Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2022) 30: 2373 – 2384
© TÜBİTAK
doi:10.55730/1300-0632.3944

Noncontact machinery operation status monitoring system with gated recurrent unit model

Jason Jing Wei LIM[1,*], **Boon Yaik OOI**[1], **Wai Kong LEE**[2],
Teik Boon TAN[1], **Soung Yue LIEW**[3]

[1]Department of Computer Science, University Tunku Abdul Rahman, Kampar, Malaysia
[2]Department of Computer Engineering, Gachon University, Seongnam, South Korea
[3]Department of Computer and Communication Technology, University Tunku Abdul Rahman, Kampar, Malaysia

Received: 24.03.2022 • **Accepted/Published Online:** 13.09.2022 • **Final Version:** 28.09.2022

Abstract: In manufacturing industry, assembly line monitoring provides statistical information about overall performance and reliability of the legacy machines, ensuring that the machines give maximum yield output. However, most legacy machines lack internet connectivity and advanced functionality, increasing the difficulty for tracking task. Therefore, this work seeks to introduce a noncontact acoustic method to track machines rather than the mainstream vibrational approach. In order to provide accurate tracking of the daily machine operation for our machine tracking system, we consider scenario of background noises such as environmental sounds from multiple sources as well as neighbouring machine's sound. Thus, several neural networks are employed to recognize the machine status accurately. The objective of our work is to investigate the effect of machine types and states on recognition performance of neural network models under extremely noisy environments as well as to demonstrate the possibility of recognizing the sound on edge device. The main contribution of this article is the proposal of lightweight recurrent and convolutional-based models for machine sound recognition. The experimental results of our extensive testing included with multiple types of machines and background noises show that the proposed system with gated recurrent unit model has the best recognition accuracy of F1 score 0.913 with standard uncertainty of 0.026 with decent inference speed on edge device.

Key words: Internet-of-things, legacy machine monitoring, operation status tracking, sound recognition

1. Introduction

Since the introduction of artificial intelligence and internet-of-things (IoT) to create the smart factory concept in Industrial Revolution (IR) 4.0, many private sectors have started to make transition towards the transformation. The gradual changes give rise to the newer generation of advanced machines with self-monitoring capabilities. Unfortunately, small and medium enterprises could not afford the inexpensive machines while containing abundance of legacy machines that are without network connectivity and advanced functionality, stunting the revolutionary growth. In manufacturing industry, assembly line monitoring ensures optimum yield output. Tracking operation states of legacy machines can then give insight on the overall performance and reliability of the machines statistically, which motivates the production of this legacy machines monitoring work.

Our former operation status monitoring work had successfully demonstrated the possibility of tracking machine via acoustic approach, replacing the conventional vibration approach to mitigate the problems of

[*]Correspondence: jwlim@utar.edu.my

vibration sensor [1]. Due to the choice of recognition method, the work also selected best feature extraction technique to accommodate with sound similarity matching. Unfortunately, similarity matching recognition method is inapplicable in real practice due to restrictions such as vulnerability to background noises, prerequisite of sound preprocessing and tuning before tracking two or more machines via single source of audio and the inevitability of constructing the correct sound template under influences of noisy environment. The weaknesses of the method heavily deteriorated the recognition accuracy, which inspires the subsequent work that proposed combining multilayer perceptron (MLP) method along with temporary wireless vibration sensor to overcome the problem of manual labelling [2]. The work also demonstrated the possibility of identifying two types of simultaneously operating machines via single recording source. However, the work focused more on identifying best MLP parameters on recognizing machine sound and lacked detail about the impact of multiple machine types and interchanging machine states towards recognition performance.

The accomplishment made by latest studies led to the conclusion that sound recognition method plays a significant role in accurately tracking machines' daily behaviors in machine tracking systems. Only MLP was considered in the past for tracking machines' states. Therefore, this work attempts to substitute MLP with two different branches of neural network models in sound recognition method. On the one hand, there is recurrent-based model specialized in dealing with time series dataset sequences. On the another hand, convolutional-based model capitalizes in handling with images data. From another point of view, sound is made up of time-series acoustic signal. That alone is sufficient for recurrent-based model to provide better performance over MLP because recurrent-based model can learn and remember historical information in time-series sequences unlike MLP. Moreover, the accumulation of sound signal can be compressed and transformed into a visual image called spectrogram which can fed into convolutional-based models as input data. Both types of models could provide even greater recognition performance in comparison to MLP.

