Received February 14, 2022, accepted March 10, 2022, date of publication March 16, 2022, date of current version April 4, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3160179

An EfficientNet-Based Weighted Ensemble Model for Industrial Machine Malfunction Detection Using Acoustic Signals

BAYU ADHI TAMA[1], MALINDA VANIA[2,3], ILJUNG KIM[4], AND SUNGHOON LIM[2,3]

[1]Center for Mathematical and Computational Sciences, Data Science Group, Institute for Basic Science (IBS), Yuseong-gu, Daejeon 34126, Republic of Korea
[2]Department of Industrial Engineering, Ulsan National Institute of Science and Technology, Eonyang-eup, Ulju-gun, Ulsan 44919, Republic of Korea
[3]Institute for the 4th Industrial Revolution, Ulsan National Institute of Science and Technology, Eonyang-eup, Ulju-gun, Ulsan 44919, Republic of Korea
[4]Manufacturing AI Big Data Centre, Korea Advanced Institute of Science and Technology, Yuseong-gu, Daejeon 34051, Republic of Korea

Corresponding author: Sunghoon Lim (sunghoonlim@unist.ac.kr)

This work was supported in part by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. 2021-0-01139, 3D Pose Estimation Motion Data Development based on the Fusion of 3D Data and AI), in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1F1A1046416), in part by the AI Collaboration Project Fund (1.220083) of UNIST (Ulsan National Institute of Science and Technology), and in part by the Institute for Basic Science (IBS) under Grant IBS-R029-C2-001.

ABSTRACT Detecting and preventing industrial machine failures are significant in the modern manufacturing industry because machine failures substantially increase both maintenance and manufacturing costs. Recently, state-of-the-art deep learning techniques that use acoustic signals have been widely applied to solve industrial machine malfunction detection problems in order to reduce maintenance and manufacturing costs. The authors of this research propose a deep learning-based industrial machine malfunction detection model that uses acoustic signals to classify normal and abnormal conditions of industrial machines. In particular, a weighted ensemble model based on EfficientNet-B0, B5, and B7 is considered to improve classification performance. Case studies involving an open dataset for Malfunctioning Industrial Machine Investigation and Inspection (MIMII) validate that the proposed EfficientNet-based weighted ensemble model provides better classification performance than individual classifiers and other ensemble models.

INDEX TERMS Weighted ensemble, convolutional neural networks, industrial machines, malfunction detection, acoustic signals.

LIST OF ABBREVIATIONS

AE	AutoEncoder.
AUC	Area Under the Curve.
CNN	Convolutional Neural Network.
CUDA	Compute Unified Device Architecture.
DCASE	Detection and Classification of Acoustic Scenes and Events.
DenseAE	Dense AutoEncoder.
DenseNet	Dense Network.
FCN	Fully Convolutional Network.
FLOPs	Floating-point Operations.
FN	False Negative.
FP	False Positive.
GPU	Graphics Processing Unit.
HPSS	Harmonic and Percussive Source Separation.
IDNN	Interpolation Deep Neural Network.
IoT	Internet of Things.
Kappa	Cohen's Kappa Coefficient.
MCC	Matthews Correlation Coefficient.
MDF	Motif Difference Field.
MFCC	Mel Frequency Cepstral Coefficient.
MIMII	Malfunctioning Industrial Machine Investigation and Inspection.
MV	Majority Voting.
PHM	Prognostics and Health Management.
ResNet	Residual Network.
ROC	Receiver Operating Characteristic.
SGD	Stochastic Gradient Descent.
SNR	Signal to Noise Ratio.
SVM	Support Vector Machine.
t-SNE	t-Distributed Stochastic Neighbor Embedding.
TN	True Negative.
TP	True Positive.

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

VAE	Variational AutoEncoder.
WMV	Weighted Majority Voting.
XAI	Explainable Artificial Intelligence.

I. INTRODUCTION

In the modern manufacturing industry, industrial machines are prone to failure due to their age-long operations. Detecting and preventing machine failures are important for the manufacturing industry, because machine failures significantly affect maintenance cost management and manufacturing cost management. Due to the development of the IoT and data-driven methods in manufacturing, utilizing large amounts of data collected from linked machines and big manufacturing data has been enabled. Several methods for monitoring industrial machines' conditions that use various sensors and microphones, such as collecting speeds from vibration signals using accelerometers [1], temperature-based sensors [2], and pressure-based sensors [3], have been proposed. Recently, research on machinery failure predictions using anomalous sounds has been developed rapidly and a large number of state-of-the-art approaches in this field have been proposed [4]–[8].

The manufacturing data can be transformed into meaningful and actionable information intelligence using data-driven approaches [9], [10]. Data-driven approaches bring a new paradigm into the modern industry for both fault detection of particular malfunctions (e.g., diagnosis) as well as PHM [9], [11]. As opposed to physics-based approaches, where modeling noisy and complex systems is not straightforward, data-driven approaches, which are more effective and flexible for fault prediction, can be updated using real-time manufacturing data [12], [13]. Data-driven approaches heavily rely upon powerful tools, machine learning techniques, to extract meaningful information from raw data [14]. Deep learning, a cutting-edge subsidiary of machine learning, possesses a remarkable role as a link that bridges big manufacturing data and fault prediction. Deep learning aims to model structured hierarchy representations underlying the data and categorize them by stacking multiple layers of hierarchical architecture-based information processing [15], [16]. Recently, the potential of deep learning has been increasing due to increased computing power, increased size of available data, and the development of state-of-the-art deep learning techniques. Since deep learning can cope with a large size of manufacturing data and learn from hierarchical representations, it is also a promising technique for fault detection [17], [18].

In this work, a deep learning-based industrial machine malfunction detection model that uses acoustic signals, which are extracted from industrial machines, is proposed to classify the machines' normal and abnormal conditions. In order to improve classification performance, an ensemble approach that integrates multiple weak learning classifiers is considered. In particular, the EfficientNet-based weighted ensemble model (hereinafter called as WMV), which provides better classification performances than other ensemble models, is used in this work. Furthermore, an open benchmark dataset (i.e., MIMII [19]) is used for the experiments that validate the classification performances of the proposed model.

In this work, a WMV model is proposed for industrial machine malfunction detection using acoustic signals with the following main contributions:

- To our knowledge, this work is the first to utilize the EfficientNet backbone network in an ensemble model for supervised industrial machine malfunction detection.
- Utilizing the EfficientNet backbone network with a weighted ensemble strategy has shown increased accuracy in supervised industrial machine malfunction detection compared to state-of-the-art models. Unlike the common ensemble method strategy, which assigns an equal weight for each ensemble member, this work utilizes different weights to determine each ensemble member's contribution that indicates the trust or expected performance of the model.
- Rather than the traditional exhaustive grid search approach, this work adopts the Dirichlet distribution process to identify and assign an appropriate weight for each model in order to adapt the contribution's importance of the weighted ensemble model at hand.
- This work demonstrates the system's capabilities using mixed ensemble learning algorithms and an EfficientNet backbone to mitigate the detrimental effects of overfitting and initialization that improve the efficiency of supervised industrial machine malfunction detection. The proposed method's efficiency is demonstrated through detailed experimental results.

The remainder of the paper is structured as follows. Section II describes the datasets utilized in this work, the data augmentation techniques, and the suggested weighted ensemble model based on EfficientNet. The experimental results and discussion are presented in Section III. Section IV details related works, and the paper concludes in Section V.

II. MATERIAL AND METHOD

This work is comprised of five primary stages (see Figure 2). First, audio data representing normal and abnormal operating conditions of industrial machines (i.e., valves and pumps) are selected. Then, audio file augmentation techniques are applied to increase the number of samples that have abnormal operating conditions, followed by a data conversion into spectrogram images. An EfficientNet-based weighted ensemble model that uses audio data is then proposed to classify normal and abnormal operating conditions. Finally, weighting strategies are considered to improve classification performance.

