those in the toy train cases were generated with

**Table 2** Details of environments with various SNRs in MIMII dataset

Environment	SNR (DB)	Normal				Abnormal			
		Fan	Pump	Slider	Valve	Fan	Pump	Slider	Valve
Environment 1/2/3	–6/0/6	400	400	400	400	400	143	356	119

**Table 3** Details of various environments in ToyADAMOS dataset

Environment	Toy car		Toy conveyer		Toy train	
	Normal	Anomalous	Normal	Anomalous	Normal	Anomalous
Case 1/2/3	12,375	12,375	11,250	11,250	12,375	12,375

different types and scales of toy trains. Some sample signals from the MFPT, MIMII, and ToyADAMOS datasets are shown in Fig. 2.

## 2.2 Hilbert transform 2D grayscale image generations

In this study, the Hilbert transform is used to generate 2D images from the original bearing fault signals. It is an effective method for performing spectrum analysis on time-domain signals. It operates on real-time time-domain signals without the need to perform transformations into the space or frequency domains, unlike the Fourier and wavelet transforms. Since the Hilbert transform is a complex operator, performing the Hilbert transform on the time-domain signal  $y(t)$  produces an analytical signal  $\bar{y}(t)$ , which has a real and imaginary part.

The Hilbert transform of a signal  $y(t)$  can be written as

$$y(t) \xrightarrow{\text{HT}} \bar{y}(t) = \bar{y}_{\text{Re}}(t) + j \cdot \bar{y}_{\text{Im}}(t) \quad (1)$$

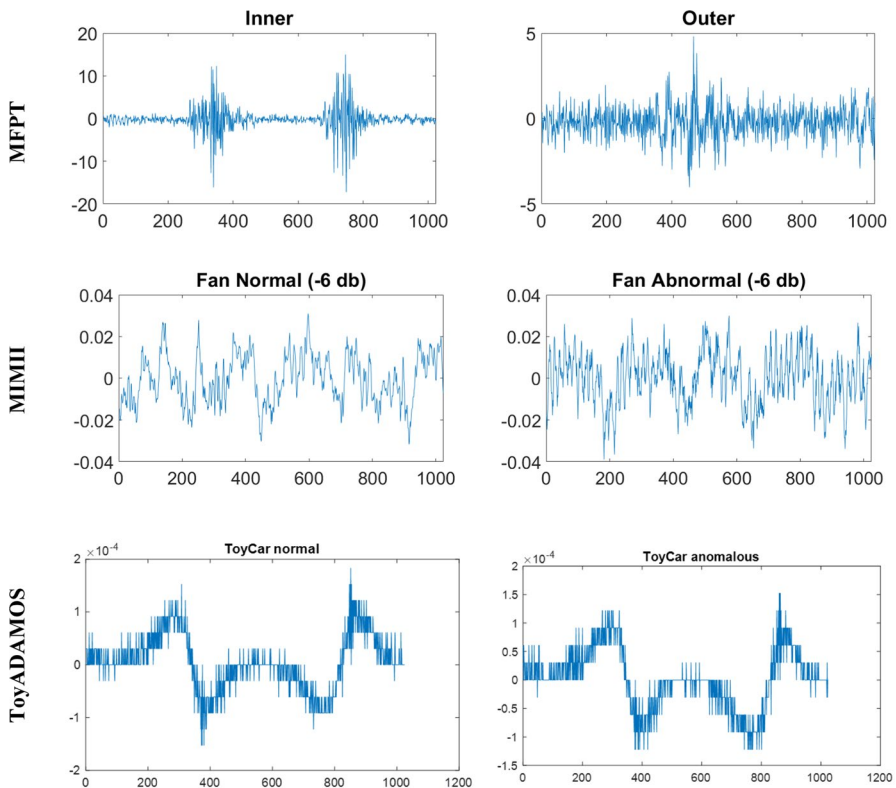


Fig. 2 Sample signals from the MFPT, MIMII, and ToyADAMOS datasets

where  $\bar{y}(t)$  is an analytic signal and  $\bar{y}_{\text{Im}}(t)$  represents the Hilbert transform of the signal  $\bar{y}_{\text{Re}}(t)$ . The amplitude  $A(t)$  of the time signal  $y(t)$  is given by

$$A(t) = \sqrt{\bar{y}_{\text{Re}}^2(t) + \bar{y}_{\text{Im}}^2(t)}. \quad (2)$$

This technique was used to extract the detailed phase shift between the real and imaginary components in an earlier study [25]. The spectrum amplitude was used to detect the existence of faults in electric machines. Since the spectrum amplitude of the analytic signal is obtained from its real and imaginary components, the amplitude of the analytical signal is used to extract the uniform texture pattern of the fault signals in this study.