The primary objective of this work is to assess the recognition performance of these recurrent-based and convolutional-based models against MLP for acoustic signals produced by the machines. The work also intends to ascertain the effect of machines' type on recognition performance as well as the effect of multiple operating states on recognition performance under presence of neighbouring machines and background noises. Additionally, the work investigates the possibility of recognizing sound on edge device for real-time monitoring use. The experiments conducted in this work will be tested with more variants of background noises in comparison to previous work to thoroughly examine the noise tolerance of the recognition models.

The paper presents work with a few contributions. Our work proposes a lightweight recurrent-based and convolutional-based neural network model that requires significantly less computational cost that can be adapted into the noncontact machinery tool acoustic tracking system for machine sound recognition application. The acoustic signature of a sound sample can be recognized without any complicated signal preprocessing or feature extraction techniques that incurs additional computational cost unlike any sound and vibration approach shown in past works. After the sound is successfully recognized by the system, it can be used by the manufacturing stakeholders in order to accurately identify the status of a machine and statistically study overall equipment effectiveness (OEE) of a legacy machine in construction industries.

The content of the paper is structured as follows. Section 2 covers related literature studies that comprise possible existing methods that can be used to recognize machine sounds. Section 3 details the methodology that includes the process workflow and experimental setup that is used to evaluate different types of sound recognition methods. Section 4 describes and analyses the experimental results. Section 5 concludes the work.

2. Related literature

Most machines emit vibration and acoustic signals during operation. Signal can be recognized using different signal recognition methods. The properties of most machine sound signals are stationary and some of them are periodic, but the sound recorded are likely filled with background noises, thus nonstationary in real practice. Thus, we refer to any signal recognition method in the literature that is usable with acoustic and vibration signal.

The first known method that was explored by past researchers is unsupervised learning. Unsupervised learning techniques usually refer to the thresholding or template matching technique [3]. The advantage of thresholding technique is no prior requirement of labelled data whilst its disadvantage is the susceptibility to background noises. Because of that, it is very difficult to construct a unique acoustic profile in a noisy environment. An extension from the thresholding-based technique is the acoustic fingerprinting with database matching. The Shazam-based technique [4] could be extended with the combination of image processing technique. The technique included the construction of a constellation map from fast Fourier transform (FFT) which constitutes numerous fingerprints which then can be stored in the database. Although the extended technique is typically used for music recognition, recent work authored by Siriphun and his colleagues [5] used the technique to classify drone types. However, the extended technique likewise possesses the same weaknesses in comparison to baseline technique. Furthermore, the extended technique requires maintenance of a fingerprint database that expends storage space, as well as relying on database searching and matching processes. The computational cost incurred by the processes increases as the amount of fingerprint steadily increases in the database, creating an obstacle to recognize sound in real time when the edge device has weak processing power. Alternate unsupervised learning approach is with clustering algorithms that can be used to group the captured audio into operation states respectively [6]. Other than clustering, there is also statistical model like Hidden Markov Model which forms a good pair with MFCC in recognizing speech [7] and also well versed in identifying underwater acoustic signals [8]. However, both clustering and statistical model approaches suffer from the presence of background noises too.

The second method is with supervised learning, which were used by most papers that monitor machine with vibration signal in the last few years. Most vibrational monitoring papers had demonstrated effective supervised learning techniques on different implementations such as troubleshooting bearing and engine fault conditions [9, 10] and estimating bearing's life expectancy [11, 12]. However, the common methodology detailed in these papers included complex signal preprocessing and feature manipulation processes which depends much on domain-related expertise knowledge due to the applied feature extraction techniques. The processes can either add to computational cost or lead to information loss in the signal, which deems those vibration-based supervised learning techniques unfitting for our work. Moreover, vibration sensor must be closely attached with target machine to obtain readings. Although there was a finding that included wireless workaround [13] for vibration sensor, the sensor can expire within a short period due to harsh industrial environment that harbors some operating machines. On top of that, there are not enough spaces to compensate for sensor retrofitting and wiring. Therefore, we dive into supervised learning techniques that are generally used for sound recognition such as support vector machine and random forest. They were used for drone and drill sound recognition, which is similar to our target [14, 15]. Additionally, the state-of-the-art sound recognition approach usually involves the use of neural network models. The advantages of using neural network models are the avoidance of expert knowledge requirement as well as feature manipulation processes and robustness to background noises. Despite all the advantages, the acoustic data must also be labelled and of excellent quality in order to provide distinguishable acoustic features for neural network models. Out of all types of neural network, recurrent neural network (RNN), especially long short-term memory (LSTM) and convolutional neural network (CNN), are mostly used for sound recognition. RNN can take account into past information while CNN is able to consider spatial information. Through vibration signal, Xiao et al. [16] used bidirectional LSTM to classify fault condition of rotating machines while Xu et al. [17] used CNN for bearing fault diagnosis. Zhao et al. [18] used both types of neural network for remaining useful life prediction (RUL) of engine. Similarly, Que et al. [19] used gated recurrent unit (GRU) for predicting bearing's RUL. Since neural networks are used more in recent years and offer better performance, our scope is limited on several neural networks-based techniques in this work.