A. DATASET

The MIMII dataset contains 26,092 sound recordings representing normal operating circumstances and 6,065 sound recordings representing abnormal operating conditions for four different kinds of industrial machinery (i.e., pumps,

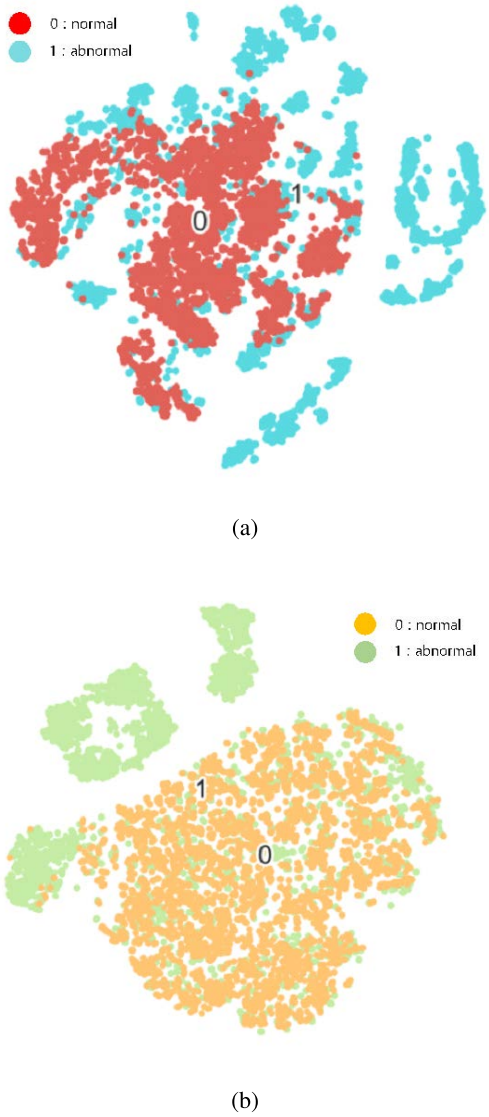

FIGURE 1. Dataset visualization in a two-dimensional space of (a) pumps and (b) valves that represent two distinct operational conditions (i.e., normal and abnormal).

valves, fans, and slide rails). Industrial machinery emits two distinct sorts of sounds: stationary and non-stationary. Additionally, these sounds vary in terms of characteristics and degree of difficulty. The sound recordings for pumps and valves are chosen for this work because of their increased difficulties in diagnosing system faults compared to other machine types (i.e., fans and slide rails). Figure 1 shows the t-SNE for sound files of pumps and valves used in this work. The t-SNE defines a soft border between the local and global structures of the data. The t-SNE determines the local neighborhood size for each data point separately based on the local density of the data by forcing each conditional probability distribution to have the same perplexity [20].

B. DATA AUGMENTATION

Data augmentation is a technique for generating additional training data from the original data. Data augmentation cannot replace actual training data. However, it may assist in generating synthetic data, which allow the model to function more effectively. Based on the audio files used in this work, several appropriate ways are considered to increase the amount of training data. To artificially increase the dataset size and diversity of the data in this work, seven well-known methods for audio file augmentation utilizing the Librosa library [21] are applied, including the following:

1) *Change pitch*: randomly change the baseline pitch in a file between −4 and 4 provided for favorable outcomes per audio files selected for this work.
2) *Change speed*: stretch the time series of audio files by a fixed rate.
3) *Change pitch and speed*: combination of the first and second methods.
4) *Random shifting*: shift the audio to the right (back forward) or left (fast forward) with a random second.
5) *Stretching*: change the speed or the duration of the sound without affecting the pitch of the sound. This method takes wave samples and a factor by which to stretch as inputs. A factor of 0.45 is used in this work, since it has a small difference from the original audio files.
6) *Value augmentation*: add some small random value into the data to alter the quality of the audio files so that they only differ by small factors from the original audio files.
7) *Add distribution noise*: add some random values into the data using NumPy and several options of data distributions, such as Gaussian distribution, Beta distribution, log-normal distribution, etc.

TABLE 1. Total amount of audio samples for each machine dataset after data augmentation.

Machine type	# of normal	# of abnormal
Pump	3,749	3,648
Valve	3,691	3,832

These augmentations will not affect the quality of the audio files and ensure that the synthetic audio files only differ by small factors from the original audio files. Two Librosa methods, shift silence and HPSS method, are not used. Shift silence is not appropriate because the data do not have a silence sequence, therefore it does not make any difference. The HPSS method divides one sound sample into harmonic or percussive sound. This means a mono sound file is changed into a stereo sound file. Stereo sound has two sound data, meaning they have a spectrogram for each type (i.e., harmonic and percussive), which will cause the data to deviate significantly from the original audio files. As a result, a total of 7,397 audio files for pumps and 7,523 audio files for valves are obtained, respectively. The details of both audio sample types are shown in Table 1.

C. EfficientNet-BASED WEIGHTED ENSEMBLE MODEL

An ensemble model can integrate multiple weak learning classifiers to obtain better predictive performance [22]. The

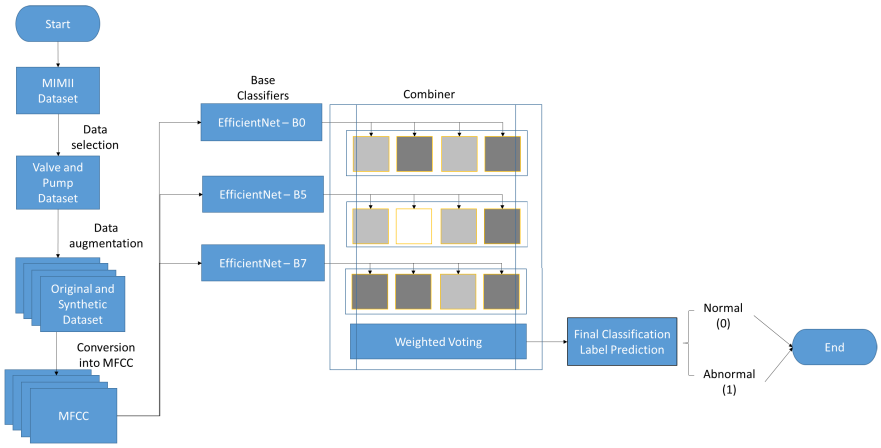

FIGURE 2. A workflow of the proposed approach.

