IDENTIFY MILITARY SOUNDS USING DEEP LEARNING

[1]Dr. Khalid M. Oqla Nahar, [2]Ali M. Ali Kharabsheh

[1]Department of Technology and Computer Sciences, Jordan

[2]Department of Technology and Computer Sciences, Jordan

E-mail: [1]khalids@yu.edu.jo, [2]alikharabsha12@gmail.com

ABSTRACT

The problem of classification of military sounds has not received attention in researchers before, The importance of this research lies in the camps when military raids occur by the enemies so that they are at a far distance and determine their direction using sound frequencies, as a result, we need to establish a sensitive model that relies on distinguishing sounds long distances. This also results in some serious security threats, that need soldiers alerts of a military raid and timely detection of the military raids to protect the security-sensitive institutions. In this paper, we propose a Convolutional Neural Network (CNN) for the detection and classification of Military sounds. To extract the necessary features from Military sounds, Mel frequency cepstral coefficients (MFCCs) feature extraction is implemented. Where the results show that the CNN model is superior in the performance and accuracy of 95.3% with testing data, It also shows that the Convolutional Neural Network (CNN) model used in this work gives very good results and can be used for classification Military sounds, especially with unseen data.

Keywords: *Convolutional Neural Network (CNN), Deep Learning, classification of military sounds,*

1. INTRODUCTION

The problem of the classification of the Military sounds hasn't received attention from the researchers in the past years. To date, signal processing and machine learning techniques haven't been applied to the problem, wavelet filterbanks [1], [2] and most recently Deep Neural Networks (DNN)[3], [4]. In particular, deep convolutional neural networks (CNN) [5] are, in principle, very well suited to the problem of sound classification: first, they are capable of capturing energy modulation patterns across time and frequency when applied to spectrogram-like inputs, which is an important trait for distinguishing between different, often noise-like, sounds such as engines and jackhammers [5]. Second, by using convolutional kernels (filters) with a small receptive field, the network should, in principle, be able to successfully learn and later identify spectro-temporal patterns that are representative of different sound classes even if part of the sound is masked (in time/frequency) by other sources (noise).

Deep neural networks, which have a high model capacity, are particularly dependent on the availability of large quantities of training data to learn a non-linear function from input to output that generalizes well and yields high classification accuracy on unseen data. A possible explanation for the limited exploration of CNNs and the difficulty to improve on simpler models is the relative scarcity of labeled data for sound classification.

In this paper, we present a deep Convolutional Neural Network (CNN) for the classification of the Military sounds e.g. weapons sounds, Sound of Soldiers marching (Regular and irregular marching), sound planes, and ...etc. We show that the proposed CNN architecture yields state-of-the-art performance for Military sounds classification.

2. RELATED WORKS

They [4] used convolutional neural networks in classifying short audio clips of environmental sounds on the dataset consists of 2000 short (5 seconds) environmental recording comprising 50 equally balanced classes of sound events in 5 major groups (animals, natural soundscapes and water sounds, human non-speech sounds, interior/domestic sounds, and exterior/urban noises).

They [6] proposed a new method for the audio classification to identify bird species. They used a convolutional neural network (CNN), which contains more than 33,000 registrations of 999 different types. Where their mean average accuracy was 0.686 when predicting the main types per unit

and scored 0.555 when using background types as additional prediction targets.

They [7] proposed applying a deep learning task to a robotic heart, that is, to recognize anomalies in the sounds of the heart. They describe an automated cardiac sound classification algorithm that combines the use of time-frequency thermal map representations with a deep convolutional neural network (CNN). Had been recording in a total of 4,430 records from 1,072 subjects, resulting in 30 hours of sound recordings of the heart. From this total dataset, 1,277 heart recordings of 308 people were removed for use as parked test data to assess challenge submissions. The overall score of 0.8399 was achieved using a single convolutional neural network.

