Acoustic Classification using Deep Learning

Muhammad Ahsan Aslam, Muhammad Umer Sarwar, Muhammad Kashif Hanif, Ramzan Talib, Usama Khalid
Department of Computer Science
Government College University
Faisalabad, Pakistan

***Abstract*—Acoustic complements is an important methodology to perceive the sounds from environment. Significantly machines in different conditions can have the hearings capability like smartphones, different software or security systems. This kind of work can be implemented through conventional or deep learning machine models that contain revolutionized speech identification to understand general environment sounds. This work focuses on the acoustic classification and improves the performance of deep neural networks by using hybrid feature extraction methods. This study improves the efficiency of classification to extract features and make prediction of cost graph. We have adopted the hybrid feature extraction scheme consisting of DNN and CNN. The results have 12% improvement from the previous results by using mix feature extraction scheme.**

Keywords*—Acoustics; deep learning; machine learning; neural networks; audio sounds

I. Introduction

The advances in the automatic recognition of voice were consolidated in industrial systems [1]. Researchers have more interest to make advancement in identification quality. It is challenging task to identify acoustics in remote situations against the noisy background [2]. In other places, the advances in retrieval of music information have provided us with systems that can transcribe the notes and chords in the music or check the name of the track and the artist from a fragment of low quality sounds. The main focus of researchers is on classification of speech and music which can be heard mostly in a typical indoor outdoor environment [3]. Sound is few times a purposeful completion of different methods such as video, which transports information that would not be present like speech, processed data and songs of birds. The voice may also be easier to collect on a cellular phone. The information collected from a pragmatics audio inquiry can be purposeful for more working such as machine exploration, alert messages for user or analysis and prediction of event arrangements [4].

In the past years, a number of sound control techniques have been proposed. Deep learning is possibly the most recently used. The term deep learning employs a high level representation of low level data by stacking multiple levels using nonlinear module. There are several variants of deep learning architectures. The convolutional neural network is a profound learning method traditionally used for image distinction because of its good performance in learning distinctive local characteristics. The first Detection and Classification of Acoustic Scenes and Events (DCASE) challenge organized by the IEEE Audio and Acoustic Signal Processing (AASP) in 2013 and then the DCASE 2016 challenge with an extend Acoustic scene classification dataset [5].

A. Acoustic Scene Classification

Acoustic is a term used in different fields. Acoustics is challenged in many terms, as in acoustic physics means the knowledge in the field of mechanical waves that are in different things like gases, liquids and solids [6]. But in general, acoustics is related to sound, vibration, etc. A person or a scientist who works in the area of acoustical fields is known as acoustician. The study of sounds, their frequency and the behavior of sound are included in acoustics. Acoustics is the science of production, control, transmission, response and the effects of sound. The classification of an acoustic scene allows devices to understand the environment and opens various applications [7]. For example, devices such as androids, IPhones, Internet devices, wearable devices, and robots prepared using artificial intelligence can benefit from the situations of the classification of the acoustic scene. Also, intelligence assistants represent another field that can benefit from the classification of sound scenes. IPA are software providers that make advice and perform action by automatically identifying different types of input data including audio, images, user input, context-based information, such as location, weather and private schedules. Now, Microsoft's Cortana and Apple's Siri are using audio inputs and the use of context-based information gathered from environmental sounds has a significant potential to recommend appropriate actions to users.

Environmental sound is a combination of several sound Sources. It has a lot of information that can help humans for feeling of environment around. Voice evaluation draws researchers' attention to the machine Learning has been implemented in the community and monitoring, information services on robotic navigation and tourism, etc. The recognition of acoustic events is aimed at the identification of voices. Classification is the detection of these sounds. This will helpful in information retrieval, with multimedia applications content analysis, context-aware and audio-based devices surveillance and monitoring systems. The classifications of acoustic events have aimed to the divisions of different acoustic events in target groups for specified area [8].

Fig. 1 gives an example of an acoustic classification system. Acoustic classification consists of different phases. The audio data can be classified using unsupervised or supervised learning based on data. This can be achieved by implementing different kinds of deep learning models like neural networks and feature extraction techniques to get the

features from that audio data. Afterwards, the performance of the model is evaluated for acoustic scene classification. Acoustic models are not only relating with sound classification it is also use for image classifications and some other kinds of classifications [9].