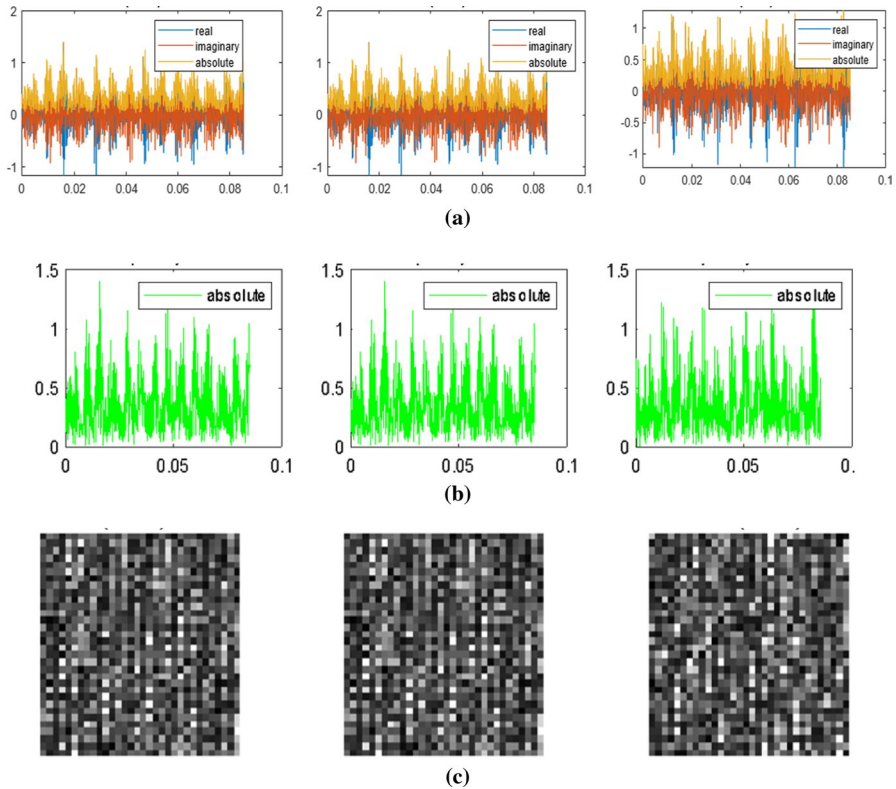
The generation of the 2D grayscale images from the original fault signals by applying the Hilbert transform is briefly described here. Since the analytical signal is complex, three different types of signals comprising the real and imaginary parts and the absolute values may be extracted from a time-series signal, as shown in Fig. 3a.

Among the three parts of the analytical signal, we consider only the conversion of its amplitude (i.e., absolute value, as depicted in Fig. 3b) to a 2D grayscale image following the approach in [8]. The 1D signal with a length of 1024 was subdivided into blocks with a length of 32 to generate a  $32 \times 32$  2D grayscale image. Sample texture images generated from the analytical inner and outer fault signals are presented in Fig. 4. The images demonstrate a uniform texture pattern for each type of signal. A 2D deep CNN–LSTM model was utilized to classify the faults represented as uniform 2D grayscale images, as discussed in the following subsection.

### 2.3 2D Deep CNN–LSTM hybrid architecture for fault recognition

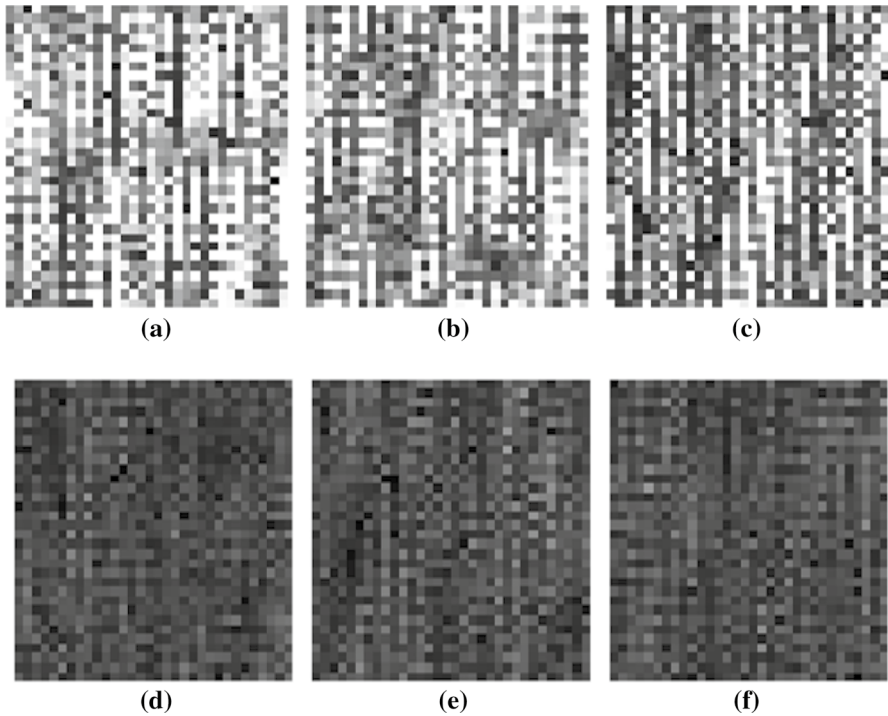
The architecture was constructed using the Keras sequential API. The signals were reshaped for image generation using the NumPy library. These images represent different signals belonging to the various classes and the image size was input into the Keras model. The input data were connected with the first layer of the entire neuron list, and the entire pixel list of all the images was forwarded to the first layer. The input shape was initially declared to avoid a merger between several images. To reduce the computation time, the input shape was set to  $32 \times 32$ . Since processing each image using a separate convolution block leads to several problems such as a long training time, the extraction of different features from additional images, and frame-to-frame changes in the characteristics of the time-series data, a series of time-distributed layers were utilized so that similar transformations could be performed across the list of the input images. Although the input shape information comprising the height, weight, and number of channels is usually used in the convolutional layer, the number of images that are inserted in each turn should be specified for the time distribution conv2D. Therefore, the input shape was set to (1, 32, 32, 1), where the first 1 indicates that one image goes through the subsequent layer in each turn.