They [8] applied two different approaches of using deep neural networks for cough detection, and a convolutional neural network and a recurrent neural network were implemented to address these problems, respectively. Where they were evaluating the performance of the two networks and compare them to other conventional approaches for identifying cough sounds. They produced an average of 40 cough sounds, yielding a total of 627 cough examples. Between the two, their convolutional network yields a higher specificity 92.7% whereas the recurrent attains a higher sensitivity of 87.7%.

They [9] present a novel machine learning-based method for heart sound classification, in the classification stage; each feature vector is classified into “normal”, “abnormal”. Their method relies on a robust feature representation generated by a wavelet-based deep convolutional neural network (CNN) of each cardiac cycle in the test recording, and support vector machine. In addition to the CNN-based features, their method incorporates physiological and spectral features to summarize the characteristics of the entire test recording, where the proposed method for them obtained a score, sensitivity, and specificity of 0.812, 0.848, and 0.776, respectively, on the hidden challenge testing set.

They [10] proposed a new musical instrument classification method using convolutional neural networks (CNNs) is presented, where the proposed classifier outperformed the baseline result from traditional handcrafted features and classifiers.

This studied [11] had two primary contributions: first, they propose a deep convolutional neural network architecture for environmental sound classification, second, they proposed the use of audio data augmentation for overcoming the problem of data scarcity and explore the influence of different augmentations on the performance of the proposed CNN architecture. Combined with data augmentation, the proposed model produces state-of-the-art results for environmental sound classification, where they used the Urban-Sound8K dataset [12]. The dataset is comprised of 8732 sound clips of up to 4 s in duration taken from field recordings, the accuracy ratio was 0.74.

They [13] propose the novel machine learning (ML) framework for the detection and classification of ADr sounds out of the various sounds like birds, airplanes, and thunderstorms in the noisy environment. And they used Support Vector Machine(SVM), where they achievement around 96.7% accuracy for ADr detection.

They [14] used the convolutional neural networks to classify the syllables of the sounds of flying motors consisting of 745 voices consisting of 9 major groups and they achieved 87% accuracy.

Table 1 Related Works

Ref	research subject	Algorithm	Dataset	Accuracy
[4]	- classifying short audio clips of environmental sounds	- Convolutional Neural Networks (CNNs).	- 2000 short (5 seconds) environmental recording	----------
[6]	- a new method for the audio classification to identify bird species	- Convolutional Neural Networks (CNNs).	- which contains more than 33,000 registrations of 999 different types.	0.686%
[7]	- applying a deep learning task to a robotic heart, that is, to recognize anomalies in the sounds of the heart.	- deep convolutional neural network (CNN)	- a total of 4,430 records from 1,072 subjects	0.8399%
[8]	- applied two different approaches to using deep neural networks for cough detection.	- Convolutional Neural Networks (CNNs). - Recurrent Neural Network (RNN).	- a total of 627 cough examples.	CNN= specificity 92.7%. RNN= sensitivity of 87.7%.
[9]	- heart sound classification, classified into “normal”, “abnormal”.	- Convolutional Neural Networks (CNNs). - Support Vector Machine(SVM).	----------	a score, sensitivity, and specificity of 0.812%, 0.848%, and 0.776%.
[10]	a new musical instrument classification method	- Convolutional Neural Networks (CNNs).	----------	where the proposed classifier outperformed the baseline result
[11]	- first, they propose a deep convolutional neural network architecture for environmental sound classification. - second, they proposed the use of audio data augmentation.	- Convolutional Neural Networks (CNNs).	- The dataset is comprised of 8732 sound clips of up to 4 s in duration	----------
[13]	the detection and classification of ADr sounds out of the various sounds like birds, aeroplanes, and thunderstorms.	Support Vector Machine(SVM)	----------	96.7%
[14]	classify the syllables of the sounds of flying motors.	- Convolutional Neural Networks (CNNs).	consisting of 745 voices consisting of 9 major groups	87%

3. METHODOLOGY

Our proposed methodology is consisting of steps the following:

1. Data Collection.
2. Data preprocessing.
3. Extract Features.
4. Training and Testing.

Figure1 graphical representation of our methodology.

