Sound Classification using Multilayer Neural Network

Manzar Iqbal, Muhammad Zakir Khan, Mumtaz Ali, NaimatUllah
City University of Science and Information Technology Peshawar, Pakistan.

Abstract: Human environment consist of a mix-up of different sounds having different frequencies and temporal structure. These sounds must be classified for better understanding. The multi-layer neural network is able to classify sounds by extracting different features of sound. A multi-layer neural network consists of 3 layers (input layer, hidden layer, and output layer). Spectrogram and time-frequency graphs are responsible to find the frequency of each sound which helps in the classifying of the sound base of their frequency and amplitude. The accuracy of the model is evaluated on a data set called UrbanSound8k which is consists of 8,733 libelled sounds divided into 10 different classes. The goal of this paper is to classify environmental sounds into many meaningful sounds. This approach can further be used in the different sound applications of daily life.

Keywords: *Multi-layer Neural Network, Spectrogram, Time-frequency graph, Classification.*

*Email: *7148@cityuniversity.edu.pk, zakir@cusit.edu.pk, mumtazali@cusit.edu.pk,*
7160@cityuniversity.edu.pk

1. Introduction

Classification of different sounds from an audio recording or an mp3 file has always been a problem. For example, different sounds from the environment like humans, animals, birds, traffic, wind, water, and music, etc. combines to form a mixture of sounds. Classification of sounds through machine learning techniques is an effective way to overcome this problem. In recent years, different researches have been made in the field of sound classification and recognition but due to the complexity of environmental sound, their performance is still far from the ideal level. In literature, different researches used different methodologies to classify sounds like Time Motif Series, Detection of Polyphonic sounds, classification of bird's sounds, environmental sounds, etc. but due to the speedy changes and complexity in the environment, these techniques are not well suited. Beside this, in the field of machine learning the domains like image recognition, text recognition, and speech recognition have gained exceptional success.

The focus of this paper is to classify different sounds and keep them in their appropriate category. To design an effective sound classification model and implementation, it must be in the notice that there are two main tasks to perform when classifying the sounds. First is the extraction of sound features and second is the classification. In this regard, the first thing is to extract different features of sounds, so we will be using different machine learning techniques to extract the features of sounds and classify them. For this purpose, we will be using open sources library Librosa in a python. It is a python package for music and audio analysis. It provides the building block necessary to create a sound information retrieval system. It provides several methods to extract different features from the sound clip. Secondly, for the classification of sounds, we will be using a two-level multi-layered neural network. These two tasks are needed to prepare the data (both features and labels) for CNN.
Convolutional Neural Network (CNN) is a classifier that classifies sounds. This is accomplished by taking raw audio information and converting it into a meaningful sound. Convolutional neural networks model has basically three-layer architecture, but with the change in domain its architecture differentiates significantly. The data set we are using is UrbanSound8k which is a labelled collection of 8,733 raw audio recordings suitable for benchmarking methods of sound classification. The data set consists of 4-seconds-long recordings organized semantically into 10 classes: air conditioner, car horn, children playing, dog barking, drilling, engine idling, gun shoot, jackhammer, siren and street music (with 800 examples per classes) loosely arranged into 10 major categories.
The main objective of the research is to bring out the accuracy in results from which the previous researches were deprived and will the answer to the question that "Can Convolutional Neural Networks be effectively used in classifying sound sources?" Sound classification can provide several benefits in different fields of life. In life-critical events, gunfire or shouting of a person can be detected using sound classification. In the forest, sounds of different species of birds can be identified.