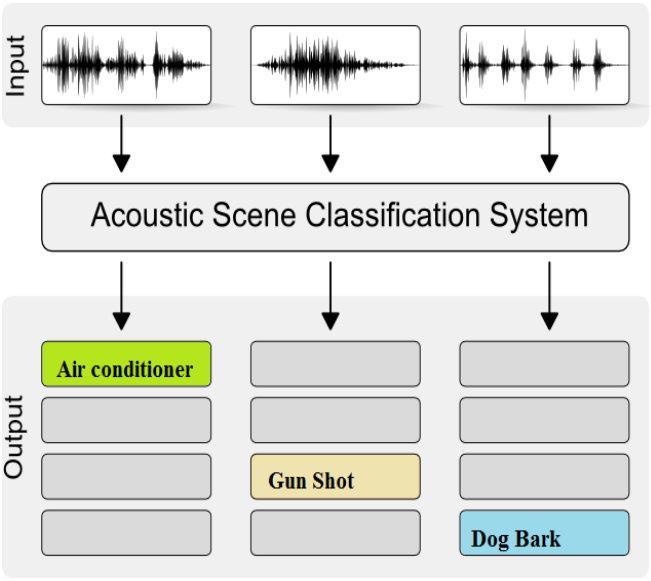

Fig. 1. Acoustic scene classification system.

B. Artificial Intellegence and Machine Learning

According to John McCarthy, artificial intelligence is the science and engineering to create the tools that are based on the sharp behavior [10]. Artificial intelligence is an intelligent behavior to make computer or software run by robots. Artificial intelligence perceives how peoples learn, how make decisions and efforts when struggling to solve the problems. The development of artificial intelligence began with the purpose of making the intelligence as same as found in the peoples. The aim of artificial intelligence is to develop the special machines (Machines that show the sharp behavior to learn, indicate, describe and advise who use it) and to perform the intelligence of peoples to systems (Develop systems which include, perceive, understand and act according to peoples). Simply Artificial Intelligence is the technique or method in which human transfer their intelligence to machines and make them intelligent to perform tasks intelligently for given data. Artificial machines get intelligence from the line of codes and also act according to the behavior of human's commands.

Artificial intelligence is scientific and technological area which has applications in other fields like Computers, Biology, Psychology, Linguistics, Mathematics and Engineering. One of the main impulses of artificial intelligence is the creation of functions for computer that are linked with the sharpness of peoples such as logical judgment, understanding and searching to solve problems. In the first half of the 20th century, people thought of artificial intelligence (AI) direct connection with the robot. Over the decades, growth in robotics proved to be enough today we have robots around us [11], but they do not stop developing at ANN.

Machine learning is an area of artificial intelligence based on the idea to make machines accessible for data and learn themselves from the accessed data. It was started of graph identification and the idea from which the systems can be studied using no program without performing certain jobs. The peoples who want to make research they specify their interest in this field want to see computers can learn from data. From initial point of learning of machines is useful due to new representations are related to models. These systems absorb from the past adaptation to make well-grounded reflection, conclusions and outcomes. This is a discipline which is not starting except who have attained freshness. However, a variety of system studying procedures is extended over a period, the capacity to implement complex numerical solutions automatically on large amount of data.

Machine learning workflow can be described through the following figure, in which completely describe how machine learning work in a well-mannered workflow. As machine learning wants to tackle huge and most complicated problems, the issue of concentrating on the most appropriate data in a conceivable overpowering measure of information. It has turned out to be progressively critical. For instance, information mining of organizations or researcher's records regularly includes managing numerous highlights and numerous cases, the Web and the Internet have put an expansive volume of low quality data in the simple access to a learning framework. Comparative issues emerge in the personalization of the separating frameworks for data recovery, email, arrange news and so forth.

Fig. 2 depicts how machine learning algorithms works. These algorithms take a dataset as input. First, it takes a dataset to extract the features from that data using different kinds of methods of features extraction. After feature extraction use the machine learning algorithms especially deep learning techniques to grouping the data objects. At the end predict the model by taking a testing dataset, it checks the performance on that test data and at the end get the results.