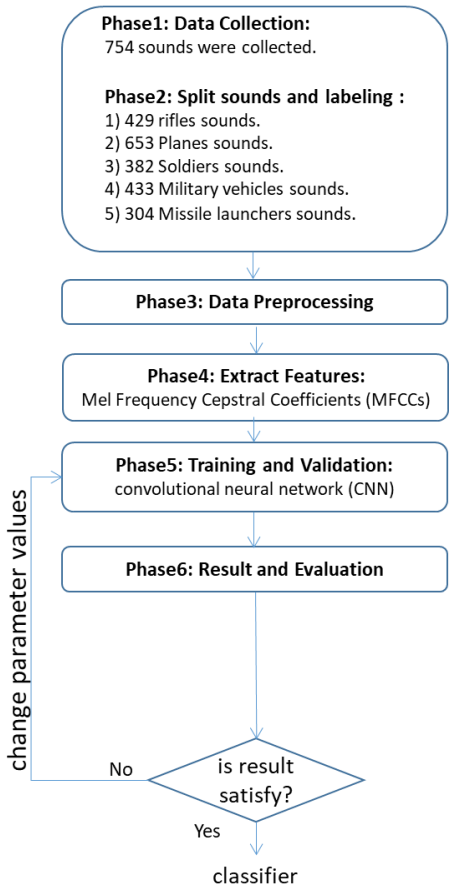

Figure 1 graphical representation of our methodology

3.1. Data Collection.

For this study, a total of 754 recordings from different subjects were obtained from multiple repositories, as described in [1] [2] [3] [4] [5] [6] [7] [8].

The data collected contained sounds of rifles, planes, marching soldiers (Regular and irregular marching), Military vehicles, and Missile launchers.

As the recordings were collected from multiple public repositories, there were variations in the recording conditions, sensors, sampling frequencies, and noise levels associated with each recording. In these recordings, military sound events were manually annotated to be used as a reference to compare the performance of the proposed military sound classification method.

As a result, a total of 2201 events were labeled as military-sounds, the military-sound events present in the recordings included sounds of rifles, planes, soldiers marching (Regular and irregular marching), Military vehicles, and Missile launchers. Table 2 shows the distribution of labels in the dataset.

Table 2 Distribution of label in the dataset

Labels	# of Recordings
rifles	429
planes	653
Soldiers marching	382
Military vehicles	433
Missile launchers	304
Total	2201

Figure2 shows the distribution of recordings according to the sum of seconds for each class.

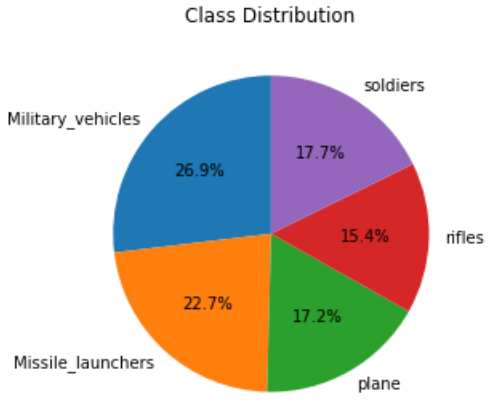

Figure 2 Distribution of recordings

[1] https://www.youtube.com/watch?v=ZJeOQ7BOkRI

[2] https://retired.sounddogs.com/results.asp?Type=1&categoryid=1005&subcategoryid=58

[3] https://www.youtube.com/watch?v=xX3MjZo4Ufg

[4] https://www.freesoundeffec[4]ts.com/free-sounds/gun-10081/

[5] http://soundbible.com/tags-gun.html

[6] https://www.fesliyanstudios.com/royalty-free-sound-effects-download/gun-shooting-300

[7] https://www.youtube.com/watch?v=YDNVI-f4CQQ

[8] https://www.youtube.com/watch?v=VHoOg4twnO0

For Deep Learning (DL), we split the dataset into testing, training, and validation by 30% from training data; Table 3 shows the split data.

