IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 7, 2020, accepted October 28, 2020, date of publication November 9, 2020, date of current version November 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3036769

Drill Fault Diagnosis Based on the Scalogram and Mel Spectrogram of Sound Signals Using Artificial Intelligence

THANH TRAN AND JAN LUNDGREN
Department of Electronics Design, Mid Sweden University, 851 70 Sundsvall, Sweden
Corresponding author: Thanh Tran (thanh.tran@miun.se)

This research was funded by EU Regional Fund and the MiLo Project (No. 20201888).

ABSTRACT In industry, the ability to detect damage or abnormal functioning in machinery is very important. However, manual detection of machine fault sound is economically inefficient and labor-intensive. Hence, automatic machine fault detection (MFD) plays an important role in reducing operating and personnel costs compared to manual machine fault detection. This research aims to develop a drill fault detection system using state-of-the-art artificial intelligence techniques. Many researchers have applied the traditional approach design for an MFD system, including handcrafted feature extraction of the raw sound signal, feature selection, and conventional classification. However, drill sound fault detection based on conventional machine learning methods using the raw sound signal in the time domain faces a number of challenges. For example, it can be difficult to extract and select good features to input in a classifier, and the accuracy of fault detection may not be sufficient to meet industrial requirements. Hence, we propose a method that uses deep learning architecture to extract rich features from the image representation of sound signals combined with machine learning classifiers to classify drill fault sounds of drilling machines. The proposed methods are trained and evaluated using the real sound dataset provided by the factory. The experiment results show a good classification accuracy of 80.25 percent when using Mel spectrogram and scalogram images. The results promise significant potential for using in the fault diagnosis support system based on the sounds of drilling machines.

INDEX TERMS Deep learning, machine fault diagnosis, machine learning, sound signal processing.

I. INTRODUCTION

A drilling machine is a kind of rotating cutting machine that is used widely in factories to drill holes in materials such as metal, wood, and plastic. The timely detection of problematic drill bits is essential to preventing damage to materials due to drilling faults. Because the cutting of a drill makes sounds, skilled technicians can distinguish the sound of a drill that is not working properly and immediately stop production so that the drill can be repaired. Recent advances in automation technology in factories have promoted the creation of an automated support system that can classify drill sounds and provide an alert when sound indicates that a drill is broken. The benefit of a drill fault diagnosis system is that it reduces production costs and manpower.

The associate editor coordinating the review of this manuscript and approving it for publication was Her-Terng Yau.

In recent years, many researchers have investigated different ways of classifying the sound signals emitted from machines. Delgado-Arredondo *et al.* [1] analyzed sound and vibration signals to detect fault induction motors. The authors used the complete ensemble empirical mode decomposition (CEEMD) to isolate important information and eliminate noise signals. The authors then calculated the frequency marginal of the Gabor representation to obtain the intrinsic mode functions (IMF) in the frequency domain. Nevertheless, it is necessary to know the motor speed or the slip to locate the fault frequency. This makes it difficult to apply widely in industry because it is not always possible to determine the specific motor speed.

Some approaches extracted statistical features from sound signals and classified them using conventional machine learning methods. For detection and diagnosis of mechanical faults in ball bearings, Kankar *et al.* [2] extracted

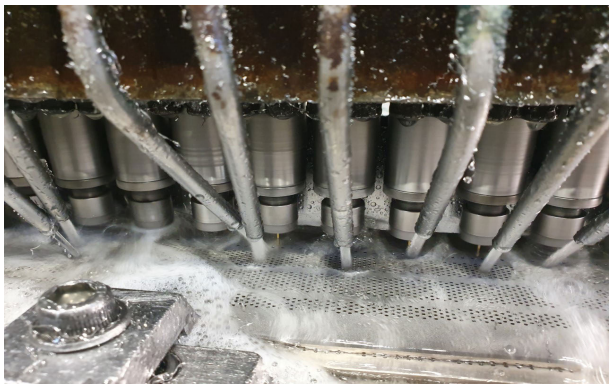
six statistical features (range, mean value, standard deviation, skewness, kurtosis, and crest factor) to train the support vector machine (SVM) and artificial neural network (ANN) classifiers. The accuracy of this research was only 71.23 percent and 73.97 percent when using ANN and SVM, respectively. Kumar *et al.* [3] developed a system for automatic drilling operations using vibration signals. The authors used low pass Butterworth filter to preprocess vibration signals before extracting eight features from the time domain, eight features from the frequency domain, and five Morlet wavelet features. These extracted features were normalized using zero one normalization and applied singular value decomposition to remove the redundant and irrelevant features. Then Support Vector Machine (SVM), Artificial Neural Network (ANN), and Bayes classifier were used for drill fault detection and recognition. Lee *et al.* [4] extracted Mel-frequency cepstrum coefficients (MFCCs) from the audio signals and also employed SVM for classification. The accuracy reached 94.1 percent on their dataset, which is collected from an NS-AM-type railway point machine at Sehwa Company in Daejeon, South Korea. The length of each sound on their dataset was around 5000 ms. However, their method did not show a promising result when applied in our drill sound dataset because each sound recording is extremely short. Kemalkar and Bairagi [5] extracted MFCCs features and made a comparison between these features and a library of features to decide on the fault or non-fault state of a bike engine. Zhang [6] used the principal component analysis (PCA) algorithm to extract and train the training samples. Then, the author used self-organizing maps (SOM) to cluster the principal component by neural network clustering into four categories and the Bayesian discriminant method to identify the testing samples. The dataset for his experiment was collected by a self-developed drilling test rig using a signal acquisition hardware system (sensors, data acquisition cards, and industrial computers).

Ince *et al.* [7] proposed using a time-domain signal as the input of a small 1-D CNN to classify motors as either healthy or faulty. The authors used a balanced dataset of 260 healthy and 260 faulty cases for training a 1-D CNN model. Luo *et al.* [8] detected the fault stage of CNC machine tools based on their vibration signals. The authors used 10 000 samples; 9000 of these were used for training and 1000 were used for testing a deep auto-encoder (DAE) model. The DAE model, which is combined between the SAE layer and the BPNN layer, is used to classify impulse and non-impulse responses. A dynamics identification algorithm was then used to identify dynamic properties from impulse responses. Finally, similarities between the dynamic properties were used to detect the health of the CNC machine tool. Similarly, Long *et al.* [9] combined a sparse autoencoder (SAE) and an echo state network (ESN) to diagnose the transmission faults of delta 3-D printers. These authors collected attitude data from an attitude sensor and used SAE to extracted features from attitude data. Then these features were used as the input of an ESN for fault recognition. Long *et al.* [10] also combined a hybrid evolutionary algorithm featuring a competitive swarm optimizer and a local search to optimize parameter values and hyperparameter settings of echo state network for intelligent fault diagnosis. To test the performance of their proposed method, the authors conducted fault diagnosis experiments for a 3-D printer and a gearbox. Paul and Lofstrand [11] proposed an interval type-2 (IT2) Takagi-Sugeno (T-S) fuzzy-based observer fault detection scheme for drill bit fault detection and raising an alarm after 45 seconds.