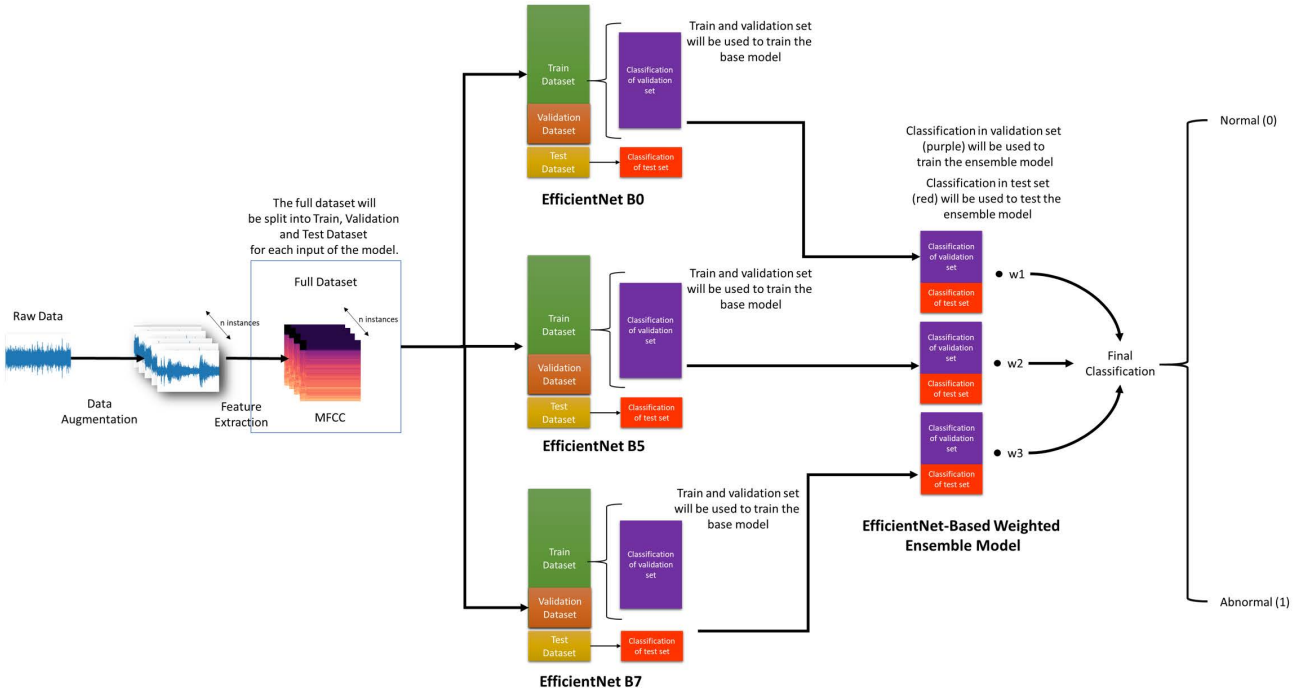

FIGURE 3. A detail framework of the proposed EfficientNet-based weighted ensemble model.

two voting models (i.e., hard-voting and soft voting) are first considered as an ensemble classifier for this work. In both voting models, the weights of the classifiers are equal. It indicates that the different predictive performances of the weak classifiers across the machine malfunction types cannot be fully used. To address this issue, a new ensemble model (i.e., WMV) is proposed in this work, as shown in Figure 3. A weighted ensemble is an extension of model averaging ensembles, where we rank members of an ensemble according to their contribution to the final prediction. Therefore, the multiple output model weights are weighted differently among the classifiers based on their predictive performance. Specifically, each classifier is assigned a distinct weight, determined by the classifier's performance.

In the experiments, the EfficientNet-B0, B5 and B7 are chosen. These architectures are chosen based on the performance results reported in the original paper [23]. The EfficientNet-B0 has the lowest performance among the EfficientNet family but outperforms ResNet-50 and Inception-v2. Additionally, EfficientNet-B5 is chosen, because it outperforms DenseNet, ResNet-152, Inception-ResNet-v2, and even AmoebaNet-A, all of which are currently employed in this field. Lastly, EfficientNet-B7

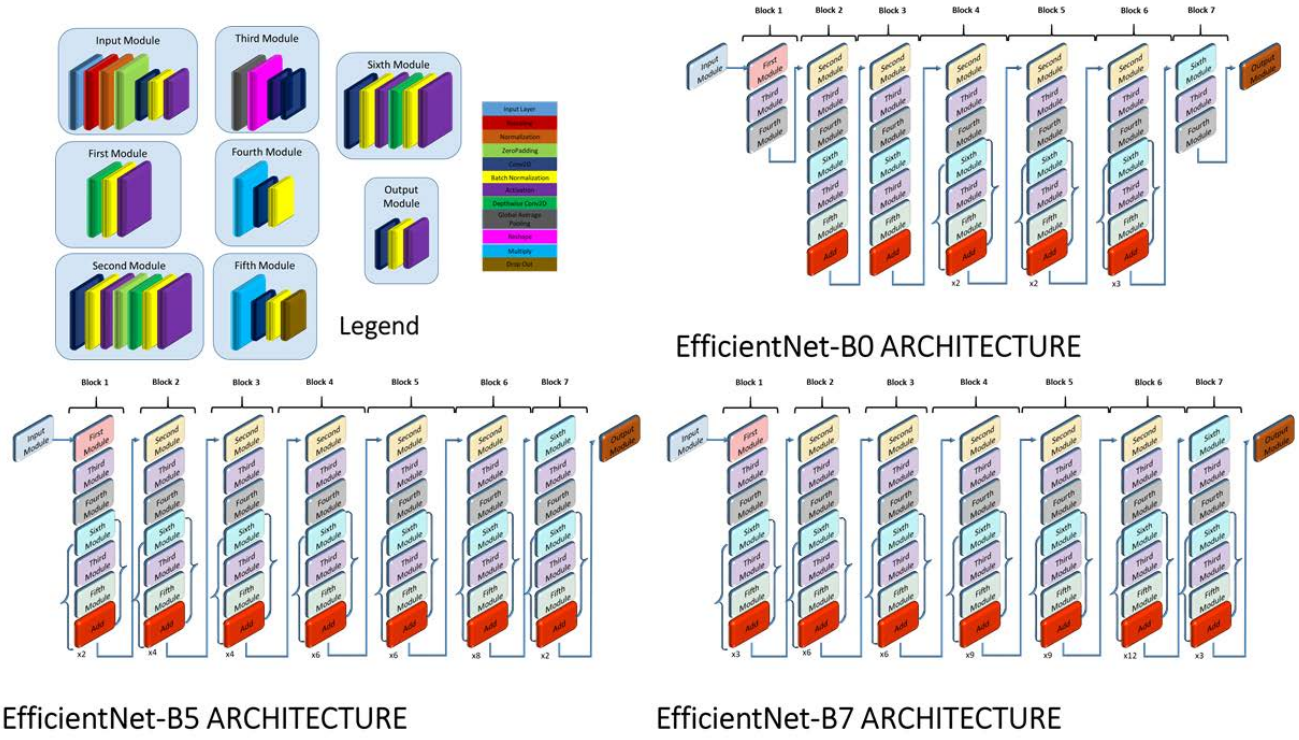

FIGURE 4. EfficientNet architecture comparison of EfficientNet-B0, EfficientNet-B5, and EfficientNet-B7.

is chosen as the best performer among the EfficientNet family, AmoebaNet family, ResNet family, and SENet [23]. EfficientNet-B7 can improve the performance of the proposed model. As a result, the proposed WMV model achieves the best performance against state-of-the-art models (see Section III).

Figure 4 illustrates the intricacies and distinctions between EfficientNet-B0, B5, and B7. The EfficientNet-B0 is the simplest among the EfficientNet family. From EfficientNet-B0 to EfficientNet-B7, the depth, width, resolution, and model size increase incrementally, but accuracy continually improves. EfficientNet-B7 is known as the most accurate and best-performing model. The EfficientNet architecture comprises seven blocks of networks, each of which is composed of several modules. As seen in Figure 4, the input module and the output module of the EfficientNet architecture are identical in all three models, with the input module handling the start of the experiment and the output modules acting as output layers. Following that, each of them has seven blocks. The number of sub-blocks/modules in such blocks varies from EfficientNet-B0 to EfficientNet-B7. The first and final blocks of the EfficientNet-B0 architecture are made up of three modules, while the EfficientNet-B5 and EfficientNet-B7 architectures are made up of six modules with varying degrees of repetition. EfficientNet-B0 has a total of 237 layers, while EfficientNet-B7 has an aggregate of 813 layers. These differences lead to different numbers of parameters and FLOPs for the three models as shown in Table 2.

TABLE 2. Number of parameters and FLOPs for EfficientNet-B0, B5, and B7 according to the observations of Russakovsky *et al.* [24].

Model	ImageNet	Ours	FLOPs
EfficientNet-B0	5.3 Millions	4 Millions	0.39 Billions
EfficientNet-B5	30 Millions	28.5 Millions	9.9 Billions
EfficientNet-B7	66 Millions	64 Millions	37 Billions

Deep learning training processes, as they are commonly known, may be quite expensive in terms of time, computational power, and limited GPU availability [25]–[27]. These difficulties and constraints frequently deter practitioners from implementing deep learning techniques. Mitigating these bottlenecks will significantly improve the use of deep learning in real-world applications, particularly in real-time industrial machine malfunction detection. Russakovsky *et al.* [24] discover that the number of FLOPs is dependent on the architecture of the deep neural network and the amount of input data. The number of FLOPs limits the execution time required by the neural network [28], [29]. Additionally, the execution time increases roughly linearly with the number of FLOPs performed. EfficientNet has been proven to have the optimum performance based on their FLOPs. Therefore, the EfficientNet backbone network is the optimum choice for the proposed weighted ensemble model.