3. Methodology

This section will be broken down into two parts. The experimental setup that is used to evaluate the performance of varying models is shown first. We later elaborate each process's workflow, parameter configuration, and specification details of the workstation used for model training.

3.1. Experimental setup

Prior IoT architecture [2] identified wireless vibration sensor node with battery supply as the medium for providing ground truth for acoustic data while main sound sensor node was used for recording acoustic dataset. To evaluate the feasibility of the proposed monitoring system, we create a series of experiments to test the neural network models on different types of machine sounds. We only focused more on recording long hours of acoustic data with a simplified setup. Thus, the vibration sensor is not included for our experimental setup but we rather use predefined class labels to annotate our dataset. The sound sensor used for recording is an external microphone board, Seeed Respeaker 2-Mics Pi HAT which is embedded upon Raspberry Pi 3A+, the edge device. The reason why the microphone board serves as the main choice is because the external board is low-cost and yet it provides portability with the edge device and great sound recording quality which is comparable to most high-end USB microphones found in the market. The device is then positioned 30 cm away from the sound source just so that the sound has significantly higher louder volume level as there are a lot of loud background sounds in a real-world environment. Additionally, the sound recorded is approximately ten times more than what is produced in a factory, which is 80 dB [20]. The position of the setup, especially the microphone board, is fixed throughout the sound recording. Figure 1 briefly describes the details of the experimental setup and process flows. Each process will be elaborated in the following subsections respectively.

3.2. Processing acoustic dataset

For our experiment, the mono channel sound recorded has 22.05 kHz sampling rate and 16-bit depth in the form of uncompressed WAV format. Multiple environmental noises are recorded and added into the training data in preprocessing step to enhance the generalization of the models to the noises. The process of adding multiple background noises and the noises overlapping each other in different time interval on the intended machine sound is done every time after long hour recording and splitting of an individual machine sound. The machine sound is only segmented into 3-s length of audio samples, which approximately 2.5 MB is produced per minute of recording. After that, the edge device will extricate the acoustical information using Mel Frequency Cepstral Coefficient (MFCC) feature extraction techniques. Compared to the wholly sound data, the extracted feature has far more reduced sizes. MFCC will return a 2-dimensional vector that has 1690 values, with 13 coefficients,

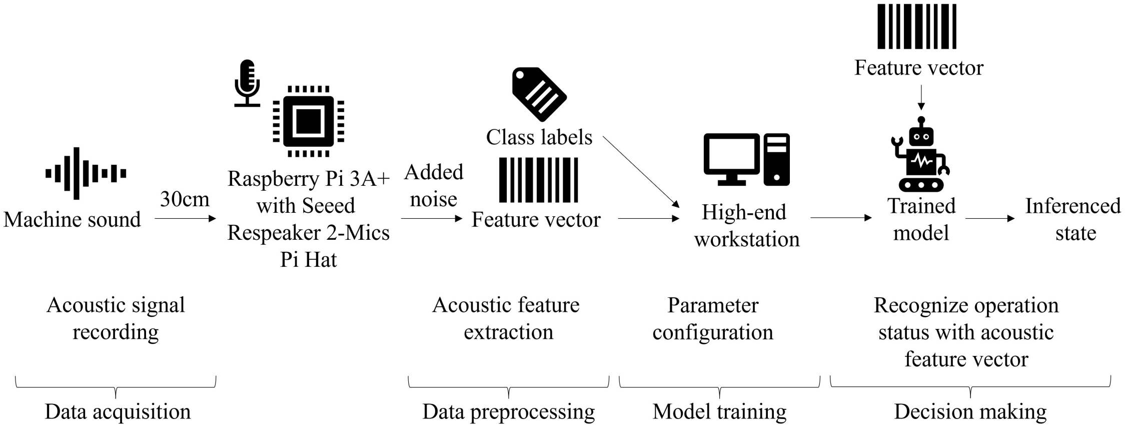

Figure 1. Experimental setup.

2048 samples (0.093s) of windows size, and 512 samples (0.023 s) of hop-length. The two-dimension vector is then flattened and fed into the neural network models as input data. For data labelling process, the numeric labels are converted into one-hot vector representations.