2. Literature study

Jiuwen Cao et al describes in this paper that the previous knowledge of the noise recognition is based on convolutional acoustic features like MFCC and LPCC which are unable to describe all the characters of the noise recognition. Convolutional Neural Network with its most reliable features like the FBank algorithm is used in this paper for the effective extraction of features used in noise recognition. A combination of CNN and FBank algorithm provides the best experimental results [10]. The Authors of this paper Huimin Zhao et al describes that deep learning techniques are used for the better performance of Environmental Sound Classification which is consists of SoundNet and EnvNet networks. Sound features are extracted from both the networks and combined together for better performance. This paper also shows that the existing methods used for sound classification do not cover all the aspects as this paper improves the accuracy by 3.2% according to the proposed experimental results [6]. Zixing Zhang et al are the authors of this paper. In this paper, the Active Learning method is introduced for the classification of bird’s sounds of different species. Two methods of AL called Sparse-instance-based AL and Leastconfidence-score-based AL are used for the discrimination of classification model. The use of both the methodologies reduce the need for human annotation [2]. Emer C,akir and Giambattista Parascandolo proposed this research which is based on the detection of Polyphonic sound events using Convolutional Recurrent Neural Network. Polyphonic Sound is the simultaneous combination of two or more sounds. CRNN is the combination of CNN and RNN. Both the methods are used combined to overcome the weaknesses of the acoustic sound recognition applications. According to the experimental results CRNN shows considerably improved over other sound recognition methods [5]. Jort F. Gemmeke et al are the authors of this paper. The main focus of the research is to collect different types of data to form an Audio data set for the availability of data for upcoming researches in the field of audiology. For this purpose, different types of researches related to sound classification and recognition are studied are their sound data sets are collected to form a large scale audio set which is consists of approximately all sound events [3]. Karol J. Piczak describes in this paper that the Environmental sounds (everyday sounds which don not contain any music or speech) are classified using Convolutional Neural Network approaches. Gaussian mixture model, SVM, and MFCC are used to extract sound features. This paper is the answer to the question that can CNN be used to classify and recognize different sounds [7]. Justin Salamon et al are the authors of this paper. The focus of the paper is to extract unsupervised features of sounds for classification. For this purpose k-mean algorithms are used for feature learning from audio signals. This paper also focuses on the temporal dynamics of the sound events as previous researches in the field of sound classification did not consider it that much important for classifying sounds [9]. Elsa Ferreira et al proposed this research. In this paper, the classification of sound occurs through the use of Time Series Motifs. TSM is the frequent repetition of the same sound in the audio. Discovering this sound and its frequency is the key to classifying the characteristics of different sounds. For this purpose, different types of sound extracting classifiers like MFCC, SVM, etc. are used. Time Series Analysis is the key to do so [4]. Jiaxing Ye et al are the authors of this paper which is about the classification of acoustic scenes using two important approaches called Sound Texture Analysis and Acoustic Event Analysis. One is used for studying the environmental nature of sound and the other is for background sounds. Experimental results show that this two-channel information is very much important for acoustic scene classification [1]. Jonathan Dennis et al in this research developed a method that minimizes degradation of the performance of sound classification in the presence of mismatched noises. This work is totally based on the visual information which is obtained through the spectrograms of the sounds. SVM is used, which is very useful in the classification accuracy of mismatched sounds [8].

3. Methodology

The methodology contains 3 major steps: Dataset, Feature extraction, and classification.

3.1 Dataset

The dataset used in this system is UrbanSound8k which is consists of 8,732 labelled sound recordings (4 seconds each) from ten classes. Dataset is divided into 10-folds and each fold contains more than 800 sound clips. Sound files are in .wav format. The dataset is trained and tested in the system using convolutional neural networks with TensorFlow and Libros to get accurate results from data.

3.2 Feature Extraction

Useful features are extracted from sound using Librosa which are listed below:

- Melspectrogram
- Mfcc
- Choroma-stft
- Spectal_contrast
- Tonnetz

Two basic function parse_audio_files and extract_features are used with librosa.load method to extract and return above mentioned features of sound.

3.3 Classification

For the training of classifier scikit-learn library is used. Other libraries can also be used but the main goal of this system the implementation of neural networks using TensorFlow for classification purposes. The dataset is trained and tested through CNN for the accuracy of results.

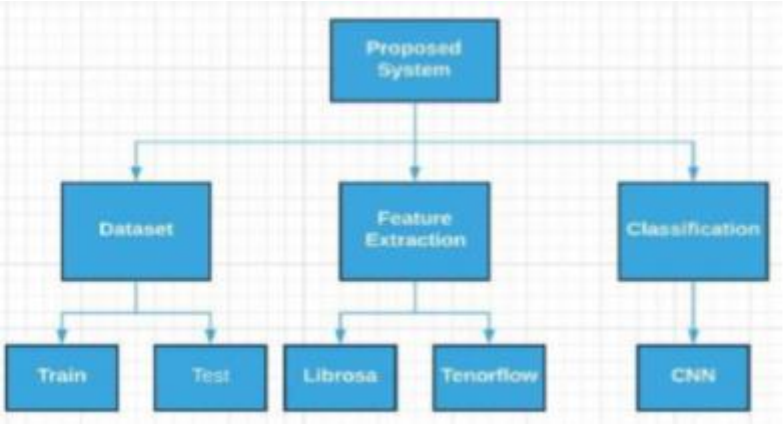


Figure 1. Methodology

3.4 Multi-layer Neural Network

A multi-layer neural network consists of units (neurons) arranged in layers, which convert an input vector into some output. Each unit takes an input, applies an (often nonlinear) function to it and then passes the output on to the next layer.