Many recent studies gained remarkable results when using image representation of the sound signal to train state-of-the-art deep learning architectures such as convolutional neural networks (CNNs) on machine fault sound diagnosis [12]–[15]. However, the vast majority of the research was conducted using a large and balanced dataset. In reality, the sounds recorded when the drills were broken occupy only a small percentage of the whole dataset compared to the normal working sound of drill machines. The imbalanced real-world dataset leads to a bias for training CNN architectures; for example, the CNN model might predict poorly on the minority class since this class has fewer data. Besides, since each sound sample in our real-world dataset is too short, around 20.83 ms and 41.67 ms, the extracted features from raw sound signals cannot carry much important information for classification. This leads to more difficulties in drill sound classification compared to previous studies. To the best of our knowledge, no studies have been conducted using such short drill sounds. Therefore, the use of an end-to-end system (only a conventional machine learning or an advanced convolutional machine learning architecture) does not meet the desired accuracy of sound classification in the industry sector. To solve the problem of drill fault sound detection, an architecture consisting of many layers is necessary (preprocessing, image conversion, feature extraction, feature selection, classification). For each layer, different algorithms that include both a deep learning architecture and machine learning methods may be used. The limitation of the traditional approach, which includes several steps/layers, is that each step requires to be optimized separately under different criteria. However, instead of losing discriminative information, optimizing each step separately under different criteria can help to improve the performance of each step and the whole system.

In this article, we propose using image representations of the sound signals instead of using raw sound signals in the classification task. Two linear transformations from audio to images are utilized for classifying drill sounds: Mel spectrogram images and scalogram images. Section II.B describes these transformation methods in detail. Furthermore, we also propose a new method combining both modern deep learning methods and traditional machine learning classifiers for drill fault sound diagnosis. Firstly, effective features are extracted automatically using a pre-trained CNN architecture instead of hand-crafted features. Secondly, neighborhood component analysis [16] is implemented to select rich features that carry

(a) A drill machine when idle.



(b) A drill machine when active.

FIGURE 1. A drill machine at Valmet AB.

the most relevant information. Afterward, these selected features are used to train machine learning classifiers to compare the performance.

As the comparative studies, we compare the performance of literature studies including (1) extracting features from raw sound signals and classifying them using conventional machine learning classification methods, and (2) using Mel spectrogram or scalogram images to train CNN architectures. We conducted the experiment before and after utilizing the NCA algorithm to assess the effectiveness of the feature selection method for the classification result. The experiment results prove the significant efficiency of the feature selection algorithm in maximizing the prediction accuracy of classification algorithms.

The rest of this article is organized as follows. Section II describes the proposed methods step-by-step, including data collection, pre-processing data, converting sound signals into images, feature extraction, feature selection, and classification. The result, comparative study, and discussion are presented in section III. Finally, the conclusion is presented in section IV.

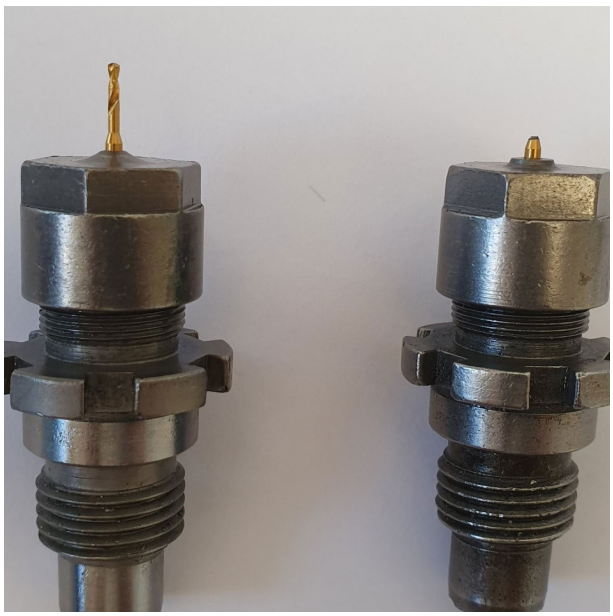


FIGURE 2. A healthy drill bit (on the left side) and a broken drill bit (on the right side).

II. PROPOSED METHOD

A. DATA COLLECTION

The dataset was recorded from a drill machine at Valmet AB, a company in Sundsvall, Sweden, using four AudioBox iTwo Studio microphones. The microphones have high-performance mic preamplifiers. Figure 1 shows a drill machine at Valmet AB when it is idle and when it is active. The sampling frequency used to record drill sounds is 96 kHz. The dataset includes two parts. One part has a total of 833 files. Each file is a vector containing 2000 samples corresponding to a time of 20.83 ms. Another part has a total of 40 417 files. Each file contains 4000 samples corresponding to a time of 41.67 ms.

Valmet AB is currently operating multiple drilling machines to drill thousands of small holes in metal plates, as shown in Figure 1. There are two types of drilling machines that are used in the factory, one comprises of 90 drill bits, the other comprises of 120 drill bits. A technician usually turns off the drill every 10 minutes to check its status. Any drill bits that are damaged must be replaced to ensure that all holes are drilled on the surface of the metal plate as intended. Stopping the production line is time-consuming and labor-intensive. Therefore, an automated system that detects broken drill bits is essential in production to reduce production and labor costs.

It is widespread to use vibration sensors for machine fault detection. However, the use of vibration sensors for the drill fault detection system at Valmet AB faces many challenges. Firstly, each drilling machine consists of 90 or 120 drill bits, as shown in Figure 1. Each drill bit needs a vibration sensor in order to detect any fracture. Mounting 90 to 120 sensors as one sensor array and classifying the vibration differences is complicated and costly. Secondly, when cutting metal, in particular, it is necessary to ensure a smooth and accurate

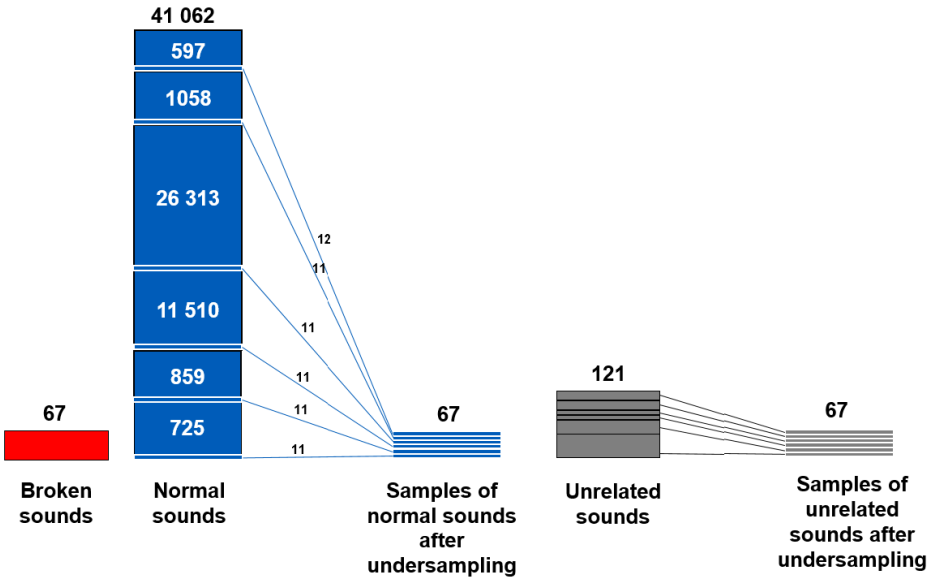


FIGURE 3. Undersampling the original dataset.

hole while preventing the metals from grabbing the drill bits. Therefore, water is used to clean the surface of the metal, as shown in Figure 1. Besides, water is also used to cool down the drill bits. It is difficult to use vibration sensors in a very wet environment. Finally, an artificial intelligence supported sound measurement system enables a more robust and accurate way to detect drill breaks in this type of machine. All of these reasons motivated us to use sound to detect the broken drill instead of using vibration.