3.3. Mel frequency cepstral coefficient computation flow

Firstly, the incoming raw audio signal is segmented into multiple overlapping short frames. Hanning windows function is then applied onto the frames to reduce spectral leakage, which is the scenario depicting the energy misdistribution from certain frequency to another frequency. We subsequently used short-time Fourier transform for creating power spectrogram rather than FFT to assume periodicity of acoustic signal over short time. Subsequently, the following Equation 1 is used for the conversion of Hertz frequency to Mel frequency, which constructs the triangular frequency band of filters. Logarithm at base power of 10 will be computed over the filter bank output, creating Mel filter banks. Discrete cosine transform is performed next in order to reduce correlation between filter bank coefficients. Finally, after mean normalization, the output will contain the selected number of coefficients, becoming MFCC feature vector representation. This further reduces the original size of 3-s audio chunk with 22.05 kHz sample rate (2.52 MB) to approximately 136 kB per minute of recording. The important flows for MFCC computation are noted as shown in Figure 2.

$$m = 2595log(1 + f/700). \tag{1}$$

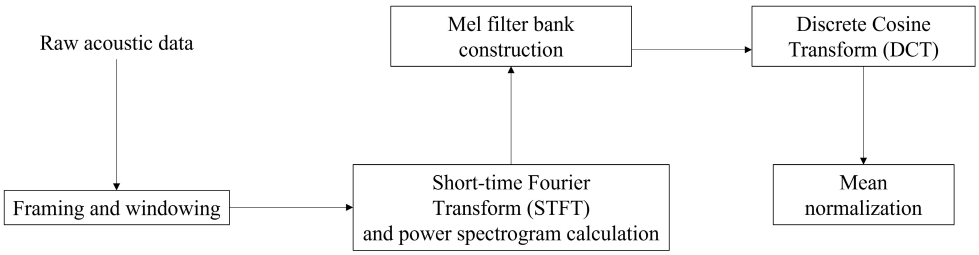
Raw acoustic data
Framing and windowing
Short-time Fourier Transform (STFT) and power spectrogram calculation
Mel filter bank construction
Discrete Cosine Transform (DCT)
Mean normalization

Figure 2. MFCC computation steps.

3.4. Neural network model training and inferencing operation statuses

There are four different neural network models available for machine sound recognition, namely MLP, LSTM, GRU, and CNN. The models are trained using Tensorflow 2.6.0 via Keras backend on a high-end processor pairing with graphics card. The models are trained in the workstation with the following specifications: i7-8700 CPU, Nvidia RTX 2070 GPU, 16 GB ram memory and 2TB storage capacity. For each sound, all the models are tested separately for the sound. The individual model has almost the same parameters in order to fairly compete with each other. Each model has two hidden layers and 52 nodes in each layer in order to equally compete with each other in terms of architecture and not affected by differences in parameter. For all the models, softmax activation function in the output nodes as categorical cross entropy loss function is used. The only difference is the activation function in the hidden layers. For MLP and CNN, the activation function is ReLU while Tanh for recurrent-based models so that the recurrent-based models can take advantage of GPU acceleration to speed up the training duration. Additionally, for CNN, the hidden layer has the kernel size of 3. There are two choices of optimizer: stochastic gradient descent (SGD) and adam. For our objective, we intend to find out the best model, which is why we are using the most basic optimizer instead of the developed one. Due to the choice of using SGD that has low convergence rate, feature scaling is performed on the train data and then the test data in order to increase the convergence rate. After the models are being trained, the trained models are downloaded from the workstation. For ease of inference, the trained models are in Tensorflow Lite format and typically has a size of less than 3 MB in total for each model. Any of the future audio recorded will be similar to the training samples, they are cut into 3-s length audios and converted into MFCC feature representation with identical parameters as acoustic data that was used for model training.

4. Experimental results and discussions

The responsibility of better handling of operation status tracking task wholly depends on finding the better neural network models, which is the biggest influencing factor towards better recognition accuracy. There are a total of three experiments we are going to conduct in order to search for better models. Prior study demonstrated the possibility of tracking both simultaneously running machines but did not consider the element of background noise heavily. Therefore, in order to better evaluate the noise robustness for the trained models, the work attempts to use noise contaminated training data as input data as well as noise contaminated testing data throughout most of the experiments as the recording obtained from actual factory environment is very noisy. The first experiment identifies the recognition performance of different types of neural network models against varying types of machines as well as different noise conditions while the second experiment focuses on distinguishing operating states of individual machines. The experiment detailed in Section 4.1 is a binary classification problem which recognizes two operating states for a machine. For the subsequent experiment discussed in Section 4.2, the recognition task is converted into a multiclass classification problem, which drives the flexibility of the model to handle the task of handling multiple operating states.