**Fig. 3** Texture image generation from sample vibration fault signals. **a** Real and imaginary parts and absolute values of Hilbert transform signals for three sample fault signals, **b** amplitude of signal components in (a), and **c** texture patterns generated from signals in (b)

The number of channels was set to one because the input images were grayscale ones. The input shape did not need to be further declared for the subsequent layers because the Keras library could guess the perfect shape for connecting with other layers. The kernel size was set to  $5 \times 5$  at the beginning and to  $3 \times 3$  subsequently. The number of filters in the first layer was set to 32 because the lower layers extract features from the smaller parts of the images and increased to 64 in the subsequent layers to detect the high-level features. The ReLu activation function was added to every time-distributed conv2D layer to introduce nonlinearity to the system. A  $2 \times 2$  pooling layer (MaxPool2D) was added after each of the conv2D layers to down-samples the filter size and select the largest value from the two neighboring pixels. A dropout layer with a dropout rate of 25% was added after each max pooling layer to drop a quarter of the neurons randomly to avoid overfitting. The DCNN contains three more conv2D layers with a filter size of 64. Further, the current output was flattened and then transformed into a 1D vector for input to the LSTM. After the time-distributed layers, the images were processed frame by frame in a chronological manner by using two LSTM layers with



**Fig. 4** Sample reconstructed  $32 \times 32$  texture images: **a** inner and **b** outer race signals

64 units or LSTM cells. The sigmoid activation function was used in the LSTM to form a smooth curve varying from 0 to 1. The return sequences were set to true in response to the output of every node to avoid generating only a single output at the final node. Two dense layers associated with 64 neurons and the ReLu activation function were added followed by a 30% dropout layer. The output size of the dense layer varied depending on the dataset. For example, since the MIMII dataset has eight classes, the sigmoid activation function was used to derive the probability of the eight neurons corresponding to these particular eight classes. A similar approach was applied for the MPFT and ToyADAMOS datasets, for which the dense layer output size was set to 2 and 6, respectively. The sigmoid activation function was used in the dense layer and the RMS prop was used as an optimizer with a learning rate of 0.0001. The categorical cross-entropy was used as a loss function to detect the classes in the model. In the experimental evaluation, all the simulations for the three datasets were run for 100 epochs. The model architecture and the detailed layer information are presented in Fig. 5.

To implement TL in the proposed model, the weights of the training dataset were saved and utilized later for testing the fault dataset. It is expected that the domain-specific pre-trained weights can have a significant positive impact on the accuracy of the test dataset.

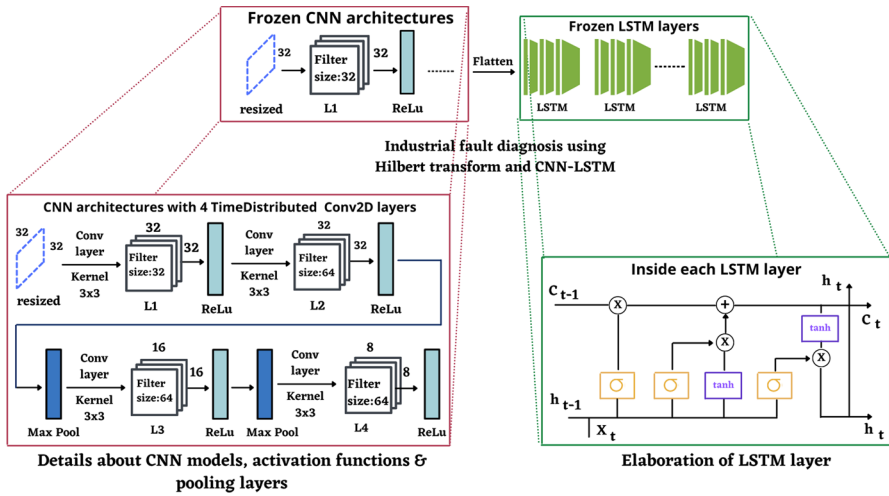


Fig. 5 Architecture of proposed hybrid model for detecting fault signals starting with image resizing followed by four time-distributed conv2D layers and LSTM layers with 64 units

### 3 Experimental results analysis

For the experimental evaluation, we used the *F1* score to numerically analyze the performance of the proposed fault diagnosis model. The *F1* score is derived from the precision (how consistent the results are over repeated measurements) and recall statistical parameters, as shown in Eqs. 3–5:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{3}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{Original Positive Number of Data}} \tag{4}$$

$$F1 \text{ Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

For consistency, the same hyperparameters such as the learning rate, batch size, and epochs were used in all the experiments to evaluate the performance of each model in a similar environment.