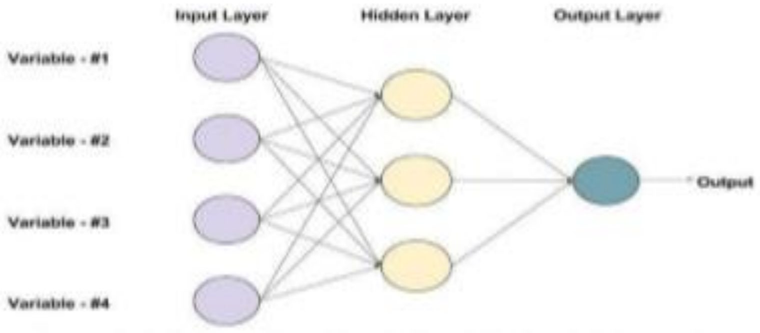
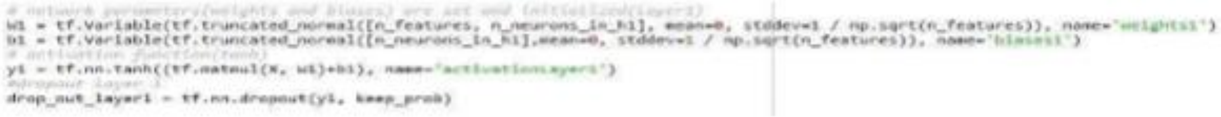


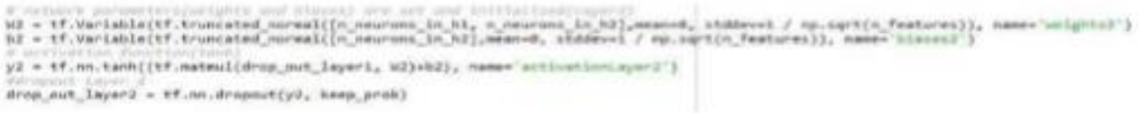
Figure 2. Multilayer Neural Network [11]

3.4.1 Layer1 (input layer)



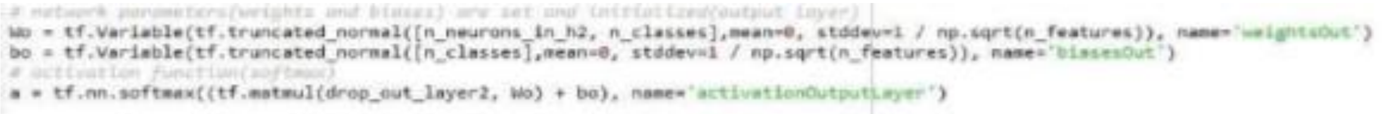
```
# network parameters(weights and biases) are set and initialized(Layer1)
W1 = tf.Variable(tf.truncated_normal([n_features, n_neurons_in_h1], mean=0, stddev=1 / np.sqrt(n_features)), name='weights1')
b1 = tf.Variable(tf.truncated_normal([n_neurons_in_h1],mean=0, stddev=1 / np.sqrt(n_features)), name='biases1')
# activation function(tanh)
y1 = tf.nn.tanh((tf.matmul(X, W1)+b1), name='activationLayer1')
#dropout Layer 1
drop_out_layer1 = tf.nn.dropout(y1, keep_prob)
```

3.4.2 Layer2 (convolutional layer)



```
# network parameters(weights and biases) are set and initialized(Layer2)
W2 = tf.Variable(tf.truncated_normal([n_neurons_in_h1, n_neurons_in_h2],mean=0, stddev=1 / np.sqrt(n_features)), name='weights2')
b2 = tf.Variable(tf.truncated_normal([n_neurons_in_h2],mean=0, stddev=1 / np.sqrt(n_features)), name='biases2')
# activation function(tanh)
y2 = tf.nn.tanh((tf.matmul(drop_out_layer1, W2)+b2), name='activationLayer2')
#dropout Layer 2
drop_out_layer2 = tf.nn.dropout(y2, keep_prob)
```