Addressing an actual demand, this research aims to detect abnormal functioning of the drills based on the sound signals. These sounds are generated when a drilling machine is operating. Based on data collected from the factory, the dataset is divided into three groups: broken sounds, normal sounds, and unrelated sounds. The broken sounds group contains all of the sounds recorded when the drill was broken. They account for approximately 0.16 percent (67 sound signal files) of all files in the dataset (41 250 sound signal files in total). The normal sounds were recorded when the drilling machine was working properly. These sounds make up most of the dataset, up to 99.54 percent (41 062 sound signal files). Figure 2 shows a normal drill bit and a damaged drill bit. Unrelated sounds were caused by the surrounding environment, such as the sound made when knocking on the microphone or knocking in the scrap box, vacuum cleaner sounds, volume control sounds, unknown sounds, etc. There are 119 sound files in this group, and they make up around 0.29 percent of all the sound files.

The actual rate of broken drill bits in the production line is only around 0.16 percent compared to the total number of drill sounds obtained (67 broken-drill sound files out of 41 250 files in total). Thus, it is difficult to detect broken drills. Consequently, creating a balanced dataset is an important step to take before proceeding further. We selected the number of sounds in majority classes (normal sounds and unrelated sounds classes) equal to the number of sounds in the minority class (broken sounds class), as shown in Figure 3. The total number of sounds in the dataset after undersampling is 201.

The balanced dataset that has 201 samples are small to work with. Therefore, training an end-to-end learning model on a small number of samples tends to overfit and produce inaccurate results. Because the more parameters the complex model has, the more susceptible it is to overfit. Besides, some conventional classifiers are good at dealing with small datasets. To overcome overfitting with the small dataset, we decided to extract features from the image representations of sounds (Mel spectrogram and scalogram images) and choose the conventional classification models for the classification task. The detail of our proposed methods is presented in the next section.

B. METHOD

In this section, the pre-processing, Mel spectrogram, scalogram, feature extraction using transfer learning, and feature selection are presented. Figure 4 shows the detailed steps of the proposed method. All drill sounds are processed step by step following the training progression. Then, the new sound signal that needs to be classified goes through the same process in the prediction phase.

1) SOUND PRE-PROCESSING

A human can hear sounds ranging in frequency from 20 Hz to 20 000 Hz. Hence, the sample rate of 44 100 Hz corresponding to a maximum sound frequency of 22 050 Hz is usually used for recording sound. However, the sample rate that is

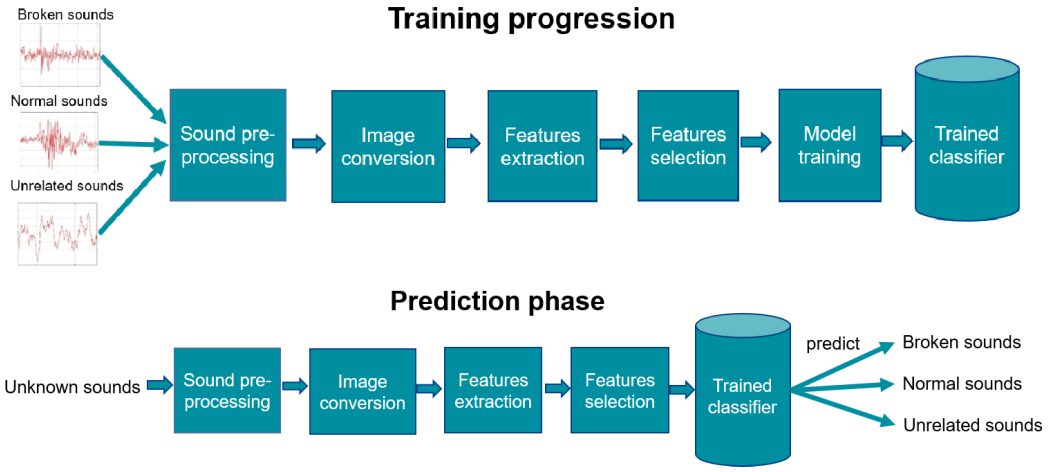


FIGURE 4. Process of representations for the proposed drill fault diagnosis method.

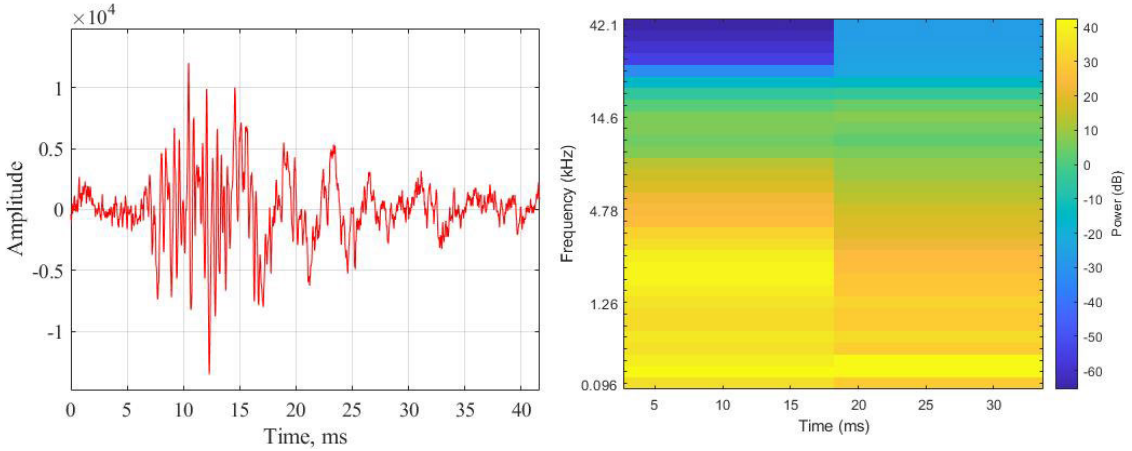


FIGURE 5. (a) The sound signal in the time domain, (b) Mel spectrogram of drill raw signal. Dark blue corresponds to low amplitudes. Brighter colors up through orange correspond to progressively stronger or louder amplitudes.

higher than 44 100 Hz makes the sound smooth. Although the drill sounds in our dataset were recorded at the sample rate of 96 000, the sound of an operating drill ranges from roughly 1000 Hz to 22 000 Hz. Thus, low pass filter and high pass filter are performed in the passband of 1000–22 000 Hz.