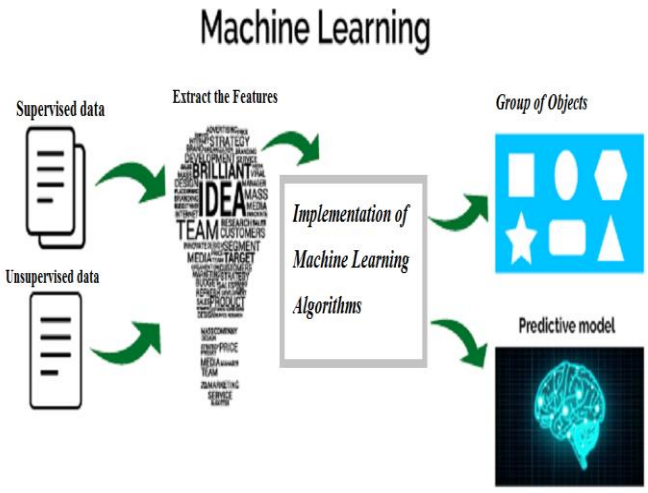

Fig. 2. Example of work-flow for machine learning.

C. Deep Learning

Deep learning is a branch of machine learning, in deep learning computational models comprising of numerous

handling layers and various reflection levels. These techniques have fundamentally enhanced the most progressive innovation in discourse acknowledgment, visual question acknowledgment, protest identification, medication revelation and genomics. Profound taking in many sided quality of expansive informational indexes, utilizing the back propagation calculation and a machine the interior parameters used to figure the impact on each layer from previous layer. Profound overlay systems have gotten leaps forward the territories of picture, video, discourse and picture handling. Voice and monotonous nets shed light on progressive articulations for example content and discourse.

Besides improving the authenticity of different patterns identification issues, one of Deep's core objectives learning, automatic machine learning revelation of numerous levels of impressions. The goal of utilizing crude information (e.g. picture pixels) as contribution to the models and let the models learn middle of the road introductions. That enables the model to learn highlight identifiers. This is particularly obvious it is imperative as demonstrated by Bengio for territories where highlights exist. It is difficult to formalize things like question acknowledgment and discourse acknowledgment errands. In the assignment of arranging woodland species, a few elective element extractors have been utilized (as noted above) demonstrates the trouble of finding a decent portrayal for inconveniences.

Recent developments in deep learning have led to a significant improvement in automatic speech recognition and music characterization. However, speech is one of many kinds of voices and people often count a variety of environmental voices to improve perception, when they are in danger and someone is walking through a busy street. It is a useful way to complement the visible information, such as more audio, videos and pictures, with the advantages that the sound can accumulate and be stored more easily. Deep learning has been a major practical success and a profound influence on machine learning and artificial intelligence literature. Practical success and deep learning in the literature have been investigated together with their theoretical features. Deep learning explores a wide range of areas for researchers to work with and learning outcomes in machine learning and artificial intelligence. Over the past decade, a number of volume control techniques have been proposed and deep learning is probably the most encouraging. As demonstrated by the deep learning, the method uses a high level representation of low level data by stacking multiple levels using a nonlinear module. Presently, deep learning architects have various variants and the convolutional neural network method is a profound learning method traditionally used for image segmentation because of its good performance in learning distinguishing local features.

Fig. 3 illustrates the layers of a neural network model. In deep learning, when it is required to learning deeply use different neural networks to classify things. These algorithms are referred to as artificial neural networks (ANNs) because the earliest use of artificial intelligence algorithms shows how learning is done in the causal mind [12]. Use different deep neural networks that can be applied for image, audio or video, speech, text, multimodal and IOT data recognition and classification while learning in depth. The deep neural network includes an input layer that contain input nodes, hidden layers with multiple hidden nodes and an output layer that contain output nodes. Initially, set the input values at the input nodes and these nodes calculate the output values by multiplying the weights and the input values which is the same for the other hidden nodes of the hidden layers and goes to the output layers. Output layer is the last layer containing output values.