3.4.3 Layer3 (output layer)



```
# network parameters(weights and biases) are set and initialized(output layer)
Wo = tf.Variable(tf.truncated_normal([n_neurons_in_h2, n_classes],mean=0, stddev=1 / np.sqrt(n_features)), name='weightsOut')
bo = tf.Variable(tf.truncated_normal([n_classes],mean=0, stddev=1 / np.sqrt(n_features)), name='biasesOut')
# activation function(softmax)
a = tf.nn.softmax((tf.matmul(drop_out_layer2, Wo) + bo), name='activationOutputLayer')
```

4. Results

Two major types of results are concluded after the successful training and testing of the data set. One is the result of training and testing of the dataset to find the accuracy of results and the other one is the spectrogram and time-frequency graph of each sound class. Each sound class has different graphs but here the results of some classes are given as an example.

4.1 Time-Frequency Graph

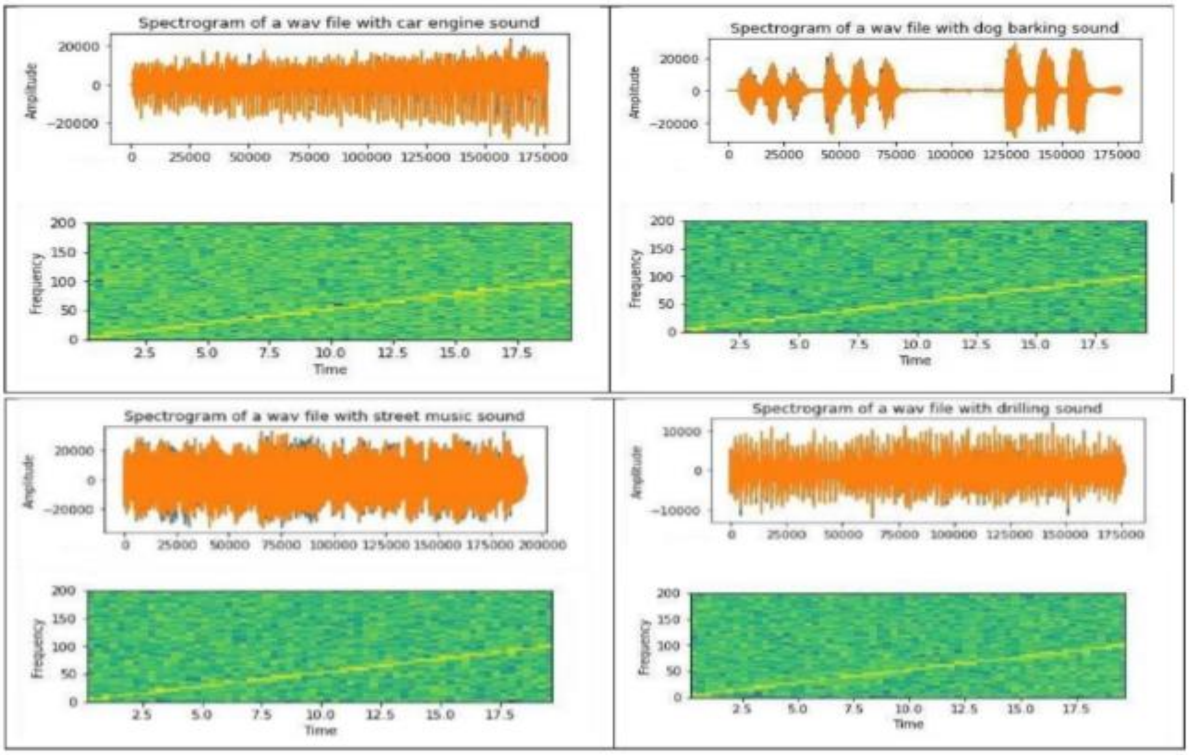
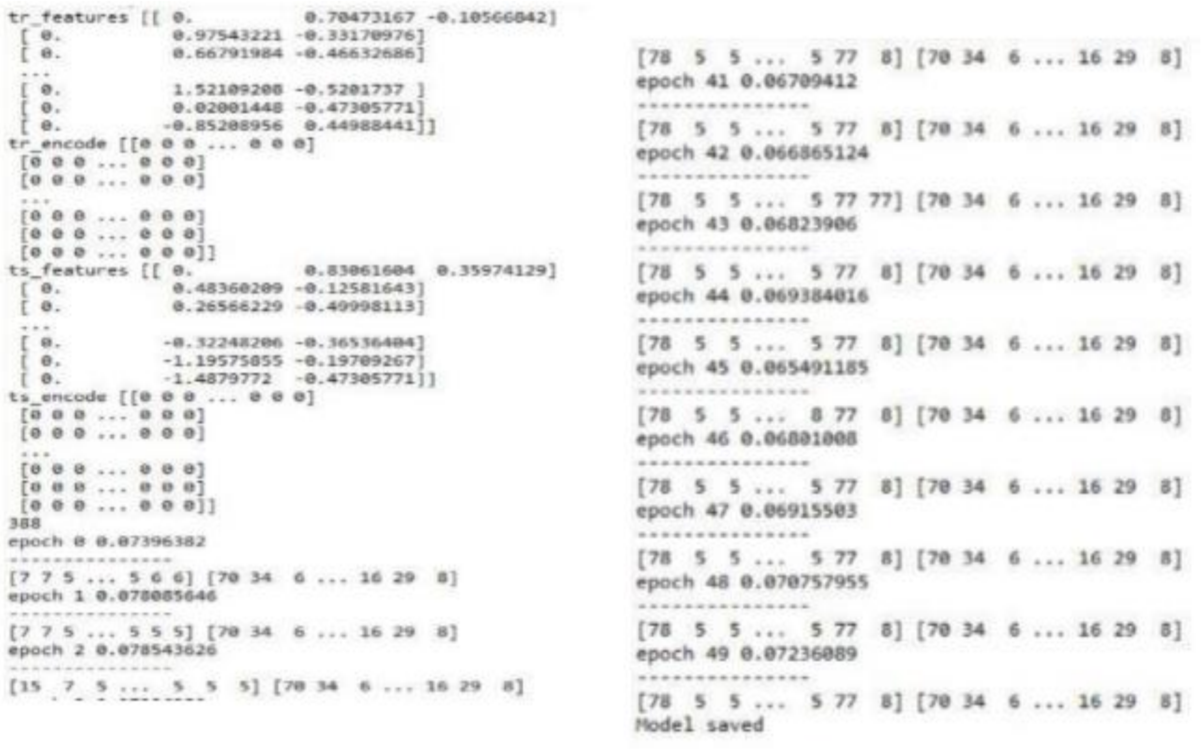