4.1. Experiment A: determining suitable neural network models for machine status tracking system

For this experiment, there are 15 different types of machine sounds, which are angle grinder, bulldozer, diesel generator, diesel locomotive, rotary tool, drill, heater, heatgun, honda generator, humidifier, makita angle grinder, nebulizer, paper machine, shredder, and wood shredder. Therefore, there is a total of 20.5 h of sound recording data available for model training and testing. The ratio proportion of individual machine audio data is broken down as the following. The data is divided into approximately 51.2% training dataset and 26.8% testing dataset. Another 22.0% is taken out to validate the models. There are only two output nodes for each neural network model in this experiment because the model is trained to decide whether a machine is running or stopped. For each machine, there will be 1 h of audio training data is made up of a) one set of 15 min audio where both machines are running simultaneously, b) one set of 30 min audio where each of the machines running separately respectively, c) one set of 15 min without the sound of both machines. We also add multiple background noises into the training data as well as other overlapped machine sound to evaluate the viability of our sound recognition models.

On the other hand, we record a total of 22 min for the testing data. it contains a) one set of 4-min audio where both machines are running simultaneously, b) one set of 4-min audio where secondary machine is running, c) one set of 7 min where the primary machine is running, d) one set of 6 min without the sound of both machines, e) one set of 1 min of silence. Similar to what we had done to the training data, we add the testing data too with background noises but the added noises for testing data are unseen and independent of the noises in the training data. The details about how the audio of the machine pairs are combined are represented in Tables 1 and 2. The audio of silence, as well as the secondary machine, is added in test data to ensure that the primary machine sound is correctly recognized despite the presence of other environmental sound. Each model is reinitialized, trained, and evaluated with shuffled data for a total of 450 times; 30 times on each of the machine sounds.

Table 1. Composition of training data.

Training duration (min)					
0–5	5–10	10–15	15–20	20–25	25–30
30–35	35–40	40–45	45–50	50–55	55–60
Primary machine (0–30), X (30–60)					
X	BGN1				
X	BGN2	BGN2	X	BGN2	BGN2
X	X	BGN3	BGN3	X	X
X	X	X	BGM1	BGM1	X

*BGN1: Background noise 1 (Environmental sound 1); BGN2: Background noise 2 (Excavator 1); BGN3: Background noise 3 (Human talking voice 1); BGM1: Secondary machine 1

Table 2. Composition of test data.

Testing duration (mins)										
0–1	1–2	2–3	3–4	4–5	5–6	6–7	7–8	8–9	9–10	10–11
11–12	12–13	13–14	14–15	15–16	16–17	17–18	18–19	19–20	20–21	21–22
Primary machine (0–11), X (11–22)										
X	BGN4									
X	BGN5	BGN5	X	BGN5	BGN5	X	X	X	X	X
X	X	BGN6	BGN6	X	X	BGN6	X	BGN6	X	X
X	X	X	BGM2	BGM2	X	X	BGM2	X	BGM2	X
X	X	X	X	X	BGN7	X	X	BGN7	BGN7	BGN7

*BGN4: Background noise 4 (Environmental sound 2); BGN5: Background noise 5 (Excavator 2); BGN6: Background noise 6 (Human talking voice 2); BGM2: Secondary machine 2; BGN7: Raining; X: Silent

Table 3 shows the recognition performance for individual machine sounds. As the table shows, all the machine sounds can be recognized by various model types correctly. Bulldozer, however, shows better performance on recurrent-based models such as LSTM and GRU. The average F1 score of each model types is then summarized into the bottom part of Table 3 which shows the overall results in monitoring operation status of the machines via the proposed sound recognition approach. Due to the total number of machine sounds being tested for each model, we can only take account of the model performance by using the F1 score. It is being summarized up in the form of average, minimum, and maximum, which the latter shows the lowest and highest recognition performance boundaries of an individual model can achieve. The last two columns show precision and recall, respectively, and they are calculated with Equations 2–4, respectively. The definitions of true-positive (TP), false-positive (FP), true-negative (TN), false-negative (FN) are provided as follows. TP indicates the running state of a machine identified correctly, FP indicates the running state of a machine misidentified, TN indicates the stopped state of a machine identified correctly, FN indicates the stopped state of a machine misidentified. In terms of recognition performance, GRU performs the best among the models. The lowest boundary of F1 score for all the models are the F1 score owned by drill due to the outdoor environmental noises, except for MLP, which has bulldozer machine as lowest boundary F1 score.

Table 3. Overall recognition performance of different neural network models on test data.