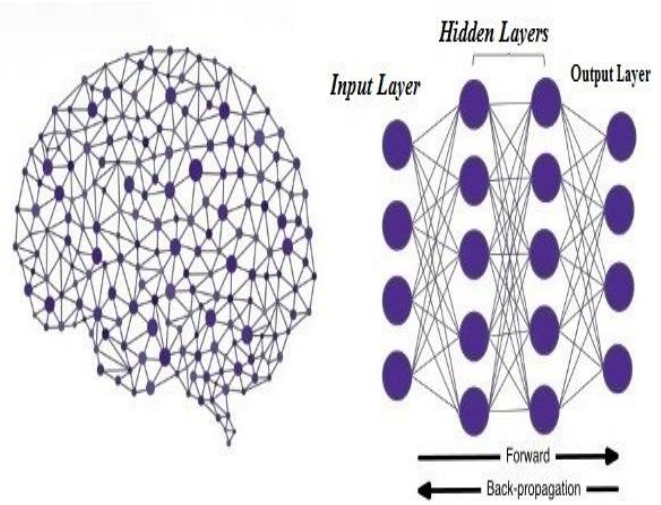

Fig. 3. Example of deep learning network [13].

II. Neural Networks

There are different kinds of neural networks that are used in deep learning for the different kinds of working, analysis of working and to perform certain kinds of artificial intelligence application in programming way. The neural networks that are included in deep learning are Convolution Neural Network (CNN), Artificial Neural Network (ANN), Long short term memory (LSTM) and Recurrent Neural Network (RNN).

A. DNN

Deep neural network (DNN) is supervised learning feedforward artificial network used in various applications like in image processing, in video recognition, automatic speech recognition and also it is trained for acoustic scene classification. It has different layers usually an input layer, several hidden layers to build a deep architecture and an output layer. Take a DNN and train it on the data taken from DCASE 2017 challenge that contain recording of different audio scenes [14]. Input layer of DNN receives the input from data and then work according to feed forward technique and pass it to its next hidden layer. In hidden layers complete its training on the provided data set then at the end find its results according to its training on training dataset. At the end to check its efficiency implants DNN on a test data and check its working according to its testing how it works? How it classifies the data?

B. CNN

CNN is also architecture of deep learning that is used for the classification of objects on the bases of layers. It also contained layers that are named as one input layer, several hidden layers, one output layer, working of CNN is same like that DNN take input from dataset and then apply functions on it at the hidden layers and find out result and show at the output layer. Commonly used CNN layers are Max pooling layer,

convolution layer and fully connected layer. In layer that is convolution, filter is convolved with input features. Max pooling layer do the job of down sampling the input and fully connected layer connects all neurons from previous layer with its every neuron.

Fig. 4 depicts the CNN model. First, the model takes the channel input. Then this input was convoluted at next stage using 1D convolution then at the next stage Max-pooling was performed after max pooling again convoluted then fully connected layer used on that data and at the end softmax is applied on that channel input at the last stage of CNN model.

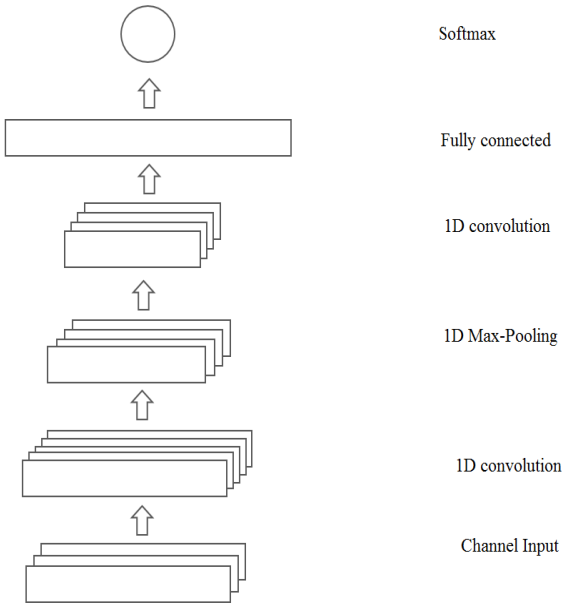

Fig. 4. CNN model.