Table 3 Split data.

Split Dataset	Training data	Testing data
rifles	294	135
planes	249	404
soldiers	202	188
Military vehicles	249	184
Missile launchers	210	94
Total	1200	1005

3.2. Data preprocessing.

Before being used for training (analysis), all the recordings were first preprocessed to minimize the variations as a result of different recording conditions in the following steps below:

1) Resampled all the recordings to 16000 Hz.
2) Subsequently, the amplitude of the recordings was scaled into a range between -1 and 1 without changing the distribution, such that important statistical parameters of the sounds were preserved.
3) Signals were then segmented into frames of 200ms width with a 50% overlap between consecutive segments.
4) Prior to the detection of military sounds, it can be useful to remove silent parts of the recordings to ensure that all further processing is performed only on parts of the signals containing a sound event. However, this requires the silence removal approach to be much simpler than the further processing stages. To achieve this, in this study, silence and low noise parts of the recordings were detected and removed using a simple threshold based on the mean of deviation in the previous samples. Thus, further processing was performed only when the average energy of a frame was above this threshold as in [15].

3.3. Extract Features

Many features can be extracted from the sound-wave like Linear Predictive Coding (LPC), Spectral Centroid, Spectral Rolloff, Mel-Frequency Cepstral Coefficients (MFCCs), and ..., etc. However, in this study, we will use the MFCCs features.

Mel Frequency Cepstral Coefficients (MFCCs) is one of the most important methods to extract a feature of an audio signal and is used majorly whenever working on audio signals. MFCCs of signals are a small set of features (usually about 10–20) which concisely describe the overall shape of a spectral envelope[9]. In Figure 3, an example of a spectrum for MFCC from our data.

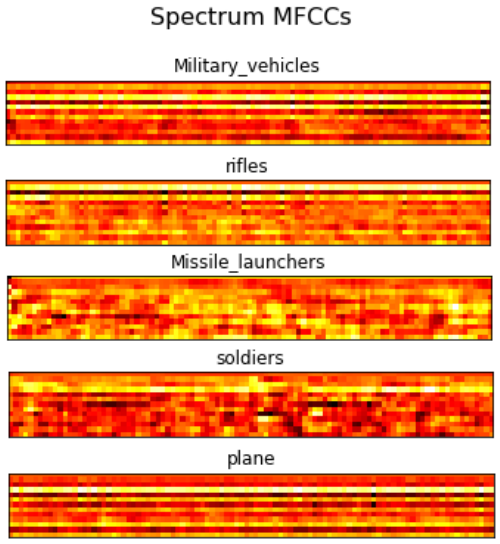

Figure 3 Spectrum MFCCs

3.4. Training and Testing.

Deep Learning (DL) is a part of the Machine Learning family, where learning can be supervised, unsupervised, and semi-supervised. DL uses several nonlinear processing layers, so the model depth will be referred to as the number of layers that the data is transformed through it.

Since DL can be implemented through several techniques, in our approach we have found that CNN is used as a trusted architecture in the field of speech recognition and sound detection, so CNN serves us well in the identification of military-sounds. In section 3.4.1, we show the detailed architecture of the CNN used in our method, and in section 3.4.2, we show phases of the detailed training and validation for each epoch.

[9] https://towardsdatascience.com/extract-features-of-music-75a3f9bc265d

3.4.1. Model Architecture

The detailed architecture of CNN is shown in Fig. 4. Activation function "RELU" is applied to each layer except for the last dense layer we applied "softmax". We use a sampling rate of 16 kHz, optimizer "Adam" with learning rate (LR=0.001), and lose function "categorical_crossentropy".