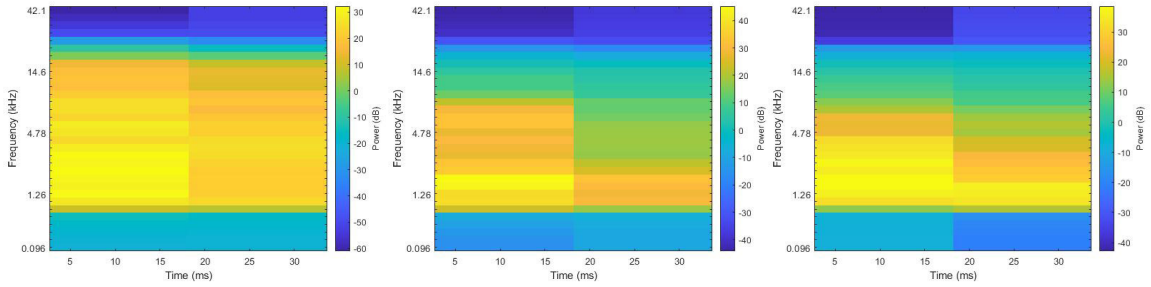
2) MEL SPECTROGRAM

Sound is usually visualized as an airwave that is a two-dimensional representation of amplitude and time. Figure 5(a) shows an example of a sound signal in the time domain. Sound can also be represented as a frequency spectrum of an audio signal as it varies with time. This is called a spectrogram. A spectrogram of sound is created from a time signal using the fast Fourier transform (FFT). ''Mel'' is short for melody. It implies that this is a perceptual scale measurement based on the comparison of the pitches.

A Mel spectrogram, a combination of the Mel scale and the spectrogram, is a visual representation of a drill sound in both frequency and amplitude by the time domains. The amplitude of a particular time is represented by colors. Brighter colors up through orange correspond to progressively stronger amplitudes, as shown in Figure 5(b). The horizontal axis presents the time from left to right. The vertical axis presents the frequency from low to high.

Frequencies in a sound signal change over time. Hence, the use of Fourier transforms on the entire audio signal results in a loss of meaningful frequency information in the time domain. Supposing the frequency of the sound signal is uniform for a very short period of time, each sound is divided into short time frames of 20 ms (2000-points windows) with a 512-point overlap between successive frames. The FFT length is 2000 points. Implementing Fourier transforms on these consecutive frames can help us obtain a good approximation of frequencies across the time domain.

A Hamming window is applied to each frame to greatly reduce spectral leakage before conducting FFT.



(a) A broken sound (b) A normal sound (c) An unrelated sound

FIGURE 6. The comparison between a broken sound, a normal sound, and an unrelated sound.

The Hamming window has the form [17]:

$$w(n) = \alpha_0 - (1 - \alpha_0)\cos(\frac{2\pi n}{N-1}), \quad 0 \le n \le N-1 \quad (1)$$

where $\alpha_0 = \frac{25}{46}$, and N is the window length.

Fast Fourier transform (FFT) using N-point is applied to calculate the power spectrum of each frame. Finally, frequency bands are extracted by applying the Mel filter bank on the power spectrum of each frame to obtain the Mel spectrogram. The Mel filter bank is composed of many triangles. Each triangle overlaps half of the next triangle. Each filter on the filter bank is triangular that has the value 1 at the center of the frequency. It is important to use the Mel filter bank because the Mel scale simulates the way a human ear reacts to a sound, by being more sensitive at the lower frequency and less so at the higher frequency. Mathematically, the Mel scale is the result of the frequency scale that is transformed in a non-linear way. The Mel scale (mel) is converted from the frequency (f) using the following formula [18]:

$$mel = 2595 \lg(1 + \frac{f}{700}) \quad (2)$$

Figure 6 shows the comparison of the Mel spectrogram of a broken sound, a normal sound, and a sound in the unrelated sound class. The stronger amplitudes corresponding to the orange region of the broken sound presented at the highest frequency. The orange region of the normal sound presented at the second-highest frequency and the unrelated sound has the lowest frequency. We can see the clear differences between the Mel spectrograms of the broken, normal, and unrelated sounds. Moreover, the normal drill sound recorded by any microphone has similarities to other normal drill sounds, regardless of the difference in volume, pitch, or timbre. Thus, Mel spectrograms of normal drill sounds are similar regardless of the length of time or the microphone used. This conclusion is also confirmed for sounds in the broken sound class. Because the sounds in the unrelated sound class are very diverse, Mel spectrograms of the sounds in the unrelated sound class may be different. However, strong power is distributed mainly at low frequencies, as shown in Figure 6(c). Consequently, the yellow region is located lower than in the Mel spectrograms of the normal sound and the broken sound. Based on the striking differences in the Mel spectrograms, we can classify the sounds of the drills.

3) SCALOGRAM

A scalogram is an image representation of the continuous wavelet transform (CWT) [19], [20]. CWT is a function that represents the frequency over time of sound waves. However, unlike a spectrogram, a scalogram is obtained by windowing a sound signal with a wavelet shifted in time. Contrariwise, a spectrogram is obtained when the sound signal is windowed with a window of constant length shifted in time and frequency. A scalogram is helpful for a short sound signal with high frequency. The CWT, $X(a, b)$ of a sound signal $x(t)$ is given by

$$X(a,b) = \frac{1}{\sqrt{2}} \int_{-\infty}^{\infty} x(t)\psi(\frac{t-b}{a})dt \quad (3)$$

where $\psi(t)$is a continuous function called an analyzing wavelet, a is a scale $a > 0$, $a \in R_+^*$, $b \in R$ is a translation value.

CWT uses a source function (mother wavelet) that is continuous in both time and frequency. There are different mother wavelets, such as the Morlet wavelet, Paul wavelet, Morse wavelet, and Bump wavelet. Because we are interested in the analysis of time-frequency, the Bump wavelet is used as a wavelet basis function (mother wavelet) in this research. The continuous function $\psi(t)$ is the Bump wavelet in our experiment and is defined as

$$\Psi(s\omega) = e^{1-\frac{1}{1-(s\omega-\mu)^2/\sigma^2}} 1_{]}\mu - \sigma/s, \mu + \sigma/s] \quad (4)$$

where $1_{]}\mu - \sigma/s, \mu + \sigma/s]$ is the indicator function for the interval $\mu - \sigma/s \le \omega \le \mu + \sigma/s$.

4) FEATURE EXTRACTION

The balanced dataset used for the experiment has only 67 Mel spectrogram images corresponding to 67 sounds for each class. Consequently, training a new CNN model from scratch would yield overfitting. To take advantage of the incredible performance of deep CNN architectures, we extracted informative features from Mel spectrogram images with a pre-trained CNN. A pre-trained network [21] is a saved

model that was trained on a large dataset such as ImageNet, MS COCO, etc. In this research, VGG19 architecture that was trained on the ImageNet dataset was utilized as the pre-trained network to extract features on our dataset. Although many state-of-the-art models have been proposed recently – ResNet, Inception V3, Xception, and so on – we chose VGG19 because this architecture is quite simple but it also performed well on our dataset for the feature extraction task.

The architecture of VGG19 is shown in Figure 7. The low-level features, such as edges, colors, and blobs are learned from the early layers. The high-level features are learned at the last three fully connected layers and a softmax layer for the specific classification task. Therefore, we extracted features at the global pooling layer that is marked as ''Pool 5'', as shown in Figure 7.