Machine sound types	Model types			
	MLP [2]	LSTM [16]	GRU [19]	CNN [17]
Angle grinder	0.995	0.978	0.984	0.984
Bulldozer	*0.675*	0.802	0.806	*0.715*
Diesel generator	0.983	0.989	0.988	0.993
Rotary tool	*1.000*	*1.000*	*1.000*	*1.000*
Drill	0.738	*0.719*	*0.753*	0.735
Heater	0.905	0.892	0.921	0.906
Heatgun	0.999	0.992	0.980	0.993
Honda generator	0.823	0.851	0.844	0.852
Humidifier	0.780	0.800	0.801	0.795
Makita angle grinder	*1.000*	*1.000*	0.998	*1.000*
Nebulizer	0.798	0.863	0.943	0.895
Paper machine	*1.000*	0.997	0.999	*1.000*
Shredder	0.718	0.750	0.776	0.742
Wood shredder	0.974	0.971	0.978	0.973
Average F1 score	0.888	0.902	0.913	0.901
Standard uncertainty	0.018	0.034	0.026	0.029
Precision	0.958	0.954	0.958	0.957
Recall	0.867	0.885	0.897	0.885

Italic font means minimum and maximum f1-scores of individual machines for the specific model.

$$f1score = 2 * (precision * recall)/(precision + recall). \tag{2}$$

$$precision = truepositive/(truepositive + falsepositive). \tag{3}$$

$$recall = truepositive/(truepositive + falsenegative). \tag{4}$$

Next, we attempt to use the best performing neural network model, which is GRU model in order to evaluate the usability of the model in recognizing the individual machine sound correctly if there is more than one machine. The performance is evaluated based on accuracy in possible scenarios that consist of different combinations of the machine sounds. Nevertheless, the background noises present are identical to the evaluation for the machine sounds recorded in Table 4. The result shown in the result of Table 4 shows decent performance across GRU model, demonstrating the adaptability of our model in tracking target machine sound with the presence of one or more than one background machines as well as handling the silent scenario.

Table 5 specifies the average training and inference duration of all machines' sounds in each model. The training duration of the individual model is calculated by averaging up the time it takes to train each machine sound dataset of 42 min, while the inference duration of the individual model is calculated by averaging up the time it takes to infer each machine sound dataset of 22 min. Among the models, unexpectedly, the MLP model has the fastest training and inference duration since it has the simplest model structure.

Table 4. Recognition performance under combination of three machine sounds with GRU model.

Combination/machine	Angle grinder (A)	Rotary Tool (B)	Shredder (C)
Silent	66.70%	100.00%	81.67%
A	100.00%	100.00%	70.00%
B	100.00%	100.00%	61.67%
C	97.50%	97.50%	80.00%
BC	72.50%	100.00%	100.00%
AC	100.00%	100.00%	22.50%
AB	100.00%	100.00%	100.00%
ABC	100.00%	100.00%	100.00%
Total average accuracy (run)	100.00%	100.00%	73.13%
Total average accuracy (stop)	84.18%	99.38%	80.84%

Table 5. Training and inference duration of different models.

Models	Training duration		Inference duration (Workstation)		Inference duration (Edge device)	
	Duration (s)	Standard uncertainty	Duration (s)	Standard uncertainty	Duration (s)	Standard uncertainty
MLP [2]	**2.591**	**0.052**	**0.010**	**0.000**	**0.292**	**0.069**
LSTM [16]	18.174	0.315	1.765	0.006	17.121	0.884
GRU [19]	15.712	0.053	1.704	0.005	15.469	0.133
CNN [17]	3.203	0.043	0.032	0.000	0.822	0.196

*Bold font means best performance for the specific model.

4.2. Experiment B: handling multiple operation states of individual machine

Different machines have different numbers of operating states. Therefore, when the operation state of a machine is changed during daily operation, the machine will exhibit different behavior, and in turn, different sounds are produced. To evaluate the viability of the model in recognizing these behaviors, two machines with multiple operating states are selected with the presence of background noises and another machine running background. Like the previous experiment, each state is treated like different machine sounds but were trained together as a single model with different number of output nodes in a model. Therefore, if there are five different operating states, there are five different datasets dedicated for each operating state, which are then accumulated together to create one complete dataset for one model. Air circulator will have 2 h of sound data, and rotary tool will have 3 h of sound data to ensure that the proportion of the dataset is equally distributed based on the number of operating states, that is half hour of data for each state.

