SOUND CLASSIFICATION SYSTEM USING MACHINE LEARNING TECHNIQUES

Rachit Srivastava, Gourav Sharma, Rahul Kumar Sharma, Nikhil Yadav	Students
Dr. Manish Kumar	Associate Professor AIET Jaipur (manishkumarmukhija82@gmail.com)

***Abstract*- Although sound is the maximum critical speaking device for all dwelling organisms, people's day by day lives have brought extra forms of sounds into the herbal environment, which can also additionally or might not be useful. As a result, separating the regular communicating noises and identifying and interpreting them is a pressing necessity of the day. This study is about the different types of sounds that can be heard in cities. As a result, sound classification may aid the machine in determining the type of sound. This study examines the numerous strategies that are used to classify sounds and train machines to learn and analyze data in order to produce appropriate output. These kind of investigations can also aid in the detection of criminal activity. In addition, the numerous types of input and other parameters that can be employed for categorization were examined in this research. The methods' benefits and drawbacks were also addressed.**

Keywords- Convolutional Neural Networks, Machine Learning, Mel-Frequency Cepstral Coefficients.

I. INTRODUCTION

From a human standpoint, urban sounds can be easily predicted. However, developing a machine that may research that sound and classify it into its suitable class continues to be a time-ingesting procedure. Any type of sound archive, whether pre-specified or defined, should be able to be grasped, classified, and outputted to the user by the system. To do so, it must first learn about the sounds and variants available to it. Machine Learning is the process of allowing a machine to learn from a sound library (ML). As a result, machine learning is a large subject with numerous strategies for solving various types of complex issues. "Feature Extraction Techniques" is one such technique that is commonly used to classify urban noises. This feature extraction technique is in charge of extracting sound, compressing it, segmenting it, and finally assessing it. All of the above-mentioned processes are carried out individually in special sub-categories of this approach. These sub-categories are used in many papers in this article, either independently or in combination.

As a result, each research has produced distinct accuracy, demonstrating that all of these strategies have both benefits and drawbacks. These strategies additionally use Convolutional Neural Networks (CNN) and Deep Neural Networks (DNN) to reveal the sound, which improves the output.

OVERVIEW ON EXISTING TECHNIQUES

With using the Tree Bagger approach, David Li et al.,(2013) [1] display the accuracy. To begin, it businesses and golf equipment all of its instructions into a set tree. These tree units have been applied as predictors, and the end result turned into generated through evaluating them to the actual archives. This approach of category produces much less errors, ensuing in better precision rates. Once the timber were categorized separately, they may be forwarded to a voting-primarily based totally device that determines which classifier has the best accuracy rate, making sure that the output is strong for the user. Ayu and Karyono (2014) [2] carried out trials to illustrate that this method is unapproachable. As a end result, through building this device, they have got refuted the speculation of the "4 circumstance classifier." A extra knowledge of city sound, in line with Jiaxing Ye et al., (2016) [3], can result in a higher environment. This studies has confirmed to be an software-orientated one, with the sensor performing for use in site visitors lighting to locate ambulance sounds and regulate the indicators at this intersection to make sure that no site visitors occurs. Another software could be to put in it in clever town lampposts, wherein it could locate gunshots and ship a region alert to the police. As a end result, it affords each surveillance and safety.

Karol and Piczak (2015) [4] published a study in which the experimental model was evaluated and validated using a 5 fold and 2 fold architecture of neural networks.

The computer performed poorly when it was manually constructed to categorize the sound, but when it was

automated with convolutional neural networks, it performed admirably. best. They chose the probability-voting method over the standard majority-voting method since it showed to be more favorable.