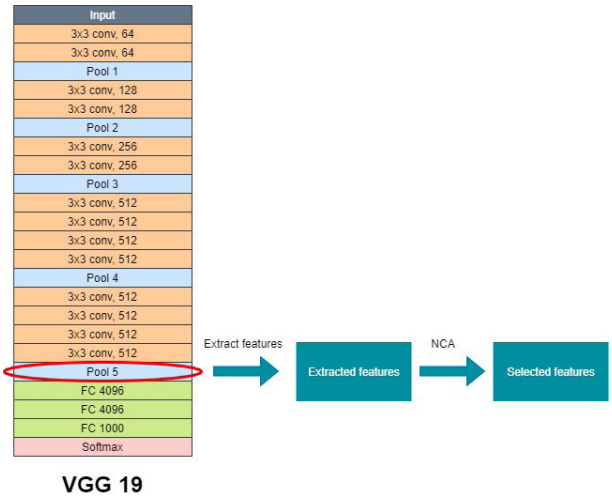


FIGURE 7. Feature extraction uses VGG19 and feature selection uses NCA.

5) FEATURE SELECTION AND CLASSIFICATION

The number of features extracted from Mel spectrogram images is 25 088. To reduce the unnecessary features, neighborhood component analysis (NCA) [6] was utilized for selecting the most relevant information for classification. Minimizing the redundancy of extracted features from VGG19 can help the model train faster and more effectively. We used classifiers provided in the Matlab toolbox to classify drill sounds using the extracted features from images.

III. RESULT

A. EXPERIMENT RESULTS

We use accuracy and F1-score to evaluate the performance of our proposed methods. The overall accuracy indicates the rate of correct classification. The F1-score is a measure arrived at by computing both the precision and the recall. In the F1-score, precision is the result of true positives divided by the sum of true positives and false positives. The recall is the number of true positives divided by the sum of true positives and false negatives. The confusion matrix shows the precise accuracy for each class in the classification task. Since the purpose is to detect damaged drills, accuracy is the most important consideration when classifying broken sounds.

In the following subsections, we show the experiment results for both approaches. The first way is to extract features from Mel spectrogram images, and the second way is to extract features from scalogram images. These extracted features are used to train various classifiers.

1) EXTRACT FEATURES FROM MEL SPECTROGRAM IMAGES AND CLASSIFY USING MACHINE LEARNING CLASSIFIERS

In this section, we converted all 201 drill sounds (67 broken sounds, 67 normal sounds, and 67 unrelated sounds) to Mel spectrogram images. In the next step, image features were extracted using the pre-trained network VGG19. The size of Mel spectrogram images is $875\times656\times3$, but the required size of input images for the VGG19 network is $224\times224\times3$. Therefore, we automatically reduced the size of images in our dataset before inputting them to the pre-trained network. Since we wanted to extract high-level features, we acquired the feature representations of the Mel spectrogram on the last max-pooling layer of VGG19. The dataset was divided into 60 percent for the training set and 40 percent for the testing set. Hence, we received 120-by-25 088 training features from the training set and 81-by-25 088 testing features from the testing set. We used NCA to reduce the number of features because we would like to use only the powerful predictive features (feature weights are higher than 2*6e-5), which carry the most meaningful information for the classification purpose. The number of features after reducing was 402. Figure 8 shows the chart of selection features where the y-axis shows the feature weights of the selected features.

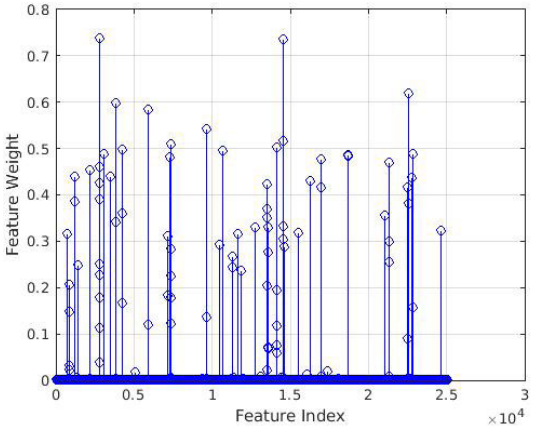


FIGURE 8. A plot of selected features. The y-axis presents feature weights. The x-axis presents feature indexes.

Afterward, these selected features were used to train various predictive models, such as SVM, KNN, Ensemble, etc. The purpose of training various classifiers was to compare the performance and select the best classifier for the classification task. Table 1 shows the experimental results of various classifiers. Medium Gaussian SVM and Quadratic

TABLE 1. The performance of classifiers using extracted features from Mel spectrogram images.

Classifier	Accuracy(%)	F1-Score
Fine KNN	75.31	0.75
Weighted KNN	75.31	0.75
Cubic SVM	76.54	0.77
Linear SVM	76.54	0.77
Ensemble Subspace KNN	76.54	0.76
Linear Discriminant	79.01	0.79
Quadratic SVM	**80.25**	**0.8**
Medium Gaussian SVM	**80.25**	**0.8**

SVM show the best overall accuracy (80.25 percent) and F1-score (0.8) when classifying three classes of drill sounds.

Table 2 shows the confusion matrix for the best classifiers, Medium Gaussian SVM and Quadratic SVM. Since the goal of this research is to detect broken drills based on their sound, the accuracy of the broken sounds class is of greatest concern. We can notice that the precision for the ''Broken sounds'' class reached 81.48 percent, which is desirable for the industry practice, as shown in Table 2. After careful analysis of the confusion matrix of eight classifiers, the drill sounds from the ''Unrelated sounds'' class were more difficult to distinguish from the ''Normal sounds'' class. The reason is that the features of the ''Unrelated sounds'' class are similar to the features of the ''Normal sounds'' class. The detail of the configuration for SVM models is described as follows. We used fitcecoc function in Matlab 2019b with the input argument was a table of selected features, a coding design was ''onevsall'' and the learner template was templateSVM function in Matlab. The parameters for templateSVM function included: PolynomialOrder was 2, kernel functions were ''polynomial'' for Quadratic SVM and ''gaussian'' for Medium Gaussian SVM, respectively, kernel scale parameters were ''auto'' for Quadratic SVM and 30 for Medium Gaussian SVM, the box constraint was set to 1, and standardize was ''true'' (it means Matlab centers and scales each column of the predictor data by the weighted column mean and standard deviation, respectively).

TABLE 2. The confusion matrix of the Quadratic SVM and Medium Gaussian SVM. The first number represents the Quadratic SVM and the second number represents the Medium Gaussian SVM, illustrated as Quadratic SVM/Medium Gaussian SVM format.

Labels	Broken sounds (%)	Normal sounds (%)	Unrelated sounds (%)
Broken sounds	**81.48/81.48**	3.704/7.407	14.81/11.11
Normal sounds	7.407/7.407	**85.19/85.19**	7.407/7.407
Unrelated sounds	7.407/3.704	18.52/22.22	**74.07/74.07**

Although Medium Gaussian SVM and Quadratic SVM have the same overall accuracy, the accuracy when classifying the broken sounds class of Medium Gaussian SVM is higher than Quadratic SVM, as can be seen in the ''Accuracy'' column of Table 3. The accuracies for the ''Broken sounds'' class reach 90.12 percent and 88.89 percent when using

TABLE 3. The accuracy, precision, recall, and F1-score per class of the Quadratic SVM and Medium Gaussian SVM. The first number represents the Quadratic SVM, and the second number represents the Medium Gaussian SVM, illustrated as Quadratic SVM/Medium Gaussian SVM format.