Figure 3. Time-Frequency Graph

The second part of the results is the dataset training and testing with the accuracy of each epoch. The results are given below:



```
tr_features [[ 0.          0.70473167 -0.10566842]
 [ 0.          0.97543221 -0.33170976]
 [ 0.          0.66791984 -0.46632686]
 ...
 [ 0.          1.52109208 -0.5201737 ]
 [ 0.          0.02001448 -0.47305771]
 [ 0.         -0.85208956  0.44988441]]
tr_encode [[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
ts_features [[ 0.          0.83061604  0.35974129]
 [ 0.          0.48360209 -0.12581643]
 [ 0.          0.26566229 -0.49998113]
 ...
 [ 0.         -0.32248206 -0.36536404]
 [ 0.         -1.19575855 -0.19709267]
 [ 0.         -1.4879772  -0.47305771]]
ts_encode [[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
388
epoch 0 0.07396382
---------------
[7 7 5 ... 5 6 6] [70 34  6 ... 16 29  8]
epoch 1 0.078085646
---------------
[7 7 5 ... 5 5 5] [70 34  6 ... 16 29  8]
epoch 2 0.078543626
---------------
[15  7  5 ...  5  5  5] [70 34  6 ... 16 29  8]
```

```
[78  5  5 ...  5 77  8] [70 34  6 ... 16 29  8]
epoch 41 0.06709412
---------------
[78  5  5 ...  5 77  8] [70 34  6 ... 16 29  8]
epoch 42 0.066865124
---------------
[78  5  5 ...  5 77 77] [70 34  6 ... 16 29  8]
epoch 43 0.06823906
---------------
[78  5  5 ...  5 77  8] [70 34  6 ... 16 29  8]
epoch 44 0.069384016
---------------
[78  5  5 ...  5 77  8] [70 34  6 ... 16 29  8]
epoch 45 0.065491185
---------------
[78  5  5 ...  8 77  8] [70 34  6 ... 16 29  8]
epoch 46 0.06801008
---------------
[78  5  5 ...  5 77  8] [70 34  6 ... 16 29  8]
epoch 47 0.06915503
---------------
[78  5  5 ...  5 77  8] [70 34  6 ... 16 29  8]
epoch 48 0.070757955
---------------
[78  5  5 ...  5 77  8] [70 34  6 ... 16 29  8]
epoch 49 0.07236089
---------------
[78  5  5 ...  5 77  8] [70 34  6 ... 16 29  8]
Model saved
```