According to Yuji Tokozume and Tatsuya Harada (2017) [5, the system's special parameter has characteristics, and the set is skilled accordingly. The static log-mel as one-channel statistics and the static blended delta log-mel [17] as second-channel enter are the 2 features. Despite the fact that the two characteristics are distinct, they share the same architecture, resulting in an accuracy of 71.0 percent, which can help enhance classification performance. They also demonstrated

that the method can extract a discriminative feature that works in tandem with the log-mel characteristics. Pooja et al., (2018) [6] used a system in which the input is fed into a feature extractor using software, and the result is a feature set. This set is given to a classification model that processes and categorizes the data using neural networks and machine learning. During the training, the above occurred. While the system is being tested, the input is processed in the same way that it was during the training, and then it is categorised and output is provided.

Minkyu Lim et al., (2018) [7] developed a system that uses a Convolutional Neural Network-based audio event classification. This system employs a feature extractor, which converts the audio signal (input in the form of an image) to PCM format.

Convolutional, pooling, and fully connected layers make up the event classifier utilized here. The convolutional layer [25] detects the overlaps that happened in the previous layer. The pooling layer reduces the dimension by combining features that are similar into one. The completely connected layer recognizes the image input's sound.

The studies were carried out by Mendoza et al., (2019) [8], who used three different types of input for sound categorization. Constant-Q transform (CQT) [13], Spectrogram features [14], and Spectrogram pictures are the three methods. CQT has been proven to be the best appropriate feature approach out of these three. The output delay is depending on the number of windows processed by the system, and the input is a three-second audio that is delivered as input via probability voting. Dropout [18] as well as Convolution Layers [19] in CNN architecture are used to determine the accuracy of classification systems.

1. RELATED WORK

According to Bruno da Silva et al., (2019) [9], the accuracy acquired for embedded gadgets declined dramatically because the model range increased. The referred to clarification has to do with the Librosa library [16] packages, that are in rate of audio characteristic extraction. The popularity of wonderful sound sorts changed into correct to the track of 50-60%. The side computation approach for embedded gadgets changed into aided via way of means of Google's Edge Tensor Processing Unit (TPU) [15]. Because every dataset had its personal versions of segregating the sound samples, it changed into located thru evaluation that every magnificence did now no longer offer the equal precision during the category system. Aditya Khamparia et al., (2019) [10] advanced a version to categorise environmental noises the usage of a Convolutional Neural Network with a two-layer architecture: absolutely linked and prediction layer [23]. With ReLU activation, the primary layer has 32 filters and the second one layer has 64 [22]. The functions are processed withinside the first layer, and the last output is anticipated withinside the prediction layer (i.e. magnificence). This can supply a degree of accuracy of as much as seventy seven percent. Tensor Deep Stacking Network Toolkit [24] is used right here because it consists of skills for checking out and training. Marc Green and Damian Thomas Murphy (2019) [11] used Augmented Reality (AR) and Machine Learning to assemble an iOS app for sound category. To show the digital items, the AR component [20] is hired. This app additionally employs a characteristic extraction approach called MFCC, which extracts audio and sends it to Core ML [21]. Core ML creates an item withinside the app and converts the version into an iOS-well matched format. The Gaussian Mixture Model and Support Vector Classifiers are hired withinside the category of sounds. In this study, Afshan Kaleem and Santi Prabha (2019) [12] look at diverse characteristic extraction techniques and category fashions. As a consequence, it famous which characteristic extraction approach produces top consequences whilst implemented to diverse categorization fashions. To make the system easier, the datasets (audio clips) are prepared into ten instructions on this study. The Librosa library is used to gather diverse audio facts factors which are idea to be valuable. They concluded that Mel-Frequency Cepstral Coefficients (MFCC) gave the most accuracy whilst hired in fashions after the experiments.

The topic of urban sound classification and identification has become a study field that has garnered increasing attention in recent years, owing to many artificial intelligence applications related to urban information processing. There have been various studies on this subject, in which different methodologies have been implemented and the results compared. The literature discusses algorithms, including their benefits and drawbacks, as well as their performance. We will present a brief literature

overview on the related work of sound identification in this part.