We apply four time-convolutional layers, in the first convolutional2D layer contain 16 filters with filter size 3*3, in the second convolutional2D layer contain 32 filters with filter size 3*3, in third convolutional2D layer contain 64 filters with filter size 3*3 and in fourth convolutional2D layer contain 128 filters with filter size 3*3. After that, we apply Max Pooling2D with pool size 2*2. After that, we apply 50% of the dropout. After that, we apply to flatten. Finally, we apply three time-Dense layers, in the first dense layer contain 128 neurons, in a second dense layer, contain 64 neurons, and in third dense layer contain 5 neurons.

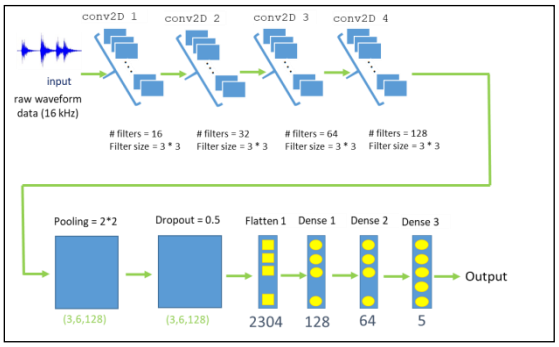

Figure 4 the detailed architecture of CNN

Table 4 contains all the parameters used in the DL models.

Table 4 the parameters used in the DL model

Model parameters	Model parameters Value
NUMBER DENSES	192
BATCH SIZE	64
NUMBER LAYERS	10
DROPOUT	0.5
OPTIMIZER	Adam
NB EPOCHS	40
ACTIVATION FUNCTION	RELU
LOSS FUNCTION	categorical_crossentropy

3.4.2. Training and Validation

Figure 5 screenshot of the training and validation in our model for the last 3 epochs, where a number of epochs equal 40 with the batch size equal to 64.

```
Epoch 37/40

76246/76246 [==============================] - 149s 2ms/step -
loss: 0.1408 - acc: 0.9482 - val_loss: 0.1342 - val_acc: 0.
9525

Epoch 00037: val_acc improved from 0.94917 to 0.95251, savi
ng model to models\conv.model

Epoch 38/40

76246/76246 [==============================] - 148s 2ms/step -
loss: 0.1370 - acc: 0.9504 - val_loss: 0.1416 - val_acc: 0.
9503

Epoch 00038: val_acc did not improve from 0.95251

Epoch 39/40

76246/76246 [==============================] - 140s 2ms/step -
loss: 0.1345 - acc: 0.9500 - val_loss: 0.1431 - val_acc: 0.
9482

Epoch 00039: val_acc did not improve from 0.95251

Epoch 40/40

76246/76246 [==============================] - 139s 2ms/step -
loss: 0.1319 - acc: 0.9511 - val_loss: 0.1504 - val_acc: 0.
9469

Epoch 00040: val_acc did not improve from 0.95251
```

Figure 5 training and validation in our model

4. MODEL EVALUATION

Evaluation of the model is an important part of any project, it shows if the model is good or not. One of the most used metrics is the accuracy score metric, it is a good metric to evaluate the model, but sometimes using the accuracy score metric alone is not enough, so we use other metrics such as F1 score, Precision, and Recall.

4.1. Precision

Precision is one of the known measures, and it is used to know the classifier's ability to return relevant instances only. The equation used to calculate it is Equation 1; the number of correct positive results is divided by the number of the positive results predicted by the algorithm.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}} \quad (1)$$

4.2. Recall

Recall (also known as sensitivity) is used to know the classifier's ability to identify all relevant instances. The equation used to calculate it is Equation 2; the number of correct positive results is divided by the number of all relevant samples.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}} \quad (2)$$

4.3. F-Measure

F-Measure is used to combine Precision and Recall into one measurement tool; it uses the harmonic mean to combine them. Equation 3 to calculate this measure.

$$F1 = 2 * \frac{1}{\frac{1}{\text{PRECISION}} + \frac{1}{\text{RECALL}}} \quad (3)$$