The result in Table 6 is again denoted by f1-score, which shows the performance evaluation against two different machine sounds with multiple operation states in the form of speed changing. Despite using different neural network models, air circulator has moderately low performance across three different states: low-, medium-, and high speed. This can be observed from the results shown in Table 6 that air circulator low-speed state is the only one that has many incorrect predictions, resulting in lower f1-score. This correlates with the statement mentioned previously in Section 3, whereby the sound sensor has limitation regarding sound volume level of the targeted machine state. It has the problem of recognising sound that has low amplitude which causes the background noises to dominate the audio source. Therefore, the recognition performance will drop tremendously. Additionally, the neural network models do not have much performance difference. The subsequent Table 7 also provides the similar observation as seen in Table 6 from the performance of individual machine sounds with every model. Nevertheless, GRU performs the best for recognising air circulator, whereas MLP performs the best for recognising rotary tool. The models are then utilized to produce the results in Table 8 that gives gradually lower recognition performance as the noises becoming louder which is indicated by decrease in signal-to-noise ratio (SNR). This leads to the conclusion that machine types is more of a primary factor rather than operating states, showing more distinct performance across different models.

Table 6. Performance comparison of different neural network models on different operating states/speed changes.

Machines' states	Models			
	MLP	LSTM	GRU	CNN
Air Circulator - Low Speed	0.372	0.368	0.381	**0.383**
Air Circulator - Medium Speed	0.598	**0.626**	0.604	0.615
Air Circulator - High Speed	0.611	**0.614**	0.604	0.607
Air Circulator - Stopped	0.872	0.862	**0.886**	0.872
Rotary Tool - Speed 1	0.893	0.864	0.879	**0.896**
Rotary Tool - Speed 2	0.913	0.909	0.914	**0.923**
Rotary Tool - Speed 3	**0.985**	0.953	**0.985**	0.980
Rotary Tool - Speed 4	**0.979**	0.945	0.970	0.956
Rotary Tool - Speed 5	**0.952**	0.882	0.951	0.931
Rotary Tool - Stopped	0.991	0.895	0.990	**0.998**

*Bold font means minimum and maximum f1-scores < tally with Table 3 of individual machines < tally with Table 3 for the specific model.

5. Conclusion and future work

The proposed work highlights the design of a scalable sound recognition model in IoT architecture which is capable of recognizing different types of machines' sound in factory. With simulation of real-life environmental scenarios: environmental noises mixed with machine sounds and targeted machine sound with overlapped neighbouring machine sounds in background, we tested out different types of models on multiple machine sounds in factory. The experimental result reveals that GRU model has the best performance in terms of recognition accuracy with moderate training time and overall fast inference time. It is also shown that even though the training data has extreme background noises as well as neighbouring machine sounds, the inference accuracy

Table 7. Overall performance comparison of machine operating states/speed changes across different models.

Models	Sounds	Average F1 score	Standard uncertainty
MLP [2]	Air circulator	0.613	0.011
	Rotary tool	0.952	0.010
LSTM [16]	Air circulator	0.617	0.017
	Rotary tool	0.908	0.043
GRU [19]	Air circulator	0.625	0.009
	Rotary tool	0.948	0.019
CNN [17]	Air circulator	0.619	0.010
	Rotary tool	0.947	0.014

Table 8. Performance comparison of machines' operating states on best performing models with difference in SNR.

Machine sounds	SNR (dB)				
	0	–5	–10	–15	–20
Air circulator (GRU)	0.625	0.377	0.300	0.317	0.312
Rotary tool (MLP)	0.952	0.834	0.623	0.406	0.272

is still performing at a promising level and suitable for real-time monitoring purpose. The only limitation is that if the volume level in sound recording is very low and the sound sensor lacks the proper sensitivity level to record the sound, the recognition accuracy significantly deteriorates as seen from the experimental result of distinguishing multiple operating states. There are two extension opportunities that stemmed from the current work in the future. It is observed that training in a high specification workstation is fast. Therefore, training process for model can also be done by the edge device, which further reduces the bandwidth consumption that requires downloading model and processing power that trains model from the cloud in order to recognize machine sounds. The proposed system is also shown to be competent for accurate and real time machine status monitoring task, however it is yet to be tested in actual factory environment, which we are keen to explore in our future work.

Acknowledgment

This research is supported by MOHE under the Fundamental Research Grant Scheme (FRGS) with project code: FRGS/1/2021/ICT08/UTAR/02/1.