The challenge of identifying and classifying urban sounds is difficult due to the lack of organisation and the presence of interfering noise. It's critical to locate discriminating and informative audio representation as a feature and apply the classification task with a robust algorithm and model to assure the efficacy of the classifier model. As a result, researchers concentrated on developing powerful machine learning and deep learning algorithms on the one hand, and high-dimensional and differentiating audio features on the other. The audio features input to the classifiers must be carefully chosen to improve classification performance. Raw waveform, spectrogram, Mel spectrogram, and Mel-frequency cepstral coefficients are some of the most frequent audio formats (MFCCs). When compared to audio waveforms, time-frequency representations hold more information and are less in size. Although a large and complicated model can learn characteristics directly from the raw waveform, it will incur a significant computational cost. In, the convolutional neural network (CNN) was employed as the classifier to compare different time-frequency representations. Mel spectrograms outperformed spectrograms and MFCC and performed well across a variety of datasets, according to the findings. The Mel spectrograms are subjected to the discrete cosine transform (DCT), which decorrelates the spectral energies and destroys the local pattern in time-frequency representation. Because of the Mel filter banks, Mel spectrogram contains more discernible features than spectrogram.

Many studies have shown that deep neural network-based models outperform classical classifiers in addressing complicated classification problems in recent years. Traditional machine learning approaches for classification tasks include support vector machines (SVM), Gaussian mixture models (GMM), and k-means clustering. However, these classifiers are prone to noise and sensitive to the temporal dynamics of the audio in sound classification applications, resulting in a lack of robustness. With the popularity of deep learning-based models, an increasing number of investigations have exploited such methods in urban sound identification tasks. The CNN, which is extensively used in computer vision and image classification applications, is the most popular and straightforward deep learning model for classification tasks. It's also a good fit for our job because sound can be represented as a 2-D time-frequency representation from which localised spectrum patterns can be learned.

In environmental sound class research, became the first actual paintings to evalu- ate the overall performance of city sound class assignment the use of CNN. Its version includes convolutional layers with max-pooling and accompanied through absolutely related layers. Log-Mel spectrogram and its delta data had been used as audio illustration characteristic to be the enter for CNN. The test became primarily based totally on 3 publicly to be had datasets, ESC-50, ESC-10, and UrbanSound8K. For every dataset, accuracies of 80.5%, 64.9% and 72.7% had been acquired respectively. a CNN structure with 8 convolutional layers, each convolutional layers had been accompanied through a max-pooling layer, the performances of this proposed CNN had been as compared with VGG. The outcomes confirmed that the proposed CNN plays higher than VGG, evaluated on 3 datasets ESC-50, ESC-10 and UrbanSound8K the use of spectrogram because the enter, the accuracies of the proposed CNN had been 76.8%, 88.7% and 74.7% respectively. Large and deep CNN fashions used withinside the picture class also are carried out to sound identification, desirable overall performance may be done as well. researchers carried out AlexNet and GoogLeNet to the spectrograms of audio, and evaluated at the datasets ESC-50, ESC-10 and UrbanSound8K. The first-rate accuracies had been given through GoogLeNet, which had been 73%, 91%, and 93% respectively on every dataset. For the identical setups, GoogLeNet done better class accuracy than AlexNet, the motive for that is that GoogLeNet is appreciably deeper and has many extra layers than AlexNet.

Some works directly used raw waveforms in time-domain as input to the classification model. first used CNN to classify raw waveforms of environmental sounds. The model consists of 34 layers, where convolutional operations were 1-D convolution. The result on UrbanSound8K was 71.8%, which was comparable to using log-Mel spectrogram inputs and 2-D convolution. However, the neural network model was much larger compared to two convolutional layers in divided the waveform into overlapped frames by sliding window and used 1-D CNN that directly learned features from waveforms. The model achieved 89% of accuracy on UrbanSound8K, competitive to other results from methods using spectrogram representations and 2-D CNN. However, the input was a long sequence of vector and the computational cost was huge.