D. WEIGHTED VOTING STRATEGY

The use of mixed ensemble learning algorithms to mitigate the detrimental effects of overfitting and initialization sensitivity on learning performance by combining individual

learners from heterogeneous and homogeneous models has attracted considerable interest [30], [31]. The classification output of each ensemble member is assigned a different weight. This approach has shown great success in a variety of areas, such as precision medicine [32], spatial prediction [33], and mortality prediction [34], [35]. The contribution of each ensemble member is weighted by a coefficient that indicates the trust or expected performance of the model. Weight values are between 0 and 1 and are expressed as a percentage, such that the total of the weights of all individual members equals one. To identify an appropriate weight for each model, the Dirichlet distribution process [36] is employed rather than the more traditional exhaustive grid search approach. The Dirichlet distribution and its compound variant, the Dirichlet-multinomial, are two of the most basic models for proportional data. Formally, the following terms are defined, where m models are trained on a dataset of n samples, and the outputs of the models are pooled to calculate the final classification (i.e., prediction) of any instance x:

$$y(x) = \sum_{j=1}^{m} \beta_j(x) h_j(x) \tag{1}$$

Here weights β_j correspond to probabilities:

$$\sum_{j=1}^{m} \beta_j(x) = 1 \quad \text{and} \quad 0 < \beta_j(x) \leq 1 \tag{2}$$

The weight optimizations are carried out via the *fit*() function, which uses a validation dataset and the Dirichlet distribution to do a greedy randomized search to optimize weights. For a more extensive description of the Dirichlet distribution, see [37]. The Dirichlet distribution is a model of how proportions vary. An experiment with possible outcomes 0, 1 and having respective probabilities of $\beta_1, \beta_2, \ldots, \beta_m$ is considered, and a probability distribution on the vector $(\beta_1, \beta_2, \ldots, \beta_m)$, $\sum_{j=1}^{m} \beta_j(x) = 1$ is assumed. Because $\sum_{j=1}^{m} \beta_j(x) = 1$, it cannot be defined a density on $\beta_1, \beta_2, \ldots, \beta_m$, but it may be defined one on $\beta_1, \beta_2, \ldots, \beta_{m-1}$ and then take $\beta_m = 1 - \sum_{j=1}^{m-1} \beta_j(x)$. Therefore, β denotes a random vector whose elements sum to 1, so that p_m represents the proportion of model m. Under the Dirichlet model with parameter vector α, the probability density at β is:

$$p(\beta) \sim Dir_{(\alpha_1,\ldots,\alpha_m)} = \frac{\Gamma(\sum_m \alpha_m)}{\Pi_k \Gamma(\alpha_m)} \Pi_k p_m^{\alpha_m - 1} \tag{3}$$

where

$$\sum_k p_m = 1 \quad \text{and} \quad p_m > 0 \tag{4}$$

The parameters α can be estimated from a validation dataset $D = \{\beta_1, \ldots, \beta_n\}$. The maximum-likelihood estimate of α maximizes $p(D|\alpha) = \Pi_i p(\beta_i|\alpha)$.

III. RESULTS AND DISCUSSION

A. EXPERIMENTAL DESIGN AND PARAMETER SELECTION

All experiments are conducted using the Keras deep learning library [38]. Experiments are conducted using the proposed ensemble model from three different individual models based on EfficientNet. There exists a total of 14,920 industrial machine audio recordings of pumps and valves allocated to the EfficientNet-B0, B5, and B7. The dataset is split into 90% for training and 10% for random testing. In the training set, the training files are split into 80% for training (10,830 files) and 20% (2,708 files) for validation. The code of the proposed method is implemented using Python 3.9.4,[1] and the deep convolutional neural network structures are established based on the Keras[2] framework with a TensorFlow backend.[3] Each training is carried out on a single GPU NVIDIA GeForce RTX 3060Ti on a Windows workstation with CUDA 11.3. It has taken nine days of training and twelve minutes and sixteen seconds of testing to complete the process. The configuration used for the individual models is as follows. The proposed network is trained using the SGD optimizer [39] with a learning rate of 1e-6 for EfficientNet-B0 and EfficientNet-B5 and 5e-5 for EfficientNet-B7, a batch size of twelve image samples for EfficientNet-B0 and EfficientNet-B5 and a batch size of eight image samples for EfficientNet-B7, and the proposed network is trained for 1,000 epochs. The best model configuration as evaluated by the loss of the test set is chosen.

B. RESULTS

A real-life machinery sound dataset from MIMII [19] is used to evaluate the effectiveness of the proposed WMV. The dataset used in this research consists of recordings from two industrial machine types (i.e., pumps and valves) under normal and anomalous operations. The anomalous recordings exhibit various scenarios, such as leakage, clogging, voltage change, a loose belt, and poor lubrication. In addition, background noise recorded in real-world factories is added to each recording according to a certain SNR. This experiment uses sounds with an SNR of −6 dB, 0 dB, and 6 dB. Therefore, the proposed WMV model represents a practical use-case scenario incorporating real-world complexity and unpreventable situations, such as when microphones capture background noises in a factory environment. Each single-channel recording is 10 seconds long and has a sampling rate of 16 kHz. The most complex datasets containing sound files of pumps and valves are selected for this experiment. Four different datasets (i.e., ID 00, ID 02, ID 04, and ID 06) are selected for each machine.

Figure 5 summarizes the performance using the selected dataset. MV and stacking are selected as ensemble models for comparison with the WMV. According to the experimental results shown in Figure 5, weighing the weighted ensemble

[1] https://www.python.org/downloads/release/python-394/
[2] https://github.com/fchollet/keras.
[3] https://www.tensorflow.org/

as a whole outperforms the best individual model and other ensemble models (i.e., MV and Stacking). As the experiment demonstrates, models with higher scores also get weighted with greater weights.

A number of measurement metrics are used in the evaluation of the experimental results to measure the performance of the proposed model as well as various individual classifiers and ensemble models. As in a classic measurement, the confusion matrix is utilized to perform classic measurements by using four variables: *TP*, *FP*, *TN*, and *FN*.

- *TP*: the number of predictions where the classifier correctly classifies the positive class as positive.
- *FP*: the number of predictions where the classifier incorrectly classifies the negative class as positive.
- *TN*: the number of predictions where the classifier correctly classifies the negative class as negative.
- *FN*: the number of predictions where the classifier incorrectly classifies the positive class as negative.

The accuracy and AUC-ROC are selected, because this experiment involves balanced datasets between the normal and anomalous datasets as well as considers both positive and negative predictions. Accuracy is the ideal choice, because it embodies simplicity and ease of interpretation. Accuracy is defined as follows.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

The results reveal that the model demonstrating the most effective performance on pumps (i.e., WMV) has an accuracy of 0.97447 and the model demonstrating the most effective performance on valves (i.e., WMV) has an accuracy of 0.98211. Additionally, an AUC metric indicates how well the model distinguishes between two conditions (i.e., normal and abnormal conditions) or how well the model performs. This metric has a value between 0 and 1. A model with 100% incorrect prediction has an AUC of 0, whereas a model with 100% correct prediction has an AUC of 1. The proposed model (i.e., WMV) achieves an AUC of 0.99810 for pumps and 0.99930 for valves. The experimental results are also supported by the F1. The F1 is the harmonic mean of precision and recall (see Equation 6). The proposed model (i.e., WMV) achieves an F1 of 0.97562 for pumps and 0.98289 for valves.