#### 3.1 Performance evaluation of proposed model

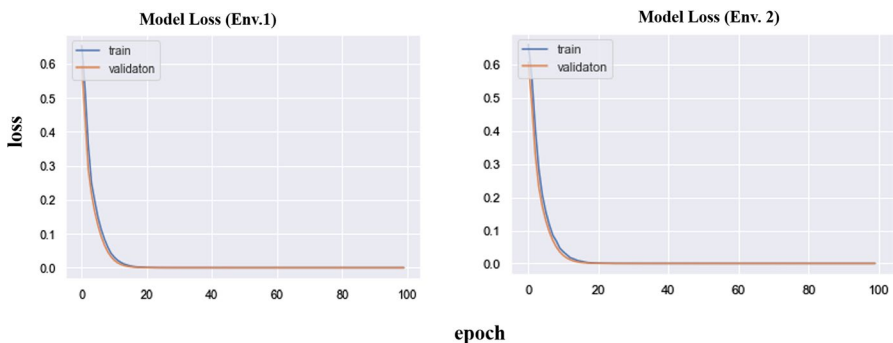
In the CNN–LSTM architecture, the front-end CNN layers and the LSTM layers (the details of the layers were presented in Sect. 2.3) function as feature extractors. In a previous study [11], the LSTM model was used as a stand-alone model

for inputting the raw data signals, and the model was trained with only sequential information. In the proposed architecture, the temporal depth and extracted features from the Hilbert transform 2D images are extracted by the deep CNN layers and fed to the LSTM layers. The results for Experiments 1, 2, and 3 performed using the MFPT, MIMII, and ToyADAMOS datasets, respectively, are discussed below.

### 3.1.1 Experiment 1: MFPT dataset

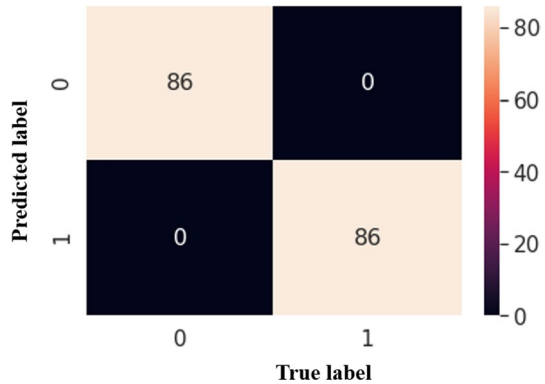
The MFPT dataset is smaller than the sub-datasets generated from the MIMII and ToyADAMOS datasets. However, this dataset has more variance because the signal loads varied from 50 to 300 lbs. Nonetheless, the hybrid architecture generalized better compared to conventional architectures and performed impressively when it was combined with the Hilbert transform 2D images. By utilizing the added information, it achieved an  $F1$  score of 1 at only 20 epochs for both environments (Environment 1 for 50–150 lbs and Environment 2 for 200–300 lbs) and maintained the score throughout the remainder of the training. This indicates that the model was able to learn very quickly without overfitting, which would have resulted in spikes in the training and validation curves at every epoch. Figure 6 shows that the training and validation loss curves converged smoothly after less than 20 epochs and maintained a net-zero loss. All the batches achieved very good results and maintained zero validation loss in both environments.

There are only two classes comprising inner and outer race faults in the two environments. 20% of the data was used for testing in the experimental evaluation. For both environments, the proposed model successfully classified both classes accurately (as shown in Fig. 7) based on the amplitudes of the analytical signal images. The good performance of the proposed model across a varied range of loads reflects its broad applicability.



**Fig. 6** Training and validation loss of the proposed model for Environment 1 (50–150 lbs) and Environment 2 (200–300 lbs) of MFPT dataset

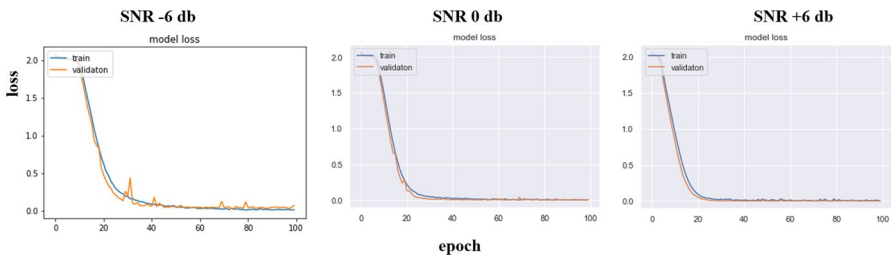
**Fig. 7** Confusion matrices of proposed model for different environments in the MFPT dataset



### 3.1.2 Experiment 2: MIMII dataset

Compare to the vibration dataset used in Experiment 1, the MIMII dataset is more complex. Various actual industrial noisy environments were captured in the collected audio signals. The MIMII dataset contains a total of 7854 sample signals across three environments with noise levels of  $-6$  dB,  $0$  dB, and  $+6$  dB. There are eight classes of data, which are denoted as 0 for fan (normal), 1 for fan (abnormal), 2 for pump (normal), 3 for pump (abnormal), 4 for slider (normal), 5 for slider (abnormal), 6 for valve (normal), and 7 for valve (abnormal) in the confusion matrix. The dataset was split in an 80:20 ratio for training and validation and run for 100 epochs for all cases and the losses are shown in Fig. 8. Figure 8 shows that for the  $-6$  dB noisy case, the hybrid model took a longer time to converge compared to the  $0$  dB and  $+6$  dB noisy cases. However, after the sixtieth epoch, it stabilized and reached its optimal state with a maximum  $F1$  accuracy of 99.6% on the test dataset. The signals with  $+6$  dB converged faster with less performance variance.