III. Audio Framework

The audio framework used in this study consists of seventeen low level descriptors for temporal and spectral sound characteristics. This set of descriptors is generic to allow a wide range of applications to use these descriptors. That can be divided into six groups (Fig. 5). In addition to these six groups, there is a very simple audio framework mute descriptor. The two basic audio descriptors are the scaled values temporarily sampled. The audio waveform descriptor describes the audio waveform envelope by sampling the maximum and minimum values in an analysis window. The audio power descriptor describes the instantaneous power levelled temporarily. These can provide a quick and effective summary of a signal, especially for visualization purposes. The four basic spectral audio descriptors have the central link deriving from an analysis of the audio signal at a temporal frequency.

The audio spectrum envelope describes the short term power spectrum of an audio signal. It can be used to show a spectrogram and to synthesize a raw naturalization of the data or as a generic descriptor for research and comparison. The center of the audio spectrum is defined as the frequency center of the power weighted record. The power spectrum can be dominated by low or high frequencies [16]. The audio spectrum extension defines the sound moment of the power spectrum of the logarithmic frequency.

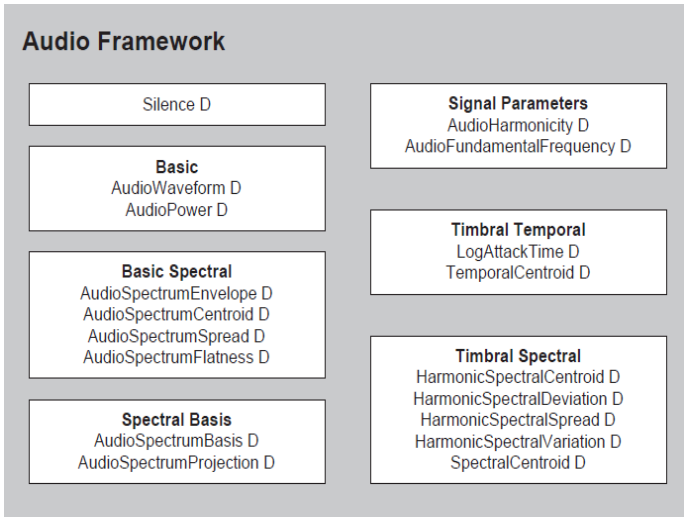

Fig. 5. Audio framework [15].

The logarithmic frequency indicates that the power spectrum is close to the centrifuge or it extends in the spectrum. The level of the audio spectrum describes the flatness characteristics of the short time power spectrum of an audio signal. This identifier refers to the scattering of the signal power spectrum from a smooth form. The spectral flatness analysis is calculated for the desired number of frequency bands. Sound signals can be used as a feature vector for robust pairing between pairs. The two spectral base identifiers represent small sized projections of a high dimensional spectral area to facilitate compactness and recognition.

The basic sound spectrum descriptor contains the basic functions used to transform high dimensional spectrum definitions to a low dimensional representation. The sound spectrum projection descriptor represents the low dimensional properties of a spectrum that is computed after projection on a reduced basis given by the sound spectrum base. The fundamental frequency is a good indication of music tone and vocal tone. Audio describes the harmonic level of an audio signal with a harmonic identifier. This makes it possible to distinguish between sounds with a harmonic spectrum and a spectrum with a non-harmonic spectrum between musical sounds and sound. The two timbral temporal identifiers describe the temporal properties of the audio segments. These are useful for describing the musical temperament. The temporal centric descriptor defines where the energy is produced over time depending on the length of the sound track. The five spectral identifiers of the timbre define the temporal characteristics of the sounds in the linear frequency domain. This makes it useful to explain the timbral spectrum coupled with the temporal identifiers especially the musical instrument tone. The spectral center identifier is very similar to the center of the sound spectrum with the use of a linear power spectrum as the only difference between them.

IV. Methodology

The methodology adopted for this research is similar to previous approaches that were based on machine learning

techniques. The most common machine learning technique used for classification of sounds or some other classifications is deep learning. In this approach, first add the dataset that contains the audio files. Afterwards, features from the dataset are extracted using the feature extraction. Then use these values in neural network layers to find out the results. The deep learning techniques are used for the classification of acoustic scenes, sounds, images or any other objects.