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + 0.5 \times (FP + FN)} \tag{6}$$

Furthermore, Kappa and MCC [40] are considered to measure performance. The MCC score analyzes the classification result and the ground truth as two sets and takes into account TP and FN to compute a correlation coefficient that ranges between -1 (complete disagreement) and 1 (complete agreement). A value of zero shows that the classification does not correlate with the ground truth. The Kappa and MCC scores can be expressed as follows.

$$\text{Kappa} = \frac{2 \times (TP \times TN - FN \times FP)}{(TP + FP) \times (FP + TN) + (TP + FN) \times (TN + FN)} \tag{7}$$

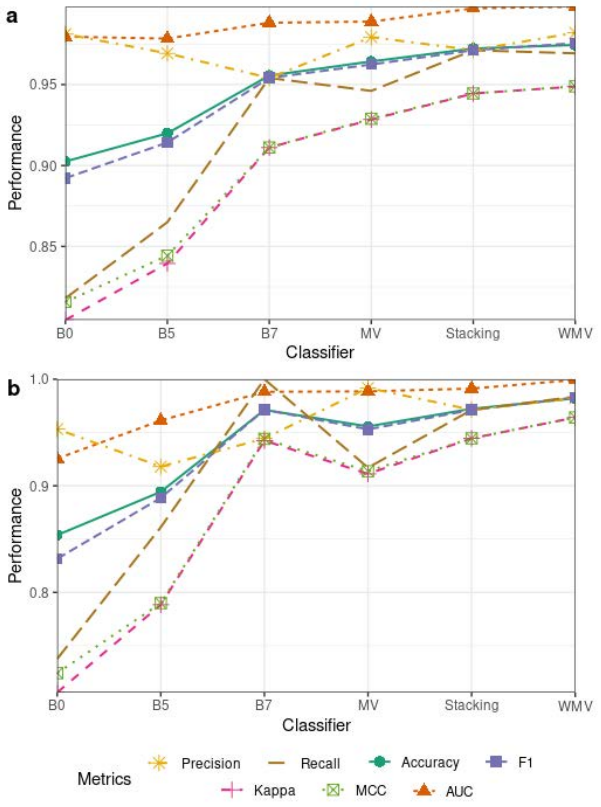

FIGURE 5. Performance comparison between various individual classifiers and ensemble models (i.e., MV, stacking, and WMV) over two distinct datasets (i.e., (a) pumps and (b) valves).

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FP) \times (TN + FN)}} \tag{8}$$

As shown in Figure 5 and Table 3, the improvement in the Kappa and MCC show that the proposed model achieves the highest agreement with the ground truth compared to the other individual classifiers and ensemble models. The Kappa score ranges between 0-1, while the MCC ranges between -1 to 1, where a score of 1 represents a perfect agreement between the observation and the results. In probabilistic measures, the best possible score is 1. The results of the proposed model (i.e., WMV) illustrate Kappa and MCC scores of 0.94884 and 0.94893 for pumps and 0.96414 and 0.96414 for valves, respectively.

C. DISCUSSION

A statistical significance test is run using the Quade test to check whether performance differences among classification models are significant. Table 4 shows that the proposed model (i.e., WMV) is significantly different (p-value $<$ 0.05) from the other classification models (i.e., EfficientNet-B0, EfficientNet-B5, EfficientNet-B7, MV, and stacking) with respect to accuracy, AUC, Kappa, F1, and MCC metrics. The result of the Friedman rank also indicates the superiority of WMV, where the proposed model maintains

TABLE 3. Precision, Recall, Accuracy, F1, Kappa, MCC, and AUC score between individual classifiers and various ensemble models.

Model	Machine	Precision	Recall	Accuracy	F1	Kappa	MCC	AUC
EfficientNet-B0	Pump	0.98122	0.81800	0.90251	0.89221	0.80455	0.81566	0.97960
	Valve	0.95321	0.73755	0.85377	0.83163	0.70606	0.72425	0.92550
EfficientNet-B5	Pump	0.96929	0.86497	0.91988	0.91416	0.83951	0.84429	0.97840
	Valve	0.91800	0.86123	0.89439	0.88871	0.78842	0.78987	0.96130
EfficientNet-B7	Pump	0.95409	0.95409	0.95560	0.95409	0.91110	0.91110	0.98820
	Valve	0.94444	1.00000	0.97120	0.97143	0.94244	0.94400	0.98820
MV	Pump	0.97934	0.94611	0.96429	0.96244	0.92841	0.92892	0.98890
	Valve	0.99184	0.91704	0.95569	0.95297	0.91120	0.91371	0.98860
Stacking	Pump	0.97126	0.97119	0.97229	0.97122	0.94449	0.94450	0.99710
	Valve	0.97126	0.97119	0.97229	0.97122	0.94449	0.94449	0.99120
WMV	Pump	**0.98198**	**0.96935**	**0.97447**	**0.97562**	**0.94884**	**0.94893**	**0.99810**
	Valve	**0.98221**	**0.98357**	**0.98211**	**0.98289**	**0.96414**	**0.96414**	**0.99930**

TABLE 4. Average Friedman rank and Quade omnibus test of the proposed model (i.e., WMV) in comparison with other models.

Performance metrics	Average Friedman rank	Quade p-value
Accuracy	1.0	$*$
Precision	1.5	NS
Recall	2.0	NS
AUC	1.0	$*$
Kappa	1.0	$*$
F1	1.0	$*$
MCC	1.0	$*$

$*$ significant at $p < 0.05$. NS: not significant.

TABLE 5. Performance improvement (%) of the proposed model (i.e., WMV) over individual classification models across different datasets and metrics.

Metrics	Classifier A	Classifier B	Pump	Valve
Accuracy	WMV	B0	7.97	15.03
		B5	5.93	9.81
		B7	1.98	1.12
Precision	WMV	B0	0.08	3.04
		B5	1.31	6.99
		B7	2.92	4.00
Recall	WMV	B0	18.50	33.36
		B5	12.07	14.20
		B7	1.60	-1.64
AUC	WMV	B0	1.89	7.97
		B5	2.01	3.95
		B7	4.14	1.12
Kappa	WMV	B0	17.93	36.55
		B5	13.02	22.29
		B7	4.14	2.30
F1	WMV	B0	9.35	18.19
		B5	6.72	10.60
		B7	2.26	1.18
MCC	WMV	B0	16.34	33.12
		B5	12.39	22.06
		B7	4.15	2.13

TABLE 6. Performance comparison between the proposed WMV model and existing models in terms of AUC metric.

Model	Approach	Year	AUC	
			Pump	Valve
Purohit et al. [19]	AE	2019	0.7233	0.6125
Perez-Castanos et al. [43]	Convolutional AE	2020	0.8261	0.8319
Koizumi et al. [44]	SPIDERnet	2020	0.952	-
Thoidis et al. [7]	Convolutional AE	2020	0.883	0.753
Kapka [45]	Class-conditioned AE	2020	0.8827	0.8456
Ribeiro et al. [6]	Convolutional AE	2020	0.7306	0.7883
Talmoudi and Hirata [42]	Real-time data tracker	2020	0.7436	0.8965
Zhang et al. [46]	MDF-FCN	2020	0.9317	0.7493
Van Truong et al. [47]	U-Net	2021	0.860	0.849
Nguyen et al. [48]	Convolutional VAE	2021	0.8127	0.5982
Müller et al. [5]	Transfer learning	2021	0.7702	0.6853
Thoidis et al. [41]	RawdNet	2021	0.903	0.967
This work	**Weighted ensemble**	**2021**	**0.9981**	**0.9993**

TABLE 7. Performance comparison between the proposed WMV model and existing models in terms of F1 metric.