Similar to Experiment 1, the proposed model correctly detected the eight types of faults in the three environments except for only a very few signals. The model accurately detected the normal and abnormal of fan, pump, and slider audio signals in most cases except only one signal in  $0$  dB and  $+6$  dB. However, four valve normal signals in the  $-6$  dB case were not detected successfully. The detailed confusion



**Fig. 8** Training and validation loss curves of proposed model for different noisy environments in MIMII dataset

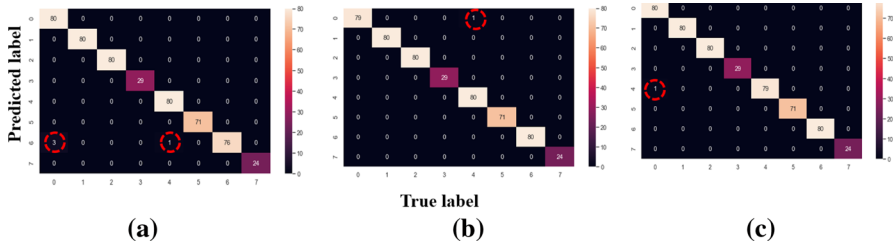


Fig. 9 Confusion matrices of proposed model for MIMII dataset for SNRs of a–6 dB, b 0 dB, and c+6 dB

matrices of the proposed model for the MIMII audio machine dataset are shown in Fig. 9.

### 3.1.3 Experiment 3: ToyADAMOS dataset

In this experiment, a complex and larger dataset, i.e., ToyADAMOS, was used to evaluate the performance of the proposed model. There are six classes of data comprising the ToyCar\_normal, ToyCar\_anomalous, ToyConveyer\_normal, ToyConveyer\_anomalous, ToyTrain\_normal, and ToyTrain\_anomalous classes and a total of 72,000 samples in the three cases in the dataset. These classes are, respectively, denoted as 1–6 in the confusion matrices in Fig. 10. Similar to Experiments 1 and 2, the dataset was split in an 80:20 ratio for training and testing. The simulations were run for 100 epochs for all the cases. The model accurately detected all kinds of fault signals in all the three cases, as shown in Fig. 10. In all the cases, the model took a few epochs to converge similarly to the other datasets presented in Experiments

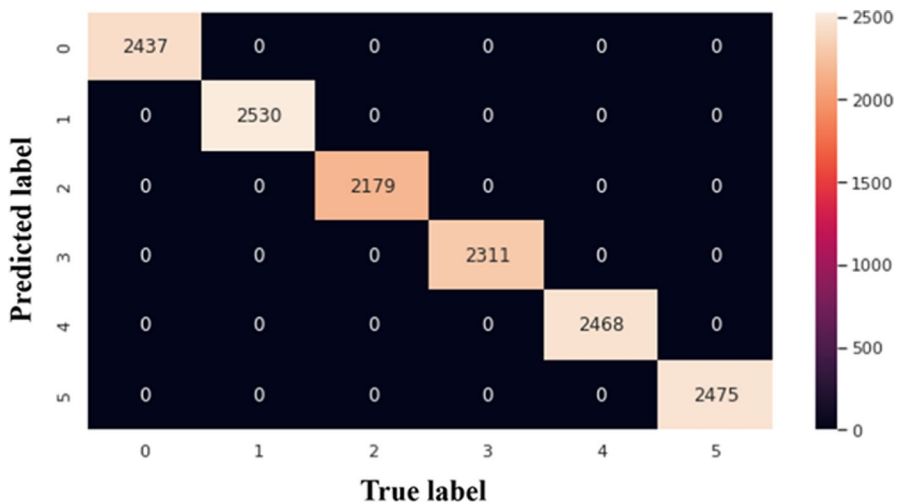
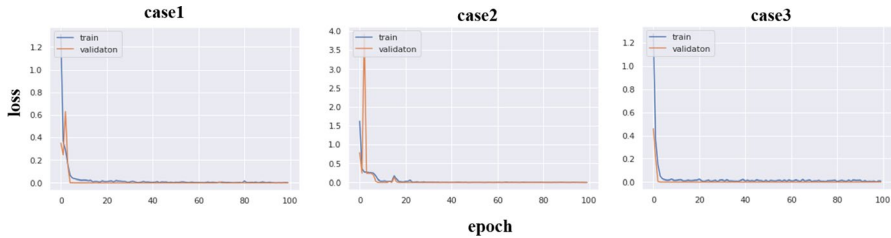


Fig. 10 Confusion matrices of proposed model for different environments in ToyADAMOS dataset



**Fig. 11** Training and validation loss curves of proposed model dataset for different environment cases in ToyADAMOS dataset

1 and 2. As shown in Fig. 11, the model stabilized after the twentieth epoch and reached its optimal state on the test dataset.

### 3.2 Comparison with other state-of-the-art models

A detailed comparison of the proposed hybrid architecture with three state-of-art transformed-signal techniques based on conventional CNN and LSTM DL models using the MFPT, MIMII, and ToyADAMOS datasets is presented in this section.