Fig. 6 elaborates the methodology adopted in this work. This research work uses a model that is based on two neural networks one is NN and second is CNN and implement separately. First of all, in these neural networks set the layers in three ways the layer that is at first end is called input layer that contain the input values next hidden layers that receive the input values and weights and then perform the function of neural networks and at the last end last layer that is called output layer and at that layer output values can be showed in the form of results. Secondly, use a dataset that contain the input in form of sounds use this dataset in the designed model and extract features from that dataset to take values in numeric form for the input and weight setting. To extract features, use the hybrid methods of feature extraction. Then feature extraction uses these values for finding the results.

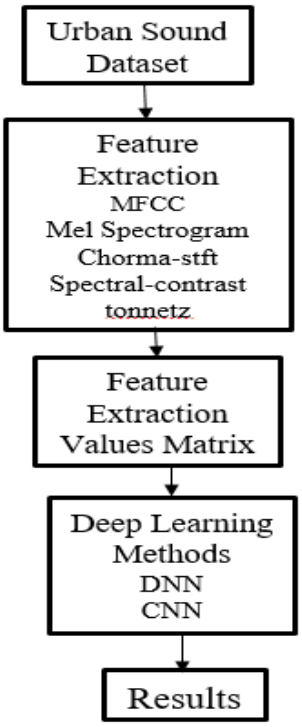

Fig. 6. Methodology adopted for this study.

A. Data Set

The dataset that used in this acoustic scene classification process was taken from the Urban sound datasets and this dataset named as Urbansound8K dataset. This dataset is divided into 10 folds and all the folds contain the audio files in the format of .wav of different kinds of sounds. These are fold1, fold2, fold3 and so on up to fold10. The dataset contains the 8732 named sound audio files and these all sounds are of length less than four seconds. The dataset consists of urban sounds including 10 types of low-level scientific classification, i.e., air conditioning, car horn, children playing, barking dogs, drilling, idling engine, blow, jackhammer, siren and street music. To avoid big contrasts of class of property, it establishes a confinement point of 1000 denominations per class, generating a total of 8732 denominations. Table I provides details of the features used in this study.

TABLE I. SOUNDS IN URBAN SOUND DATASET

Sr. No.	Sound	Format	Duration
1	Air conditioner	.wav	<=4 sec
2	Car horn	.wav	<=4 sec
3	Children playing	.wav	<=4 sec
4	Dog bark	.wav	<=4 sec
5	Drilling	.wav	<=4 sec
6	Engine idling	.wav	<=4 sec
7	Gunshot	.wav	<=4 sec
8	Jackhammer	.wav	<=4 sec
9	Siren	.wav	<=4 sec
10	Street music	.wav	<=4 sec

B. Feature Extraction

To extract the useful features from sound information, utilize the Librosa library. It gives a few strategies to extract diverse features from the sound files. Following are strategies to extract different highlights:

- **Spectrogram Mels**: calculates an energy spectrogram on the Mel scale
- **mfcc**: coefficients of cepstral Mel frequency
- **Chorma Stft**: calculates a chromatogram from a waveform or a power spectrogram
- **Spectral Contrast**: calculates the spectral contrast
- **Tonnetz**: calculate the characteristics of the tonal centroid (tonnetz)

Fig. 7 illustrates that how to simplify the procedure for highlighting the extraction of audio locks, two assistance strategies have been identified. First, parse audio files which takes the name of the main catalog, the subdirectories within the main index and the expansion of the document (the default is .wav) as information. At that time, it emphasizes each of the documents within the subdirectories and calls the second job called the associated function. Take the document as information, read the record by calling the librosa, load technique, concentrate and return the salient points mentioned above. In general, these two strategies are necessary to move from the raw sound clicks to the salient points of the instructions. The class name of each solid closure is in the document name.