Model	Approach	Year	F1	
			Pump	Valve
Koizumi et al. [44]	SPIDERnet	2020	0.880	-
This work	**Weighted ensemble**	**2021**	**0.9756**	**0.9829**

the top position in rank order across all performance metrics. In addition, Table 5 summarizes the relative improvement that Classifier A (i.e., WMV) produces over Classifier B (i.e., base classification models) on two acoustic signal datasets. Overall, WMV offers a performance enhancement (with at least 0.08%) over the base classification models in all performance metrics. The maximum enhancement (i.e., 36.55%) is achieved on valves in terms of the Kappa metric.

The proposed WMV model is compared with several existing models, such as MIMII AE, DCASE AE, DenseAE, VAE, and several other recently used models. In unsupervised and semi-supervised AE, the best performing models are compared with the proposed model. This can be seen in Tables 6 and 7. Table 6 shows the value of an AUC achieved by several existing models and the proposed model. The performance results of the proposed model achieve significant improvement compared to existing models, with an AUC of 0.7233 to a near perfect performance of an AUC of 0.9981 for pumps. The AUC for valves also improves from 0.6125 into 0.9993. This illustrates that the proposed model significantly outperforms all existing models for both tested machines (i.e., pumps and valves). The malfunction detection

TABLE 8. Summary of previous works (in chronological order) that consider similar acoustic signal datasets.

Ref.	Year	Method	Input feature	Performance measures	Remarks
[19]	2019	AE	Mel spectrogram	AUC	Performance results, i.e., valve: 0.67, pump: 0.81, fan: 0.94; and slide rail: 0.90
[49]	2020	IDNN	Mel spectrogram	AUC	Achieve 27% improvement against non-stationary machine sounds.
[46]	2020	FCN	MDF	AUC	Performance results, i.e., valve: 0.7362, pump: 0.9996, fan: 0.9978, and slide rail: 0.9646
[44]	2020	One-shot learning	Spectrogram	AUC and F1 scores	Utilizing a neural network-based feature extractor and attention mechanism.
[47]	2021	Fully connected U-Net	Mixed features, i.e., MFCC, chroma feature, Mel spectrogram, spectral contrast, and Tonnetz	AUC and pAUC	The proposed model achieve 83.38% AUC and 64.51% pAUC on average overall machine types.
[50]	2021	CNN	Mel spectrogram	Precision, sensitivity, and accuracy	The designed model has only 1-3 convolutional layers and gives higher accuracy than AlexNet.
[48]	2021	Convolutional VAE	Mel spectrogram	AUC and pAUC	The use of fully-connected models yields better performance in comparison with convolutional models.
[5]	2021	ResNet, Gaussian mixture models, and One-class SVM	Mel spectrogram	AUC	The use pre-trained model on the task of image classification.
[51]	2021	SVM and MLP	Mel spectrogram	AUC	The proposed approach improves the performance by up to 39.5% compared with the baselines.
[41]	2021	RawdNet	Mel spectrogram	AUC	The proposed method is the fusion of supervised feature learning and unsupervised deep one-class neural network

in terms of the performance of an AUC increases by up to 9.51% and 3.23% for the pumps and valves, respectively, compared to the newest model developed this year by Thoidis *et al.* [41]. This improvement is also supported by the F1 value of pumps compared to a previous model that increased malfunction detection by 9.56%.

In this experiment, it is shown that the proposed model is effective with both pump and valve machine types due to balanced AUC between pumps and valves (less than 1%), while other existing models, such as SPIDERnet, AE by Purohit *et al.* [19], and MDF-FCN, are more suited for pumps than valves. Talmoudi and Hirata [42], Ribeiro *et al.* [6] and RawdNet by Thoidis *et al.* [41] are more suited for valves. These are shown due to their performance of an AUC that differs from 3.7% up to 21.5% for pumps compared to valves. In general, most models performs better for pumps than valves. This finding indicates that malfunction detection for valves is more challenging compared to pumps, because valves have non-stationary signals, while pumps have stationary signals. Pumps' stationary signals have a constant time period, frequency and spectral content, while valves' non-stationary signals do not have these fundamental assumptions. Therefore, the proposed model demonstrates its robustness to process both stationary and non-stationary signals.

IV. RELATED WORKS

Industrial machine malfunction detection using acoustic signals has garnered considerable interest in recent years. A sound anomaly might signal an issue or malfunction; therefore, early recognition of the anomaly can avert a variety of problems, including predictive monitoring of industrial equipment and auditory monitoring of highways [4], [52], [53]. Detecting anomalies include supervised, unsupervised, and semi-supervised approaches. Supervised anomaly detection is a form of a binary classification problem that needs the full dataset to be labeled ''normal'' or ''abnormal.'' An unsupervised method uses unlabeled data, as it will understand which data is normal and abnormal. Semi-supervised anomaly detection needs only ''normal'' data to be annotated so the model can learn which data are ''normal''.

A brief overview of current machine malfunction detectors that use acoustic sound signals, particularly the MIMII dataset, is presented as follows. Table 8 presents a summary of existing works that utilize the MIMII dataset for fault detection. The MIMII dataset is originally introduced by [19]. The authors then use AE for detecting the anomaly. Different AEs are constructed for each machine type and model ID using a training dataset consisting only of normal data. Nguyen and Huang [50] utilize CNN for machine fault detection based on sound signals analysis. The findings demonstrate that, despite their modest structure (e.g., 1–3 convolutional layers), the developed models achieve a high level of accuracy and surpassed ALexNet in most classification tests. When the training and testing data come from the same platform, a basic CNN model with a single convolution layer performs very well (96.51%–99.52%). An AE/VAE interpolation error is considered an anomaly score for anomalous sound detection, which overcomes the problem of predicting edge frames [49]. It is shown that

the suggested method achieves a 27% improvement in the conventional AUC score, particularly when it comes to non-stationary industrial machine noises.

An enhanced AE model named convolutional VAE is used for industrial machine anomalous sound detection [48]. The proposed technique is well-suited for increasingly complicated applications in which signal characteristics are distributed over temporal and frequency domains. Zhang *et al.* [46] solve an image encoding problem in time series representations by using MDF. Then, a FCN is applied to classify the MDF images in the MIMII dataset. Furthermore, Koizumi *et al.* [44] design SPIDERnet, a similarity function for one-shot anomaly detection in sounds. An embedded spatial similarity measure using a neural network and attention mechanisms that absorb time-frequency stretching are the two main components of SPIDERnet's detection system for anomalous sounds. Instead of utilizing deep AEs, Müller *et al.* [5] suggest neural networks that are pre-trained on the image classification task to extract features. Anomaly detection models are subsequently trained using these characteristics. The proposed approach outperforms convolutional AEs in noisy samples of four distinct industrial machines. Similarly, Van Truong *et al.* [47] utilize normal sound samples and a fully connected U-Net architecture to generate an acoustic representation that is the closest match to the input sound for comparison with an anomalous sound.

In contrast with the previous works, Gantert *et al.* [51] use spectral feature extraction techniques. In order to distinguish between normal and abnormal operations, spectral characteristics derived from industrial sounds are employed as inputs for supervised machine learning algorithms. Classification results (e.g., SVM) demonstrate a superior AUC score when compared to classical machine learning models applied to the same dataset while maintaining a small model parameter set for acceptable generalization. Finally, Thoidis *et al.* [41] provide a method for extracting discriminative embeddings from multi-channel raw audio for diverse machinery sounds. To identify problems on specific machines, a deep CNN learns machine embeddings and transfers them to a deep one-class neural network. The proposed model surpasses state-of-the-art audio-driven fault detection approaches and is much more resilient in noisy conditions.

V. CONCLUSION

The objective of this research is to propose a deep learning-based industrial machine malfunction detection model that uses acoustic signals to classify normal and abnormal conditions of industrial machines. In particular, a WMV model is proposed for improving classification performance.

This research consists of five primary stages. First, audio data that represent normal and abnormal operating conditions for industrial machines are selected. Then, various well-known methods for audio file augmentation are applied to increase the number of samples that have abnormal operating conditions, and a data conversion into spectrogram images is considered. A WMV model is then proposed to classify normal and abnormal operating conditions using audio data. Finally, weighting strategies are considered to improve classification performance.