4.4. Accuracy

Accuracy is the most popular used performance measure, it is known as the ratio of correctly predicted observation to the total observations. Accuracy evaluates the model well only if the dataset was balanced. Since our data is not balanced, we used other measurement tools to evaluate the model, equation 4 to calculate this measure, which is the same as Equation 5.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad (4)$$

$$\text{Accuracy} = \frac{\text{TP + TN}}{\text{TP + TN + FP + FN}} \quad (5)$$

5. RESULTS AND DISCUSSION

In our experiments, According to the performance measures, the results showed that the CNN model gives very good results in solving the problem of military sound classification. CNN achieved an accuracy of 95.1% with training, achieved an accuracy of 95.2% with validation, and achieved an accuracy of 95.3% with testing (unseen data). Table 5 shows more details (Performance measures).

Table 5 Performance measures

Performance measures	Testing Accuracy
Precision	94%
Recall	97%
F-Measure	95%
Accuracy	95.3%

After several experiments using different parameters, higher accuracy achieved with a better score. We achieved an accuracy of 95.3% using CNN with unseen data.

One way to review and visualize the performance of DL models is by using plots, Figure 6 shows the training and validation accuracy for each epoch. From the plot and based on several experiments, we stopped training at epoch number 40 for the CNN model, because after these numbers the models start to overfit the data.

Figure 7 we see the training and validation loss for each epoch, it shows that validation and testing loss keep decreasing in the CNN model until we reach 40 epochs, after 40 epochs the validation loss starts to increase so we stopped training this model at epoch number 40.

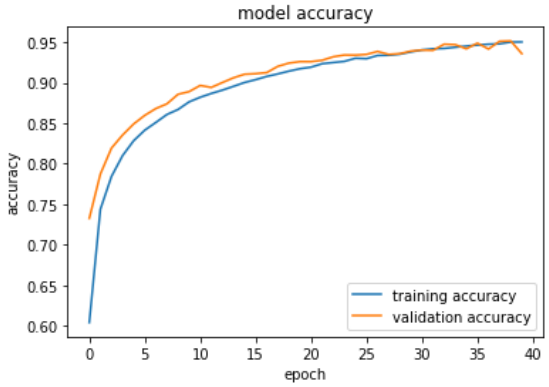

Figure 6 Training and Validation accuracy

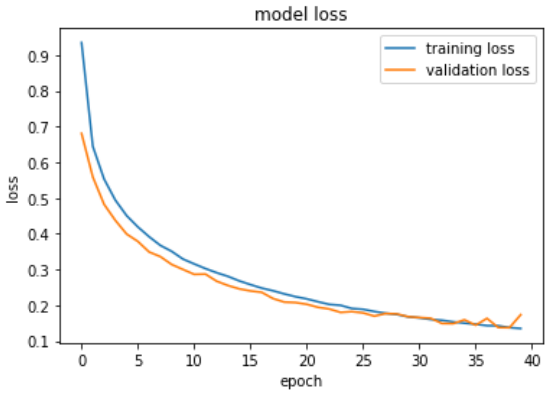

Figure 7 Training and Validation loss

A confusion matrix is a table that is used to describe the performance of a classifier on the test dataset. As shown in Figure 8 for the CNN model there are (47) wrong predictions and (958) right predictions; as we mentioned previously, the total of testing data was 1001 recordings.

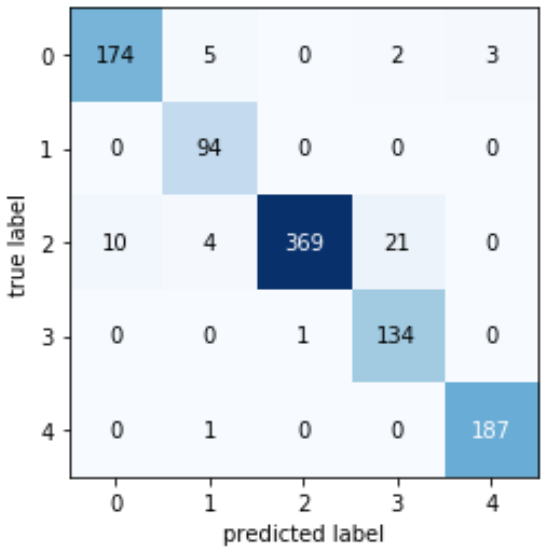

Figure 8 Confusion Matrix

As we see most of the classifications were centered on the diagonal, which means that the classification achieves a high degree of accuracy.