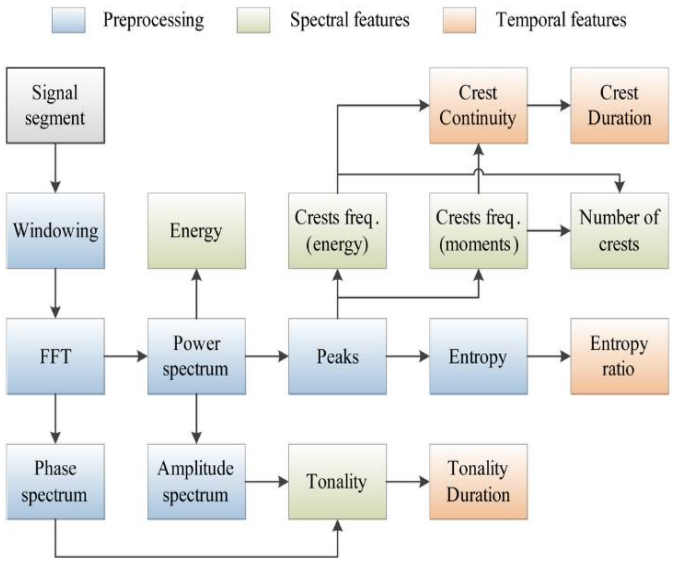

Fig. 7. Feature extraction from audio sounds.

The channel bank has a generally trapezoidal size reaction, with a significant portion between adjacent channels. In each channel, a flag envelope setting shown significantly with changing wave and low pass screening. Each flag is standardized on the calculation of the normal level on a complete articulation and isolate this quantity. Moderate adjustments in each standardized dissected indicator envelope then sieving flag through an unpredictable band passage channel and taking the log expansion yield.

C. Required Tools

The tools that are used for this acoustic scene classification process of sounds are as following:

- Anaconda
- Spyder
- Python

D. Required Libraries

Libraries that are used in this research work as follows:

- Numpy
- Scipy
- Pandas
- Matplotlib
- Plotly
- Theano
- Tensorflow
- Keras
- Librosa
- FFMPEG

V. Results and Discussion

Two neural networks are implemented in this research work with different feature extraction techniques to get the results over the Urban dataset 8k. Extract the features from the dataset that putted in the working code after feature extraction set the hidden layers and weights on those layers by using the feature values that extracted from the data. Data set contain the sounds of each classes and in all the fodders sounds of all classes are presents so selection of folds is very sensitive matter. The two neural networks that are used in this research work are DNN and CNN.

In deep neural network use the three folds of dataset and then parse that sound files and parsing that sound files extract the features from the sound files. By extracting features use the values of that features in the form weights on hidden layers and then apply the formula of DeepNN and get the results. Fig. 8 depicts the F-score obtained from that DeepNN. Approximately 80% accuracy was achieved using this approach.

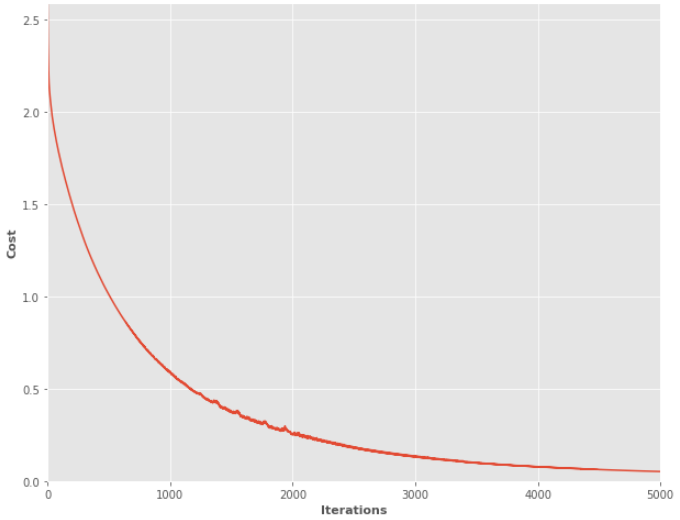

Fig. 8. Resulting curve of DNN.

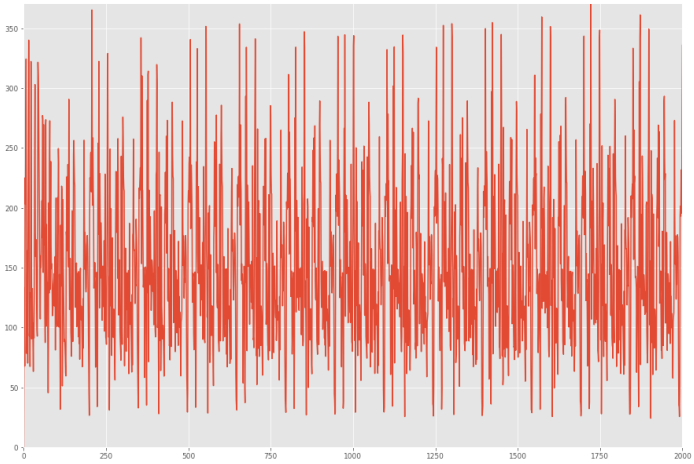

Fig. 9. Resulting graph of CNN.