Class	Accuracy (%)	Precision	Recall	F1-Score
Broken sounds	88.89/**90.12**	0.81/**0.81**	0.85/**0.88**	0.83/**0.85**
Normal sounds	87.65/**85.19**	0.85/**0.85**	0.79/**0.74**	0.82/**0.79**
Unrelated sounds	83.95/**85.19**	0.74/**0.74**	0.77/**0.8**	0.75/**0.77**

Medium Gaussian SVM and Quadratic SVM, respectively. Table 3 shows the accuracy, precision, recall, and F1-score for the Quadratic SVM and Medium Gaussian SVM. These evaluation metrics are parameters to measure the performance of the Quadratic SVM and Medium Gaussian SVM. For example, precision for the ''Broken sounds'' class measures how often a classifier correctly predicts broken sound for a group of different sounds. The precision for the ''Broken sounds'' class in Table 3 is 0.81. However, we want to predict the broken sound even when we are not sure it is a broken sound. Hence, recall is a good evaluation metric for fault sound detection. The recall for the ''Broken sounds'' class of Medium Gaussian SVM reaches 0.88. It means the false negative (predicted as not a broken sound but where the drill actually had broken) is low. Because both precision and recall are important metrics in evaluating the performance of a classification model, F1-score is a balancing metric between precision and recall.

2) EXTRACT FEATURES FROM SCALOGRAM IMAGES AND CLASSIFY USING MACHINE LEARNING CLASSIFIERS

Figure 9 shows the magnitude scalogram (CWT) with the cone of influence (COI) of drill sounds using the bump wavelet. The cone of influence and the scalogram is displayed during the time period (the x-axis), as shown in Figure 9. The gray region from the dashed white line to the x-axis and y-axis is the cone of influence. This indicates areas in the scalogram chart that may be affected by edge effects. The edge effects are strongly significant in the shaded gray region outside of the white dashed line. Edge effects are effects in scalograms that arise from regions where extended wavelengths expand beyond the edges of the observation interval. Therefore, the gray side area outside the dashed white line is an unreliable representation of a scalogram, whereas the area inside the cone represents information that is reliable and accurate in a scalogram.

Similar methods to those described above are used for scalogram images. Pre-trained VGG19 is also utilized to extract features from 201 scalogram images (67 broken sounds, 67 normal sounds, and 67 unrelated sounds). As a result, we obtained 25 088 training features from the training set. After using NCA to reduce the number of features, there

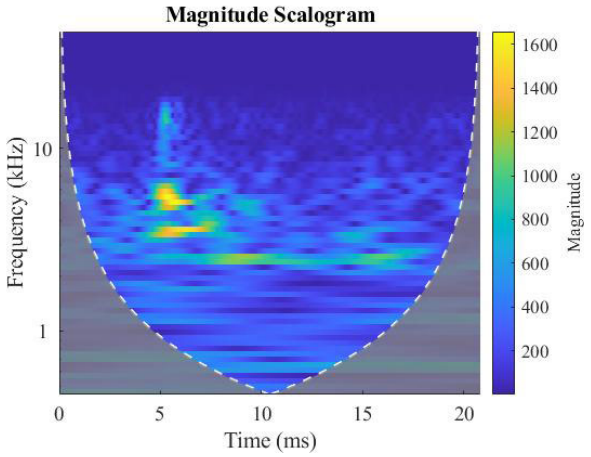


FIGURE 9. The magnitude scalogram (CWT) of a broken drill sound.

are 448 features selected. The chart of selected features is shown in Figure 10.

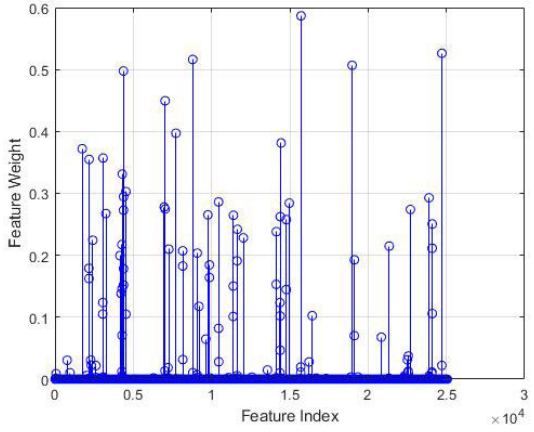


FIGURE 10. A plot of selected features. The *y*-axis presents feature weights. The *x*-axis presents feature indexes.

Table 4 shows the accuracy and F1-score of different classifiers. It is clear that a scalogram is a good image representation for a sound signal on the classification task. These features, which are extracted from scalograms of drill sounds, are meaningful for diverse classifiers. The overall accuracy rises from approximately 75.31 percent (Linear Discriminant) to 80.25 percent (Subspace KNN).

The confusion matrix for the best classifier is shown in Table 5, which shows that the accuracy for the ''Broken sounds'' class reaches 81.48 percent. The accuracy of the ''Normal sounds'' class is the highest in the confusion matrix. The precision, recall, and F1-score for each class are shown in Table 6. The experiment results prove that the Ensemble Subspace KNN classifier provides the best result for the drill sound classification using scalograms, as shown in Table 4. Detail configuration for the Ensemble Subspace KNN model is described as follows. We used the fit ensemble function in Matlab 2019b with the input argument as a table of selected features. The ensemble aggregation method was ''subspace'',

TABLE 4. The best performance of classifiers using extracted features from scalogram images.

Classifier	Accuracy(%)	F1-Score
Linear discriminant	75.31	0.75
Linear SVM	76.54	0.77
Medium KNN	76.54	0.77
Weighted KNN	76.54	0.77
Quadratic SVM	77.78	0.78
Fine KNN	77.78	0.78
Cubic SVM	79.01	0.79
Medium Gaussian SVM	79.01	0.79
Ensemble Subspace KNN	**80.25**	**0.80**

TABLE 5. The confusion matrix of the best classifier (Subspace KNN).

Labels	Broken sounds (%)	Normal sounds (%)	Unrelated sounds (%)
Broken sounds	**81.48**	11.11	7.407
Normal sounds	3.704	**85.19**	11.11
Unrelated sounds	3.704	22.22	**74.07**

TABLE 6. The precision, recall, and F1-score per class of the best classifier (Subspace KNN).

Class	Precision	Recall	F1-score
Broken sounds	0.81	0.92	0.86
Normal sounds	0.85	0.72	0.78
Unrelated sounds	0.74	0.8	0.77

the number of ensemble learning cycles were 30, the weak learner to use in the ensemble is k-nearest neighbors, and the number of predictors to sample was the number of selected features.

B. COMPARATIVE STUDY

Most previous works were conducted on different datasets or different kinds of signals. For example, approaches utilized ultrasonic data or vibration signals. Moreover, some previous work experimented with large and balanced datasets. In addition, the accuracy of the system can be affected by the length of the sound. As a result, we do not provide a precise and flawless comparison. The aim of this comparison part is to provide the discrimination potential of different machine fault detection methods, in particular, the conventional machine learning method and deep learning CNN architecture. The conventional machine learning methods were conducted using manually extracted features as the input of the traditional machine learning classifiers, as presented in section III.B.1. We also examined the use of Mel spectrogram images or scalogram images as an input to a CNN architecture, as figured out in section III.B.2.