6. CONCLUSION

The importance of this research lies in the camps when military raids occur by the enemies so that it is at a far distance and determines their direction using sound frequencies; as a result, we need to establish a sensitive model that relies on distinguishing sounds long distances. This also results in some serious security threats, that need soldiers alerts of a military raid and timely detection of the military raids to protect the security-sensitive institutions.

Where the results showed that the CNN model is superior in the performance and accuracy of 95.3% with testing data, It also shows that the Convolutional Neural Network (CNN) model used in this work gives very good results and can be used for classification Military sounds, especially with unseen data.

REFERENCES:

[1] J. Salamon and J. P. Bello, "Feature learning with deep scattering for urban sound analysis," *2015 23rd Eur. Signal Process. Conf. EUSIPCO 2015*, no. June, pp. 724–728, 2015.

[2] J. T. Geiger and K. Helwani, "Improving event detection for audio surveillance using Gabor filterbank features," *2015 23rd Eur. Signal Process. Conf. EUSIPCO 2015*, pp. 714–718, 2015.

[3] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2015-Septe, 2015.

[4] K. J. Piczak, "2015 Ieee International Workshop on Machine Learning for Signal Processing Environmental Sound Classification With Convolutional Neural Networks," *IEEE Int. Work. Mach. Learn. signal Process. Boston, USA*, 2015.

[5] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "A B7CEDGF HIB7PRQTSUDGQICWVYX HIB edCdSISIXvg5r ` CdQTw XvefCdS," *proc. IEEE*, 1998.

[6] E. Sprengel, M. Jaggi, Y. Kilcher, and T. Hofmann, "Audio based bird species identification using deep learning techniques," *CEUR Workshop Proc.*, vol. 1609, pp. 547–559, 2016.

[7] J. Rubin, R. Abreu, A. Ganguli, S. Nelaturi, I. Matei, and K. Sricharan, "Recognizing abnormal heart sounds using deep learning," *CEUR Workshop Proc.*, vol. 1891, pp. 13–19, 2017.

[8] J. Amoh and K. Odame, "Cough Sounds," pp. 1–9, 2016.

[9] M. Tschannen, T. Kramer, G. Marti, M. Heinzmann, and T. Wiatowski, "Heart sound classification using deep structured features," *Comput. Cardiol. (2010).*, vol. 43, pp. 565–

568, 2016.

[10] T. Lee, Taejin. Park, “MUSICAL INSTRUMENT SOUND CLASSIFICATION WITH DEEP CONVOLUTIONAL NEURAL NETWORK USING FEATURE FUSION APPROACH Taejin Park and Taejin Lee Electronics and Telecommunications Research Institute (ETRI), Republic of Korea,” *Electron. Telecommun. Res. Inst. (ETRI), Repub. Korea*.

[11] J. Salamon and J. P. Bello, “Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification,” *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, 2017.

[12] J. Salamon, C. Jacoby, and J. P. Bello, “A dataset and taxonomy for urban sound research,” *MM 2014 - Proc. 2014 ACM Conf. Multimed.*, no. 1, pp. 1041–1044, 2014.

[13] M. Z. Anwar, Z. Kaleem, and A. Jamalipour, “Machine Learning Inspired Sound-Based Amateur Drone Detection for Public Safety Applications,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2526–2534, 2019.

[14] R. Girshick, “Fast R-CNN,” *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2015 International Conference on Computer Vision, ICCV 2015, pp. 1440–1448, 2015.

[15] R. X. Adhi Pramono, S. Anas Imtiaz, and E. Rodriguez-Villegas, “Automatic Identification of Cough Events from Acoustic Signals,” *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 217–220, 2019.