The mfcc technique was used to extract the features from the sound files, two folds of dataset, 41 frames, 60 bands, 2460 feature size, 10 class labels, 2 channels, 50 batch size, 30 kernel size, 20 depths, 150 hidden units, 0.01 learning rate and 2000 training iterations. In CNN also weights are set by using the values of feature extraction that are extracted using the librosa library. Fig. 9 depicts the resulting plot that is obtained from the CNN.

Accuracy obtained from the CNN is 0.153 nearly about to 87% that is enhanced from the previous approach. In CNN working is slower but the results are accurate when compared to the DeepNN.

VI. Conclusion

The results after implementation of different hybrid feature extraction techniques on Urban dataset 8K improved in both machine learning techniques CNN and DNN. The fundamental task of this research work was that how improve the results from the previous methodology that is adopted on that Urban sound dataset 8K, so use many feature extraction methods to extract the feature values from sound data then apply deep learning methods DNN and CNN to get the results from extracted features. After implementation of this hybrid feature extraction method improve the efficiency of the neural networks on the described dataset.

References

[1] Barker, J., Vincent, E., Ma, N., Christensen, H., & Green, P. (2013). The PASCAL Chime speech separation and recognition challenge. Computer Speech & Language, 27(3), 621-633.

[2] Benetos, E., & Dixon, S. (2013). Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model. The Journal of the Acoustical Society of America, 133(3), 1727-1741.

[3] Ramona, M., & Peters, G. (2013, May). Audio Print: An efficient audio fingerprint system based on a novel cost-less synchronization scheme. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (pp. 818-822). IEEE.

[4] Marques, P. A., & De Araujo, C. B. (2014, October). The need to document and preserve natural Soundscape recordings as acoustic memories. In Proceedings Invisible Places Conference. Draft version available: http://invisibleplaces. org/pdf/ip2014-marques. pdf. Accessed (Vol. 13).

[5] Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., & Plumbley, M. D. (2015). Detection and classification of acoustic scenes and events. IEEE Transactions on Multimedia, 17(10), 1733-1746.

[6] Stowell, D., & Plumbley, M. D. (2014). Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. PeerJ, 2, e488.

[7] Bisot, V., Serizel, R., Essid, S., & Richard, G. (2016, March). Acoustic scene classification with matrix factorization for unsupervised feature learning. In Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on (pp. 6445-6449). IEEE.

[8] Boudreaux, K. (2018, March). Commonsense and Artificial Intelligence Systems. In Society for Information Technology & Teacher Education International Conference (pp. 21-24). Association for the Advancement of Computing in Education (AACE).

[9] Barchiesi, D., Giannoulis, D., Stowell, D., & Plumbley, M. D. (2015). Acoustic scene classification: Classifying environments from the sounds they produce. IEEE Signal Processing Magazine, 32(3), 16-34.

[10] Mesaros, A., Heittola, T., & Virtanen, T. (2016, August). TUT database for acoustic scene classification and sound event detection. In Signal Processing Conference (EUSIPCO), 2016 24th European (pp. 1128-1132). IEEE.

[11] Giannoulis, D., Stowell, D., Benetos, E., Rossignol, M., Lagrange, M., & Plumbley, M. D. (2013, September). A database and challenge for acoustic scene classification and event detection. In Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European (pp. 1-5). IEEE.

[12] Russell, S. J., & Norvig, P. (2016). Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited.

[13] Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (2013). Machine learning: An artificial intelligence approach. Springer Science & Business Media.

[14] Hovy, E., Navigli, R., & Ponzetto, S. P. (2013). Collaboratively built semi-structured content and Artificial Intelligence: The story so far. Artificial Intelligence, 194, 2-27.

[15] Andersson, Tobias. Audio classification andcontent description. 2004.

[16] Zaremba, W., Sutskever, I., & Vinyals, O. (2014). Recurrent neural network regularization. arXiv preprint arXiv:1409.2329.