Table 4 shows a performance comparison of the Hilbert transform-based texture extraction method with methods based on DWT, FFT, and the gammatone spectrogram on the MIMII dataset. The FFT-based method exhibited the worst performance, while the Hilbert transform-based method outperformed the other models. The overall performance of DWT depends on the selected kernel function and has poor directionality is shift-invariant and does not contain phase information. We utilized the db4 kernel function for the evaluation. In contrast, similar to the Hilbert transform, the transformed signal after FFT also contains magnitude and phase information. However, the domain conversion results in high latency as the data is not processed in the same order as the input data. The application of several filter banks in the gammatone spectrogram leads to higher computational complexity than the other transform techniques. For all three environments in the MIMII dataset, the Hilbert transform-based texture feature extraction achieved better  $F1$  scores than the average  $F1$  scores of 0.94, 0.88, 0.90, and 0.99 for the DWT, FFT, and gammatone spectrogram-based transform methods, respectively.

The hybrid model has fewer higher trainable parameters than the stand-alone DCNN, LSTM, and state-of-art TL architectures. More specifically, the DCNN, LSTMs, and DCNN–LSTM architectures have 328,102, 62,406, and 1,168,294 trainable parameters, respectively, while the VGG16 architecture has 14,779,974 parameters, which is more than 10 times that of the hybrid architecture.

The experimental results in Table 5 show that all the architectures successfully detected the fault signals in the MFPT dataset under various load conditions. Although all the models detected the fault signals accurately, the train and validation loss curves of the proposed architecture were smoother and converged earlier than those of the other state-of-art models. In contrast, for the MIMII and ToyADAMOS datasets, the deep LSTM model exhibited lower training accuracy than its validation

**Table 4** *F1* scores of DWT, FFT, and Hilbert transform analytical signal-based texture extraction with deep CNN–LSTM model for MIMII dataset

Environment	DWT with hybrid CNN–LSTM	FFT with hybrid CNN–LSTM	Gammatone spectrogram with hybrid CNN–LSTM	Hilbert transform (absolute part) analytical signal with hybrid CNN–LSTM
Environment 1 (–6 dB)	0.920	0.820	0.918	0.996
Environment 2 (0 dB)	0.999	0.91	0.812	1.00
Environment 3 (+6 dB)	0.89	0.92	0.994	1.00
<b>Average</b>	0.936	0.883	0.908	0.999

and testing accuracies because it could not extract deeper features as a stand-alone model. The LSTM model failed to learn and predict several cases that the DCNN and proposed architecture managed to successfully. The proposed model outperformed the other models under the different environments of the three datasets by achieving higher accuracy and smoother training and validation curves than those of the state-of-art models.

The results demonstrate that the proposed model is efficient not only for a particular dataset but also for the different environments in the three datasets—the MFPT dataset covers loads with large variances, the MIMII dataset contains both negative-scale and positive-scale noisy cases, and ToyADAMOS contains data from different machines with various specifications. This demonstrates an important aspect of the model performance under different conditions that may be present in real-life situations. It can thus be concluded that the proposed model classifies faults efficiently and makes accurate predictions in the given environments with varying data complexities.

### 3.3 Implementation of transfer learning in proposed model

The environment for rotatory machine fault detection can vary because of environmental variations and the physical characteristics of the machines. This study is therefore limited to environment-specific conditions. To reduce the gap between the different environments, TL was implemented by interconnecting the various environments.

To reduce the training time, several researchers have recently used ImageNet pre-trained weights to test for fault signals, as discussed in Sect. 1.1. In this study, the trained weights obtained using a source fault dataset were saved and used for training/testing the target datasets with a completely different set of conditions. For example, we trained the model using case 1 of the ToyADAMOS dataset. The model took more than 50 epochs to converge. The weights were then saved, the model was retrained using case 2, and the previous weights were updated according to the new environment samples. The model took only approximately



**Table 5** Accuracies of deep CNN, deep LSTM, and proposed model for MFPT, MIMII, and ToyADAMOS datasets

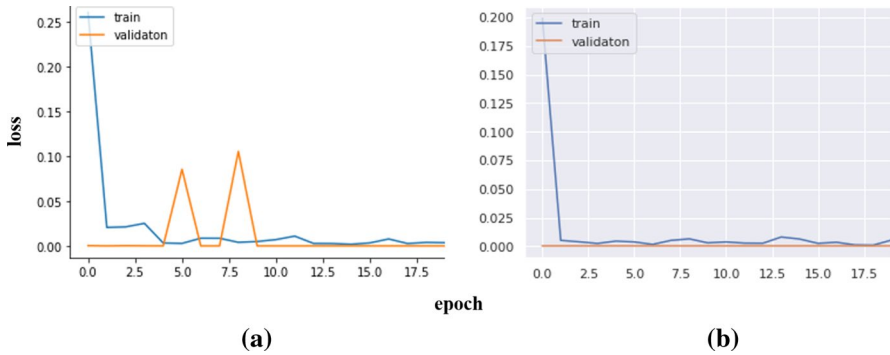
Dataset	Environments	Deep CNN	Deep LSTM	Hybrid DCNN-LSTM
MFPT	Load 50–150	1.00	1.00	1.00
	Load 200–300	1.00	1.00	1.00
	<b>Average</b>	1.00	1.00	1.00
MIMII	– 6 dB	0.98	0.99	0.996
	0 dB	0.99	0.998	1.00
	+ 6 dB	1.00	0.97	1.00
	<b>Average</b>	0.99	0.986	0.998
ToyADAMOS	Case1	0.979	0.991	0.997
	Case2	0.972	0.994	0.998
	Case3	0.995	0.982	1.00
	<b>Average</b>	0.982	0.989	0.998