References

[1] Ooi BY, Lim JJW, Lee WK, Shirmohammadi S. Non-intrusive operation status tracking for legacy machines via sound recognition. In: 2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC). IEEE; 2020. p. 1–6. doi:10.1109/I2MTC43012.2020.9129526

[2] Ooi BY, Lim JJW, Liew SY, Shirmohammadi S. Remote operation status tracking for manufacturing machines via sound recognition using IoT. In: 2022 IEEE International Instrumentation and Measurement Technology Conference (I2MTC). IEEE; 2022. p. 1–6. doi:10.1109/I2MTC48687.2022.9806481

[3] Li P, Wu Q, Wu C, Yuan C. A denoising method of frequency spectrum for recognition of dashboard sounds. In: 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). IEEE; 2017. p. 1–5. doi:10.1109/CISP-BMEI.2017.8302139

[4] Wang AL, Street KH. An industrial-strength audio search algorithm. International Conference on Music Information Retrieval ISMIR. 2003. http://www.ee.columbia.edu/~dpwe/papers/Wang03-shazam.pdf

[5] Siriphun N, Kashihara S, Fall D, Khurat A. Distinguishing drone types based on acoustic wave by IoT device. In: 2018 22nd International Computer Science and Engineering Conference (ICSEC). IEEE; 2018. p. 1–4. doi:10.1109/ICSEC.2018.8712755

[6] Shuyang Z, Heittola T, Virtanen T. Active learning for sound event classification by clustering unlabeled data. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2017. p. 751–755. doi:10.1109/ICASSP.2017.7952256

[7] Kuamr A, Dua M, Choudhary T. Continuous Hindi speech recognition using gaussian mixture HMM. In: 2014 IEEE Students' Conference on Electrical, Electronics and Computer Science. IEEE; 2014. p. 1–5. doi:10.1109/SCEECS.2014.6804519

[8] Diamant R, Kipnis D, Zorzi M. A clustering approach for the detection of acoustic/seismic signals of unknown structure. IEEE Transactions on Geoscience and Remote Sensing. 2018;56 (2):1017–1029. doi:10.1109/TGRS.2017.2758162

[9] Soualhi A, Medjaher K, Zerhouni N. Bearing health monitoring based on hilbert–huang transform, support vector machine, and regression. IEEE Transactions on Instrumentation and Measurement. 2015;64 (1):52–62. doi:10.1109/TIM.2014.2330494

[10] Tao J, Qin C, Li W, Liu C. Intelligent fault diagnosis of diesel engines via extreme gradient boosting and high-accuracy time-frequency information of vibration signals. Sensors. 2019;19 (15):3280. doi:10.3390/s19153280

[11] Mao W, He J, Zuo MJ. Predicting remaining useful life of rolling bearings based on deep feature representation and transfer learning. IEEE Transactions on Instrumentation and Measurement. 2020;69 (4):1594–1608. doi:10.1109/TIM.2019.2917735

[12] Yongjiao C, Wei D. Early diagnosis of processing faults based on machine online monitoring. In: 2016 Prognostics and System Health Management Conference (PHM-Chengdu). IEEE; 2016. p. 1–4. doi:10.1109/PHM.2016.7819956

[13] Goyal D, Dhami SS, Pabla BS. Non-contact fault diagnosis of bearings in machine learning environment. IEEE Sensors Journal. 2020;20(9):4816–4823. doi:10.1109/JSEN.2020.2964633

[14] Zakeri V, Hodgson AJ. Automatic identification of hard and soft bone tissues by analyzing drilling sounds. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2019;27 (2):404–414. doi:10.1109/TASLP.2018.2880336

[15] Anwar MZ, Kaleem Z, Jamalipour A. Machine learning inspired sound-based amateur drone detection for public safety applications. IEEE Transactions on Vehicular Technology. 2019;68 (3):2526–2534. doi:10.1109/TVT.2019.2893615

[16] Xiao L, Duan F, Tang J, Abbott D. A noise-boosted remaining useful life prediction method for rotating machines Under different conditions. IEEE Transactions on Instrumentation and Measurement. 2021;70:1–12. doi:10.1109/TIM.2021.3064810

[17] Xu G, Liu M, Jiang Z, Söffker D, Shen W. Bearing fault diagnosis method based on deep convolutional neural network and random forest ensemble learning. Sensors. 2019;19 (5):1088. doi:10.3390/s19051088

[18] Zhao C, Huang X, Li Y, Yousaf Iqbal M. A double-channel hybrid deep neural network based on CNN and BiLSTM for remaining useful life prediction. Sensors. 2020;20 (24):7109. doi:10.3390/s20247109

[19] Que Z, Jin X, Xu Z. Remaining useful life prediction for bearings based on a gated recurrent unit. IEEE Transactions on Instrumentation and Measurement. 2021;70:1–11. doi:10.1109/TIM.2021.3054025

[20] Ooi BY, Beh WL, Lee WK, Shirmohammadi S. A parameter-free vibration analysis solution for legacy manufacturing machines' operation tracking. IEEE Internet of Things Journal. 2020;7 (11):11092–11102. doi:10.1109/JIOT.2020.2994395