1) EXTRACT FEATURES FROM RAW SOUND SIGNALS AND CLASSIFY USING A MACHINE LEARNING CLASSIFIER

This section demonstrates a machine learning approach to classify drill sounds based on the extracted features of the raw sound signals. The dataset was also split into two

parts so that 60 percent for each class was used for training, and the remaining 40 percent was used for testing. Pitch and 13 MFCCs were extracted from the sounds by a HelperComputePitchAndMFCC function in the library of Matlab 2019b. The extracted features were normalized by subtracting the mean and dividing the standard deviation. These features were used to train a K-nearest neighbor (KNN) classifier by fitcknn function in Matlab 2019b. KNN proves to be a suitable classifier for the classification of multiple classes, as seen in the results from the experimental evidence in section III.A, in which the number of nearest neighbors is 5. Euclidean distance was the distance metric to compute the distance to the neighbor. The distance weight was calculated by the inverse of distance squared (Distance weighting function was set to ''squaredinverse''). The standardized parameter of fitcknn function was set to ''False''. The overall validation accuracy reaches 67.49 percent. The confusion matrix of KNN is visualized in Table 7.

TABLE 7. The confusion matrix of KNN classifier.

Labels	Broken sounds (%)	Normal sounds (%)	Unrelated sounds (%)
Broken sounds	**67.62**	18.57	13.81
Normal sounds	8.095	**76.9**	15
Unrelated sounds	13.98	28.07	**57.95**

2) USING MEL SPECTROGRAM OR SCALOGRAM IMAGES AS THE INPUT OF THE CNN ARCHITECTURE

Because the small number of Mel spectrogram or scalogram images is not enough to train an end-to-end CNN model, transfer learning with GoogLeNet is applied for the classification task. GoogLeNet is a pre-trained image classification network that has been trained on the ImageNet dataset to classify 1000 object categories. The dataset was also divided into two parts, 70 percent for the training part and 30 percent for the testing part (60 images). Because the size of the input images for GoogLeNet is $224\times224\times3$, all images in the dataset need to be resized to $224\times224\times3$. Data augmentation methods were used to increase the number of images in the training part: reflexed images via x-axis, translated images randomly via x-axis and y-axis in a range of [-30 30] pixels, and scaled images by a random scale in the range of [0.9 1.1]. The fully connected layer of GoogLeNet was replaced with a new fully connected layer in which the number of outputs is three because there are three classes in which to classify. We set the learning rates in earlier layers to zero so the network did not update the parameters of these layers. This helps reduce the training time and prevents overfitting to the new small dataset. The initial learning rate was set to a small number of $3e^{-4}$ to slow down the training in the transferred layers. The batch size was 20. The maximum epochs were 20 because we did not need to train many epochs for transfer learning.

The overall accuracy when using transfer learning to retrain GoogLeNet to classify Mel spectrogram and scalogram datasets reached 70 and 75 percent, respectively. The confusion matrices for Mel spectrogram and scalogram datasets are shown in Table 8 and Table 9.

TABLE 8. The confusion matrix when using transfer learning to retrain GoogLeNet to classify Mel spectrogram dataset.

Labels	Broken sounds (%)	Normal sounds (%)	Unrelated sounds (%)
Broken sounds	**75**	20	5
Normal sounds	0	**80**	20
Unrelated sounds	10	35	**55**

TABLE 9. The confusion matrix when using transfer learning to retrain GoogLeNet to classify scalogram dataset.

Labels	Broken sounds (%)	Normal sounds (%)	Unrelated sounds (%)
Broken sounds	**80**	15	5
Normal sounds	5	**75**	20
Unrelated sounds	10	20	**70**

C. DISCUSSION

The experiment results for all mentioned methods are shown in Figure 11. The method, which extracts pitch and MFCC features and classifies three classes using KNN, shows the worst result. The accuracy only reaches 67.49 percent for this method. For the Mel spectrogram images, the classification accuracy reaches 70 percent with the deep learning CNN architecture GoogLeNet, and 80.25 percent with our proposed procedures (feature extractions, feature selection, and Weighted KNN classifier). Similarly, for the scalogram images, the classification accuracy reaches 75 percent and 80.25 percent when classifying using GoogLeNet and our proposed procedures, respectively. It is clear that our proposed methods reach outstanding results.

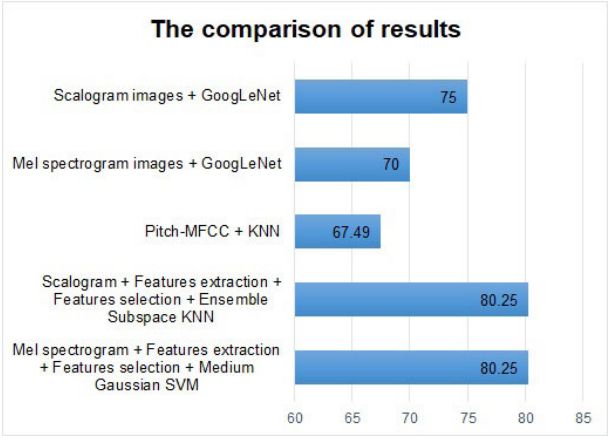


FIGURE 11. Results of different methods.

Moreover, for the comparison between scalogram images and Mel spectrogram images, we use both of them as the input

of GoogLeNet. The scalogram images achieved 75 percent accuracy, 5 percent higher than Mel spectrogram images. When using scalogram and Mel spectrogram images as the input of our proposed procedures (feature extractions, feature selection, and the classifier), both of them achieve 80.25 percent accuracy. In conclusion, scalogram images prove more effective for classifying drill sound signals when used as inputs to both a CNN architecture and our own proposed procedures.

IV. CONCLUSION

In this research, we propose a novel approach for drill sound classification that consists of using image representations of sounds (Mel spectrogram and scalogram) and extracting features from these images based on deep learning CNN. We also utilize NCA to reduce the number of extraction features from CNN before inputting them to the machine learning classifiers. Our proposed methods achieve 80.25 percent on scalogram and Mel spectrogram images. The obtained result is promising with regard to applications for early fault drill detection in the industry.

For the comparison study, we compare two time-frequency analyses (Mel spectrogram and scalogram) when using these images as the input of our proposed methods. Scalogram images obtained using the bump wavelet with the COI contribute to boosting the performance of our procedures. Moreover, we also experimented with the traditional method (classifying drill sounds using pitch and 13 MFCCs features as the input of a KNN classifier) and the state-of-the-art deep convolutional neural network on our dataset as the comparison. The experiment results prove the robustness of using image representation of drill sounds for classifying drill sounds whether using a CNN architecture such as GoogLeNet as a classifier or using our proposed procedures.