10 epochs to converge this time because it has already learned features from the previous training. The model required even fewer epochs to converge in case 3, as shown in Fig. 12. Similarly, for MFPT, we trained the model using the data for 50–150 lbs loads. Although we ran the model for 100 epochs, the best model was obtained after the twentieth epoch. The best weights were then utilized for training using the data for 200–300 lbs loads. This time, the model took only seven epochs to converge with the trained weights. Since the MFPT dataset has less variance and more differentiable features between its two classes, it converged more quickly compared to the other two datasets.

The same approach was applied for the MIMII dataset, as presented in Table 6. The model was pre-trained with – 6 dB noisy data and subsequently trained with the two remaining sets of environmental data (0 dB and + 6 dB noisy). The experimental results show that as the noise increased, the model took less time to converge and better models were achieved. After implementing TL on the proposed model, it took less than 20 epochs of the faulty signals in all datasets to be classified accurately instead of the 100 epochs required without implementing TL. Since TL-based hybrid DL significantly reduces the training time, it is highly suitable for real-time industrial fault diagnosis in various environments.

## 4 Conclusion

This paper presented an industrial fault diagnosis model with Hilbert transform and a 2D deep CNN–LSTM architecture. The model was used to classify faults in different environments with loads ranging from 0 to 300 lbs and noises from – 6 dB to + 6 dB. Two environments were included in the MFPT dataset and three environments each



**Fig. 12** Training and validation loss curves for proposed model under different environments in the ToyADAMOS dataset: **a** case 1 as source task and case 2 as target task; **b** case 1 as source task and case 3 as target task

**Table 6** Computational complexity with several epochs of the proposed model with and without TL

	Proposed hybrid (HT, Deep CNN–LSTM) (Epochs)	Proposed hybrid model with transfer learning	
		Epochs	<i>F1</i> score
MFPT dataset (training with 50–150 lbs load)	20 ( <i>F1</i> score 1.00)	7	For 200–300 lbs load, 1.00
MIMII dataset training with –6 dB SNR dataset	100 ( <i>F1</i> score 0.996)	20	For 0 dB SNR, 1.00
		20	For +6 dB SNR, 1.00
ToyADAMOS dataset training with –case1 dataset	100 ( <i>F1</i> score 0.997), case1	10	For case 2, 0.998
		7	For case3, 1.000

in the MIMII and ToyADAMOS datasets. The Hilbert transform analytical signal-based texture extraction method was compared with the state-of-the-art DWT, FFT, and gammatone spectrogram-based methods. The Hilbert transform-based 2D image generation outperformed the state-of-the-art transform methods because it extracted more efficient features. The *F1* score was used as a performance metric to evaluate the performance of the proposed and state-of-the-art models. The state-of-the-art models did not perform consistently well as the motor load and noise increased. In contrast, the proposed model exhibited consistent performance in all environments with varying loads, RPMs, and noise levels. The proposed model also had 12 times less trainable parameters than state-of-art TL models. Implementing TL with domain-specific fault datasets reduced the average training time over five times. This reduced the time required compared with training for every machine-specific environment from scratch. It is therefore expected that the proposed model can play a significant role in real-time industrial fault diagnosis in environments with various loads, RPMs, and noise levels.

The performance of the proposed hybrid DTL model was evaluated under various environments in a single machine. In future work, incremental learning techniques can be evaluated with more complex fault datasets from different machines.

**Acknowledgment** This research was supported and funded by the Korean National Police Agency. [Project Name: XR Counter-Terrorism Education and Training Test Bed Establishment/Project Number: PR08-04-000-21].

**Funding** The funding was provided by the Korean National Police Agency.

**Data availability** Three publicly available datasets (MFPT [22], MIMII [23], and ToyADAMOS [24]) are used to validate the models. Sample datasets are available in the following GitHub link: <https://github.com/MaheZ20Kaist/Fault-Datasets/blob/main/Dataset.txt>