Case studies involving audio data of two industrial machine types (i.e., pumps and valves) from an open dataset (i.e., MIMII) are used to verify the proposed WMV model. It is concluded that the proposed EfficientNet-based ensemble model provides better classification performance than individual classifiers (i.e., EfficientNet-B0, EfficientNet-B5, EfficientNet-B7) for all selected measures (i.e., precision, recall, accuracy, F1, Kappa, MCC, AUC) for both machine types (i.e., pumps and valves). In addition, the weighted ensemble applied for the proposed WMV model provides better classification performance than other ensemble models (i.e., MV and stacking) for all selected measures for both machine types as well. The experimental results indicate that the proposed WMV model can be used to detect industrial machine malfunctions using acoustic signals with high accuracy.

The authors will consider using audio data for other types of industrial machines (i.e., fans and slide rails) from the MIMII dataset as well as collecting and using other acoustic signals of industrial machines (e.g., washing machines). Future works also consider how the proposed WMV model that uses acoustic signals of industrial machines can be combined with existing industrial machine malfunction detection models that use different types of data (e.g., image data and tabular data) collected from industrial machines in order to improve malfunction detection performance.

Most existing sound analysis research has been focused on the detection power of algorithms rather than the understanding behind these detections. Therefore, in the future, the authors will seek to develop an industrial machine malfunction detection system capable of explaining the transparency of the features of normal and abnormal conditions of industrial machines. This feature is essential for artificial intelligence models due to the weight of human factors, where human lives are in danger because of malfunctioning industrial machines. There have been extensive noteworthy works and interests in the field of XAI covering different gaps in different domains [54]–[56], but there is still a struggle with the applicability of XAI that needs further research. It is important to better understanding artificial intelligence models' decisions and why they happen. This understanding will increase trust in artificial intelligence models with positive expected outcomes, enabling decisions regarding whether to fully rely on artificial intelligence models or consider human factors to address security attacks on artificial intelligence model-based techniques.

ACKNOWLEDGMENT

(Bayu Adhi Tama and Malinda Vania contributed equally to this work.)

REFERENCES

[1] C. Peeters, Q. Leclere, J. Antoni, and P. Lindahl, "Review and comparison of tacholess instantaneous speed estimation methods on experimental vibration data," *Mech. Syst. Signal Process.*, vol. 129, pp. 407–436, Aug. 2019.

[2] L. A. Gupta and D. Peroulis, "Wireless temperature sensor for condition monitoring of bearings operating through thick metal plates," *IEEE Sensors J.*, vol. 13, no. 6, pp. 2292–2298, Jun. 2013.

[3] R. F. Salikhov, Y. P. Makushev, G. N. Musagitova, L. U. Volkova, and R. S. Suleymanov, "Diagnosis of fuel equipment of diesel engines in oil-and-gas machinery and facilities," in *Proc. Nanosci. Nanotechnol., Nano-Scitech*, 2019, Art. no. 050009.

[4] T. K. Chan and C. S. Chin, "A comprehensive review of polyphonic sound event detection," *IEEE Access*, vol. 8, pp. 103339–103373, 2020.

[5] R. Müller, F. Ritz, S. Illium, and C. Linnhoff-Popien, "Acoustic anomaly detection for machine sounds based on image transfer learning," in *Proc. 13th Int. Conf. Agents Artif. Intell.*, 2021, pp. 1–4, doi: 10.5220/0010185800490056.

[6] A. Ribeiro, L. Miguel Matos, P. Jose Pereira, E. C. Nunes, A. L. Ferreira, P. Cortez, and A. Pilastri, "Deep dense and convolutional autoencoders for unsupervised anomaly detection in machine condition sounds," 2020, *arXiv:2006.10417*.

[7] I. Thoidis, M. Giouvanakis, and G. Papanikolaou, "Audio-based detection of malfunctioning machines using deep convolutional autoencoders," in *Proc. Audio Eng. Soc. Conv.*, 2020, p. 10330.

[8] B. A. Tama, S. Y. Lee, and S. Lee, "An overview of deep learning techniques for fault detection using vibration signal," in *Proc. Congr. Conf. Proc.*, vol. 261, no. 1, 2020, pp. 5701–5706.

[9] S. Yin, X. Li, H. Gao, and O. Kaynak, "Data-based techniques focused on modern industry: An overview," *IEEE Trans. Ind. Electron.*, vol. 62, no. 1, pp. 657–667, Jan. 2015.

[10] S. Jeschke, C. Brecher, T. Meisen, D. Özdemir, and T. Eschert, "Industrial Internet of Things and cyber manufacturing systems," in *Proc. Ind. Internet Things*. Cham, Switzerland: Springer, 2017, pp. 3–19.

[11] J. Lee, F. J. Wu, W. Y. Zhao, M. Ghaffari, L. X. Liao, and D. Siegel, "Prognostics and health management design for rotary machinery systems-reviews, methodology and applications," *Mech. Syst. Signal Process.*, vol. 42, nos. 1–2, pp. 314–334, Jan. 2014.

[12] V. Stojanovic, N. Nedic, D. Prsic, L. Dubonjic, and V. Djordjevic, "Application of cuckoo search algorithm to constrained control problem of a parallel robot platform," *Int. J. Adv. Manuf. Technol.*, vol. 87, nos. 9–12, pp. 2497–2507, Dec. 2016.

[13] N. Nediá and V. Stojanovié, "A nature inspired optimal control of pneumatic-driven parallel robot platform," *Proc. Inst. Mech. Eng., C, J. Mech. Eng. Sci.*, vol. 231, no. 1, pp. 59–71, Jan. 2017.

[14] G. Kim, J. G. Choi, M. Ku, H. Cho, and S. Lim, "A multimodal deep learning-based fault detection model for a plastic injection molding process," *IEEE Access*, vol. 9, pp. 132455–132467, 2021.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[16] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[17] K. H. Sun, H. Huh, B. A. Tama, S. Y. Lee, J. H. Jung, and S. Lee, "Vision-based fault diagnostics using explainable deep learning with class activation maps," *IEEE Access*, vol. 8, pp. 129169–129179, 2020.

[18] G. W. Song, B. A. Tama, J. Park, J. Y. Hwang, J. Bang, S. J. Park, and S. Lee, "Temperature control optimization in a steel-making continuous casting process using a multimodal deep learning approach," *steel Res. Int.*, vol. 90, no. 12, Dec. 2019, Art. no. 1900321.

[19] H. Purohit, R. Tanabe, T. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," in *Proc. Detection Classification Acoustic Scenes Events Workshop (DCASE)*, 2019, pp. 209–213.

[20] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-SNE," *J. of Mach. Learn. Res.*, vol. 9, pp. 2579–2605, May 2008.

[21] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf.*, 2015, pp. 18–25.

[22] J. G. Choi, C. W. Kong, G. Kim, and S. Lim, "Car crash detection using ensemble deep learning and multimodal data from dashboard cameras," *Expert Syst. Appl.*, vol. 183, Nov. 2021, Art. no. 115400.

[23] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*.

[24] O. Russakovsky, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[25] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, May 2019.

[26] S. Pang, C. Pang, L. Zhao, Y. Chen, Z. Su, Y. Zhou, M. Huang, W. Yang, H. Lu, and Q. Feng, "SpineParseNet: Spine parsing for volumetric MR image by a two-stage segmentation framework with semantic image representation," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 262–273, Jan. 2021.

[27] M. Vania and D. Lee, "Intervertebral disc instance segmentation using a multistage optimization mask-RCNN (MOM-RCNN)," *J. Comput. Des. Eng.*, vol. 8, no. 4, pp. 1023–1036, Jun. 2021, doi: 10.1093/jcde/qwab030.