ABBREVIATIONS

The following abbreviations are used in this manuscript:

ANN	Artificial neural network
CEEMD	Complete ensemble empirical mode decomposition
CNN	Convolutional neural network
COI	The cone of influence
CWT	Continuous wavelet transform
DAE	Deep auto-encoder
ESN	Echo state network
FFT	Fast Fourier transform
IMF	Intrinsic mode functions
KNN	K-nearest neighbor
MFCCs	Mel-frequency cepstrum coefficients
MFD	Machine fault detection
NCA	Neighborhood component analysis
SAE	Sparse autoencoder
SOM	Self-organizing maps
STFT	Short-time Fourier transform
SVM	Support vector machine
PCA	Principal component analysis

ACKNOWLEDGMENT

This research was funded by EU Regional Fund and the MiLo Project (No. 20201888). Thanh Tran thanks Valmet AB for providing the drill sound dataset.

REFERENCES

[1] P. A. Delgado-Arredondo, D. Morinigo-Sotelo, R. A. Osornio-Rios, J. G. Avina-Cervantes, H. Rostro-Gonzalez, and R. D. J. Romero-Troncoso, "Methodology for fault detection in induction motors via sound and vibration signals," *Mech. Syst. Signal Process.*, vol. 83, pp. 568–589, Jan. 2017, doi: 10.1016/j.ymssp.2016.06.032.

[2] P. K. Kankar, S. C. Sharma, and S. P. Harsha, "Fault diagnosis of ball bearings using machine learning methods," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 1876–1886, Mar. 2011, doi: 10.1016/j.eswa.2010.07.119.

[3] A. Kumar, J. Ramkumar, N. K. Verma, and S. Dixit, "Detection and classification for faults in drilling process using vibration analysis," in *Proc. Int. Conf. Prognostics Health Manage.*, Jun. 2014, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/7036393/

[4] J. Lee, H. Choi, D. Park, Y. Chung, H.-Y. Kim, and S. Yoon, "Fault detection and diagnosis of railway point machines by sound analysis," *Sensors*, vol. 16, no. 4, p. 549, Apr. 2016. [Online]. Available: http://www.mdpi.com/1424-8220/16/4/549

[5] A. K. Kemalkar and V. K. Bairagi, "Engine fault diagnosis using sound analysis," in *Proc. Int. Conf. Automat. Control Dyn. Optim. Techn. (ICACDOT)*, Sep. 2016, pp. 943–946. [Online]. Available: http://ieeexplore.ieee.org/document/7877726/

[6] N. Zhang, "Research on automatic fault diagnosis system of coal mine drilling rigs based on drilling parameters," in *Proc. IEEE 4th Adv. Inf. Technol., Electron. Automat. Control Conf. (IAEAC)*, Dec. 2019, pp. 2373–2377. [Online]. Available: https://ieeexplore.ieee.org/document/8997875/

[7] T. Ince, S. Kiranyaz, L. Eren, M. Askar, and M. Gabbouj, "Real-time motor fault detection by 1-D convolutional neural networks," *IEEE Trans. Ind. Electron.*, vol. 63, no. 11, pp. 7067–7075, Nov. 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7501527/

[8] B. Luo, H. Wang, H. Liu, B. Li, and F. Peng, "Early fault detection of machine tools based on deep learning and dynamic identification," *IEEE Trans. Ind. Electron.*, vol. 66, no. 1, pp. 509–518, Jan. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8294247/

[9] J. Long, Z. Sun, C. Li, Y. Hong, Y. Bai, and S. Zhang, "A novel sparse echo autoencoder network for data-driven fault diagnosis of delta 3-D printers," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 3, pp. 683–692, Mar. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8699100/

[10] J. Long, S. Zhang, and C. Li, "Evolving deep echo state networks for intelligent fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4928–4937, Jul. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8827669/

[11] S. Paul and M. Lofstrand, "Intelligent fault detection scheme for drilling process," in *Proc. 7th Int. Conf. Control, Mechatronics Automat. (ICCMA)*, Nov. 2019, pp. 347–351. [Online]. Available: https://ieeexplore.ieee.org/document/8988616/

[12] A. Khamparia, D. Gupta, N. G. Nguyen, A. Khanna, B. Pandey, and P. Tiwari, "Sound classification using convolutional neural network and tensor deep stacking network," *IEEE Access*, vol. 7, pp. 7717–7727, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8605515/

[13] J. Wang, J. Zhuang, L. Duan, and W. Cheng, "A multi-scale convolution neural network for featureless fault diagnosis," in *Proc. Int. Symp. Flexible Automat. (ISFA)*, Aug. 2016, pp. 65–70. [Online]. Available: http://ieeexplore.ieee.org/document/7790137/

[14] H. Liu, L. Li, and J. Ma, "Rolling bearing fault diagnosis based on STFT-deep learning and sound signals," *Shock Vib.*, vol. 2016, pp. 1–12, Jul. 2016. [Online]. Available: http://www.hindawi.com/journals/sv/2016/6127479/

[15] O. Janssens, V. Slavkovikj, B. Vervisch, K. Stockman, M. Loccufier, S. Verstockt, R. Van de Walle, and S. Van Hoecke, "Convolutional neural network based fault detection for rotating machinery," *J. Sound Vib.*, vol. 377, pp. 331–345, Sep. 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0022460X16301638

[16] W. Yang, K. Wang, and W. Zuo, "Neighborhood component feature selection for high-dimensional data," *J. Comput.*, vol. 7, no. 1, pp. 162–168, Jan. 2012. [Online]. Available: http://ojs.academypublisher.com/index.php/jcp/article/view/5076

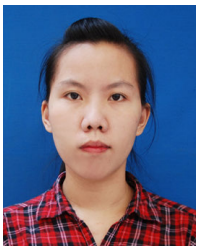
[17] R. W. Schafer and V. A. Oppenheim, *Discrete-Time Signal Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 1999.

[18] D. O'Shaughnessy, "Hearing," in *Proc. Speech Commun., Hum. Mach.*, 2000, pp. 109–139, doi: 10.1109/9780470546475.ch4.

[19] M. Stéphane, "Sparse representations," in *A Wavelet Tour of Signal Processing*, S. Mallat, Ed. Amsterdam, The Netherlands: Elsevier, 2009, ch. 1, pp. 17–18. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9780123743701000057

[20] J. M. Lilly, "Element analysis: A wavelet-based method for analysing time-localized events in noisy time series," *Proc. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 473, no. 2200, Apr. 2017, Art. no. 20160776. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rspa.2016.0776

[21] F. Chollet, *Deep Learning With Python*. Shelter Island, NY, USA: Manning Publications Co., 2018. [Online]. Available: https://dl.acm.org/doi/book/10.5555/3203489#cited-by-sec



THANH TRAN received the M.S. degree from Pukyong National University, Busan, South Korea, in 2019. She is currently pursuing the Ph.D. degree with Mid Sweden University. Her research interests include digital signal and image processing, industrial sound measurements, bioinformatics, and deep learning.

JAN LUNDGREN received the Ph.D. degree from Mid Sweden University, Sundsvall, Sweden, in 2007.

After receiving his Ph.D. degree, he has worked as an Assistant Professor with Mid Sweden University, where he has been an Associate Professor, since 2013. He currently leads the Research Group, Mid Sweden University, focusing on AI-supported sensor systems, including industrial sound and imaging measurements. His work and research interests include industrial sound measurements, high-resolution real-time surface characterization, industrial soft sensors, and on-chip noise characterization.