Figure 3. Dataset Modeling

5. Discussion

Sounds are divided into 10 major categories. From them, each of the sounds has a different amplitude and time-frequency graph which differentiates every sound from the other. These differences are shown above in the result section in the form of spectrogram graphs of sounds. The classification of sounds is done with the help of these graphs because each sound has a different amplitude and frequency. The spectrogram is an important feature of sound. There are two important libraries matplotlib and numpy are used. Matplotlib plot the required graphs and numpy is used to do the required calculations. Dataset has been trained and tested to get maximum accuracy. In the above-given results, 50 epochs are trained which gives different values of accuracy every time. The data has been trained using TensorFlow and sklearn.model_selection and sklearn.preprocessing libraries.

6. Future work

After the successful classification of sound events, it can be used in different daily life applications like automatic music tagging, audio segmentation, audio source separation, music recommendation, and onset detection, etc. Sound classification can be implemented in the future using the combination CNN and RNN for the much higher accuracy of results and more feature extraction of sounds because CNN can extract only the higher-level feature of sounds. The purpose of the sound classification is to classify the sounds from an audio file containing mix sounds. The approach can further be used to identify male and female voices.

7. Conclusion

The main objective of this research was to classify sounds from an audio file containing raw sound and to check whether the multi-layer neural network model can solve the problem of sound classification in an effective way or not, so we proposed a system which classifies different sounds from an audio file. After conducting the experiment we proposed that convolution neural networks can efficiently be applied in sound classification even when we have less number of dataset values. Moreover, maximizing the dataset can lead to more accurate results. The question raised which was raised above in the introduction can now be answered that using the multi-layer technique we can easily classify sounds in different categories.

8. Acknowledgment

This research was supported by Mr Muhammad Zakir Khan. We thank our colleagues from City University of Science and Information Technology, who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper.

9. References

[1] Jiaxing Ye, Takumi Kobayashi, Masahiro Murakawa, Tetsuya Higuchi, "Acoustic Scene Classification based on Sound Textures and Events ", MM'15, October 26-30, 2015, Brisbane, Australia.

[2] Kun Qian1, 2), Zixing Zhang2), Alice Baird2), Bjorn Schuller2, 3), "Active Learning for Bird Sound Classification", ACUSTICA UNITED WITH ACUSTICA, vol. 103 (2017).

[3] Jort F Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Palkal, Marvin Ritter, "Audio Set: An Ontology and HumanLabeled Dataset for Audio Events", IEEE ICASSP 2017.

[4] Elsa Ferreira Gomes1, 2, Fabio Batista2, "Classifying Urban Sound using Time Series Motifs". Advanced Science and Technology Letters, vol. 97 (SUComS 2015), pp. 52-57.

[5] Emre Cakir, Giambattista Parascandolo, Toni Heittola, Heikki Huttunen, Tuomas Virtanen, "Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection." IEEE 2017, pp: 2329-9290.

[6] Huimin Zhao1,*, Xianglin Huang2, Wei Liu3, Lifang Yang4, "Environmental Sound Classification based on feature fusion", MATEC Web of Conference 173, 03059(2018), SMIMA 2018.

[7] Karol J. Piczak, "Environmental Sound Classification with Convolutional Neural Networks", 2015 IEEE INTERNATIONAL. WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 17-20, 2015, BOSTON, USA.

[8] Jonathan Dennis, Huy Dat Tran, Haizhou Li, “Spectrogram Image Feature for Sound Event Classification in Mismatched Conditions”, IEEE 2010.

[9] Justin Salamon1, 2, Juan Pablo Bello2, “Unsupervised Feature Learning for Urban Sound Classification”, IEEE 2015.

[10] Jiuwen Cao1, Min Cao1, Jianzhong Wang1, Chun Yin2, Danping Wang1,3, Pierre Paul Vidal1,4, “Urban Noise Recognition with Convolutional Neural Network”, Springer Science+Business Media, LLC, part of Springer Nature 2018.

[11] Vikas Gupta October 9, 2017, an example of Feed-forward Neural Network with one hidden layer, Neural Network diagram by Vikas Gupt, accessed November 8, 2019, < https://www.learnopencv.com/understanding-feedforward-neural-networks>.