## Declarations

**Conflict of interest** The authors declare that no conflicts of interest are associated with this publication.

**Consent for publication** All authors have agreed and given their consent for submission of this paper to the Journal of Supercomputing.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Park YJ, Fan SKS, Hsu CY (2020) A review on fault detection and process diagnostics in industrial processes. *Processes* 8(9):1123. <https://doi.org/10.3390/pr8091123>
2. Islam MR, Jia U, Jong MK (2016) acoustic emission sensor network based fault diagnosis of induction motors using a gabor filter and multiclass support vector machines. *Adhoc Sens Wirel Netw* 34(1):273–287
3. Wang Y, Orchard J (2009) Fast discrete orthonormal Stockwell transform. *SIAM J Sci Comput* 31(5):4000–4012
4. Grover C, Turk N (2022) A novel fault diagnostic system for rolling element bearings using deep transfer learning on bispectrum contour maps. *Eng Sci Technol* 31:1–12. <https://doi.org/10.1016/j.jestech.2021.08.006>
5. Hasan MJ, Manjural MM, Jong-Myon K (2019) Acoustic spectral imaging and transfer learning for reliable bearing fault diagnosis under variable speed conditions. *Measurement* 138:620–631
6. Kaplan K, Kaya Y, Kuncan M, Minaz MR, Ertunc M (2020) An improved feature extraction method using texture analysis with LBP for bearing fault diagnosis. *Appl Soft Comput* 87:106019
7. Hasan MJ, Islam MM, Kim JM (2021) Multi-sensor fusion-based time-frequency imaging and transfer learning for spherical tank crack diagnosis under variable pressure conditions. *Measurement* 168:108478
8. Islam R, Jia U, Kim JM (2018) Texture analysis-based feature extraction using Gabor filter and SVD for reliable fault diagnosis of an induction motor. *Int J Inf Technol Manag* 17(1–2):20–32
9. Xia M, Li T, Xu L (2018) Fault diagnosis for rotating machinery using multiple sensor and convolutional neural networks. *IEEE/ASME Trans Mechatron* 23(1):101–110

10. Zhao G, Zhang G, Ge Q, Liu X (2016) Research advances in fault diagnosis and prognostic based on deep learning. In: Prognostics and system health management conference, China, pp 1–6. <https://doi.org/10.1109/PHM.2016.7819786>
11. Lei J, Chao L, Dongxiang J (2019) Fault diagnosis of wind turbine based on long short-term memory networks. *Renew Energy* 133:422–432
12. Peng DD, Liu ZL, Wang H (2019) A novel deeper one-dimensional CNN with residual learning for fault diagnosis of wheelset bearings in high-speed trains. *IEEE Access* 7:10278–10293
13. Wen L, Li X, Gao L (2019) A transfer convolutional neural network for fault diagnosis based on ResNet-50. *Neural Comput Appl* 32:6111–6124. <https://doi.org/10.1007/s00521-019-04097-w>
14. Wen L, Li X, Li X, Gao L (2019) A new transfer learning based on VGG-19 network for fault diagnosis. In: IEEE 23rd international conference on computer supported cooperative work in design, Portugal, pp 205–209. <https://doi.org/10.1109/CSCWD.2019.8791884>
15. Chen W, Qiu Y, Feng Y, Li Y, Kusiak A (2021) Diagnosis of wind turbine faults with transfer learning algorithms. *Renew Energy* 163:2053–2067
16. Xu W, Wan Y, Zuo TY, Sha XM (2020) Transfer learning based data feature transfer for fault diagnosis. *IEEE Access* 8:76120–76129
17. Ping M, Zhang HL, Fan WH, Wang C (2020) A diagnosis framework based on domain adaptation for bearing fault diagnosis across diverse domains. *ISA Trans* 141:553–559
18. Wen L, Gao L, Li X (2019) A new deep transfer learning based on sparse auto-encoder for fault diagnosis. *IEEE Trans Syst Man Cybern Syst* 49(1):136–144
19. Shao S, McAleer S, Yan R, Baldi P (2019) Highly accurate machine fault diagnosis using deep transfer learning. *IEEE Trans Ind Inform* 15(4):2446–2455. <https://doi.org/10.1109/TII.2018.2864759>
20. Fan H, Xue C, Zhang X, Cao X, Gao S, Shao S (2021) Vibration images driven fault diagnosis based on CNN and transfer learning of rolling bearing under strong noise. *Shock Vib*. <https://doi.org/10.1155/2021/6616592>
21. Hasan MJ, Manjural I, Kim JM (2021) Multi-sensor fusion-based time-frequency imaging and transfer learning for spherical tank crack diagnosis under variable pressure conditions. *Measurement* 168:108478. <https://doi.org/10.1016/j.measurement.2020.108478>
22. Bechhoefer E (2013) Condition based maintenance fault database for testing diagnostics and prognostic algorithms. MFPT Data. <https://www.mfpt.org/fault-data-sets/>
23. Purohit H, Tanabe R, Ichige K, Endo T, Nikaido Y, Suefusa K, Kawaguchi Y (2019) MIMII Dataset: sound dataset for malfunctioning industrial machine investigation and inspection [arXiv:1909.09347](https://arxiv.org/abs/1909.09347)
24. Koizumi Y, Saito S, Uematsu H, Harada N, Imoto K (2019) ToyADMOS: a dataset of miniature-machine operating sounds for anomalous sound detection. In: 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, pp 313–317. [arXiv:1908.03299](https://arxiv.org/abs/1908.03299)
25. Medoued A, Lebaroud A, Sayad D (2013) Application of Hilbert transform to fault detection in electric machines. *Adv Differ Equ*. <https://doi.org/10.1186/1687-1847-2013-2>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Authors and Affiliations

Mahe Zabin<sup>1</sup> · Ho-Jin Choi<sup>1</sup>  · Jia Uddin<sup>2</sup>

Mahe Zabin  
mahezabin@kaist.ac.kr

Jia Uddin  
jia.uddin@wsu.ac.kr

<sup>1</sup> School of Computing, Korea Advanced Institute of Science and Technology, KAIST, Daejeon, South Korea

<sup>2</sup> AI and Big Data Department, Endicott College, Woosong University, Daejeon, South Korea