[28] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, and S. Bates, "In-datacenter performance analysis of a tensor processing unit," *ACM SIGARCH Comput. Archit. News*, vol. 45, no. 2, pp. 1–12, 2017, doi: 10.1145/3140659.3080246.

[29] D. Justus, J. Brennan, S. Bonner, and A. Mcgough, "Predicting the computational cost of deep learning models," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2018, pp. 3873–3882.

[30] Y. Yang and J. Jiang, "Adaptive bi-weighting toward automatic initialization and model selection for HMM-based hybrid meta-clustering ensembles," *IEEE Trans. Cybern.*, vol. 49, no. 5, pp. 1657–1668, 2019.

[31] C. Ju, A. Bibaut, and M. V. D. Laan, "The relative performance of ensemble methods with deep convolutional neural networks for image classification," *J. Appl. Statist.*, vol. 45, no. 15, pp. 2800–2818, 2017.

[32] A. R. Luedtke and M. J. van der Laan, "Super-learning of an optimal dynamic treatment rule," *Int. J. Biostatist.*, vol. 12, no. 1, pp. 305–332, May 2016.

[33] M. M. Davies and M. J. van der Laan, "Optimal spatial prediction using ensemble machine learning," *Int. J. Biostatist.*, vol. 12, no. 1, pp. 179–201, May 2016.

[34] A. Chambaz, W. Zheng, and M. J. Van Der Laan. (2016). *Data-Adaptive Inference of the Optimal Treatment Rule and its Mean Reward the Masked Bandit*. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01301297

[35] R. Pirracchio, M. L. Petersen, M. Carone, M. R. Rigon, S. Chevret, and M. J. van der Laan, "Mortality prediction in intensive care units with the super ICU learner algorithm (SICULA): A population-based study," *Lancet Respiratory Med.*, vol. 3, no. 1, pp. 42–52, Jan. 2015.

[36] W. Roth and F. Pernkopf, "Bayesian neural networks with weight sharing using Dirichlet processes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 246–252, Jan. 2018.

[37] N. Balakrishnan and E. Hashorva, "Scale mixtures of Kotz-Dirichlet distributions," *J. Multivariate Anal.*, vol. 113, pp. 48–58, Oct. 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0047259X11001710

[38] F. Chollet. (2015). *Keras*. [Online]. Available: https://github.com/fchollet/keras

[39] M. Hardt, B. Recht, and Y. Singer, "Train faster, generalize better: Stability of stochastic gradient descent," in *Proc. 33rd Int. Conf. Mach. Learn.*, New York, NY, USA, vol. 48, Jun. 2016, pp. 1225–1234.

[40] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, Jan. 2020, doi: 10.1186/s12864-019-6413-7.

[41] I. Thoidis, M. Giouvanakis, and G. Papanikolaou, "Semi-supervised machine condition monitoring by learning deep discriminative audio features," *Electronics*, vol. 10, no. 20, p. 2471, 2021.

[42] S. Talmoudi and Y. Hirata, "An unsupervised big data visualization-based scheme for anomalous sound detection of facilities," *Res. Square*, vol. 4, pp. 1–43, Dec. 2020.

[43] S. Perez-Castanos, J. Naranjo-Alcazar, P. Zuccarello, and M. Cobos, "Anomalous sound detection using unsupervised and semi-supervised autoencoders and gammatone audio representation," 2020, *arXiv:2006.15321*.

[44] Y. Koizumi, M. Yasuda, S. Murata, S. Saito, H. Uematsu, and N. Harada, "SPIDERNet: Attention network for one-shot anomaly detection in sounds," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 281–285.

[45] S. Kapka, "ID-conditioned auto-encoder for unsupervised anomaly detection," 2020, *arXiv:2007.05314*.

[46] Y. Zhang, F. Gan, and X. Chen, "Motif difference field: An effective image-based time series classification and applications in machine malfunction detection," in *Proc. IEEE 4th Conf. Energy Internet Energy Syst. Integr. (EI2)*, Oct. 2020, pp. 3079–3083.
[47] H. Van Truong, N. C. Hieu, P. N. Giao, and N. X. Phong, "Unsupervised detection of anomalous sound for machine condition monitoring using fully connected U-Net," *J. ICT Res. Appl.*, vol. 15, no. 1, pp. 41–55, Jun. 2021.
[48] M.-H. Nguyen, D.-Q. Nguyen, D.-Q. Nguyen, C.-N. Pham, D. Bui, and H.-D. Han, "Deep convolutional variational autoencoder for anomalous sound detection," in *Proc. IEEE 8th Int. Conf. Commun. Electron. (ICCE)*, Jan. 2021, pp. 313–318.
[49] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 271–275.
[50] M. T. Nguyen and J. H. Huang, "Fault detection in water pumps based on sound analysis using a deep learning technique," *Proc. Inst. Mech. Eng., E, J. Process. Mech. Eng.*, vol. 4, Oct. 2021, Art. no. 09544089211039304.
[51] L. Gantert, M. Sammarco, M. Detyniecki, and M. E. M. Campista, "A supervised approach for corrective maintenance using spectral features from industrial sounds," in *Proc. IEEE 7th World Forum Internet Things (WF-IoT)*, New Orleans, LO, USA, vol. 14, 2021, pp. 1–5.
[52] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 1, pp. 279–288, Jan. 2016.
[53] Y. Kawaguchi and T. Endo, "How can we detect anomalies from subsampled audio signals?" in *Proc. IEEE 27th Int. Workshop Mach. Learn. for Signal Process. (MLSP)*, Sep. 2017, pp. 1–6.
[54] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
[55] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.
[56] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2021. [Online]. Available: https://www.mdpi.com/1099-4300/23/1/18

BAYU ADHI TAMA received the Ph.D. degree from Pukyong National University, Republic of Korea, in 2018. He is currently a Senior Research Fellow with the Data Science Group, Institute for Basic Science (IBS), Republic of Korea. Previously, he worked as a Postdoctoral Researcher with the Pohang University of Science and Technology and the Ulsan National Institute of Science and Technology, Republic of Korea, in 2018 and 2020, respectively. His research interests include data-driven computational approaches for healthcare, manufacturing, cybersecurity, and cultural analytics.

MALINDA VANIA received the joint Ph.D. degree from the University of Science and Technology (UST) and the Korea Institute of Science and Technology (KIST), Republic of Korea. She is currently working as a Postdoctoral Researcher with the Unstructured Data Mining and Machine Learning Laboratory, Department of Industrial Engineering, Ulsan National Institute of Science and Technology (UNIST), Republic of Korea. Previously, she worked as an Assistant Researcher with the Center for Healthcare Robotics, Korea Institute of Science and Technology, Seoul, Republic of Korea. Her research interests include applied deep learning for healthcare, medical image processing, image registration, image segmentation, image synthesis, explainable artificial intelligence, and sounds analysis.

ILJUNG KIM received the Ph.D. degree in management information systems (MIS) from Hanyang University, Republic of Korea, in 2018. He worked for the Ministry of Foreign Affairs and Trade (MOFAT) of Republic of Korea as the IT Manager and then worked for the SAF, U.S. government chaptered non-profit public organization, as the Chief Researcher. Since 2018, he has been a Professor and the Head of the Manufacturing AI Big data Centre, Korea Advanced Institute of Science and Technology (KAIST). His research interests include high-tech industry policy, cognitive science, and artificial intelligence in manufacturing industry.

SUNGHOON LIM received the B.S. and M.S. degrees in industrial engineering from KAIST, Republic of Korea, in 2005 and 2009, respectively, and the Ph.D. degree in industrial engineering from The Pennsylvania State University, University Park, PA, USA, in 2018. Since 2018, he has been an Assistant Professor with the Department of Industrial Engineering, Ulsan National Institute of Science and Technology (UNIST), Republic of Korea, where he has been the Head of the Institute for the 4th Industrial Revolution, since 2021. His research interests include machine learning/deep learning, industrial artificial intelligence, and smart manufacturing.