DATA AUDIO AND FEATURES:

The dataset that will be utilised in this project, the Urbansound 8k, will be introduced in this chapter. The many representations of sound signals that can be utilised as input to deep neural networks are then described and justified, with visualisations supplied at the same time. Finally, we go over the audio signal processing techniques and activities.

AUDIO DATASETS:

There are various open-source datasets available for the research of detection and classification of acoustic scenes

and events (DCASE). ESC-50, ESC-10, and UrbanSound8K are the most popular and commonly used datasets for this research area. They are all sourced from, which is a website that collects field recordings posted by various participants. Because we are interested in the urban sound environment, we will use the Urban- Sound8K dataset as our main emphasis in this project. has proposed and motivated this dataset, as well as a complete taxonomy of urban noises, which may be found and downloaded at. This dataset contains 8732 sound snippets, each of which is no longer than 4 seconds and is organised into 10 folders. Table 2.1 shows the quantity of audio clips in each folder.

Fold Number	Number of Samples
Fold 1	873
Fold 2	888
Fold 3	925
Fold 4	990
Fold 5	936
Fold 6	823
Fold 7	838
Fold 8	806
Fold 9	816
Fold 10	837

Table 2.1: Number of audio samples in each fold.

The fact that the audio snippets are labelled is a key feature of this dataset, since it allows us to perform supervised learning. Furthermore, the audio classes provided in this dataset are related to the sound we're interested in, which is common urban noise. Air conditioner, car horn, children playing, dog bark, drilling, engine running, rifle shot, jackhammer, siren, and street music are the ten classes of identified urban noises, each with a unique number class identity (ID) ranging from 0 to 9.The class names and the corresponding numeric class IDs, together with the number of samples for each class are listed in table 2.2. These 10 classes of sounds were evenly pre-sorted into the 10 folds, which allows us to perform cross-validation or choose the training set and testing set base on the fold number.

ClassName	NumericClassID	NumberofSamples
Airconditioner	0	1000
Carhorn	1	42 9
Childrenplaying	2	1000
Dogbark	3	1000
Drilling	4	1000
Engineidling	5	1000
Gunshot	6	374
Jackhammer	7	1000
Siren	8	929
Streetmusic	9	1000

Table 2.2: Class name, class IDs and number of samples.

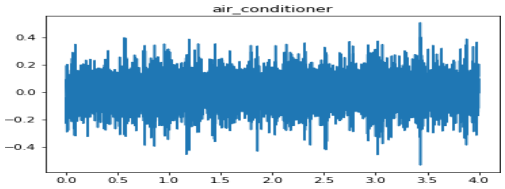
Audio Feature:

Some typical audio signal representations are explained in this section. These terms are studied in this order: Audio waveform, spectrogram, Mel spectrogram, delta feature. Each subsection examines and visualises the definitions of various features, as well as the benefits and drawbacks of using each feature as a neural network input.

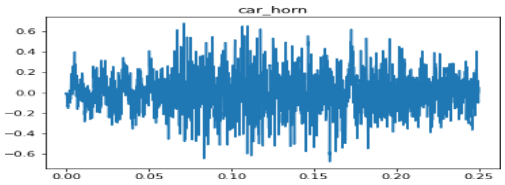
Waveform

A signal's waveform is a graph that depicts its shape as a function of time, demonstrating how the signal's amplitude changes over time. The audio waveform depicts how the loudness of a sound changes over time from the perspective of an audio signal. Figure 2.1 depicts the waveform visualisation of ten separate classes, with each audio captured at a sampling frequency of 22.05kHz.

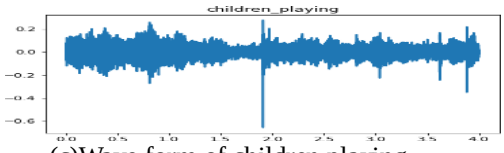
The waveform only contains information about how audio changes over time, but it lack frequency information, making it a less discriminating representation than the spectrogram. Consider a deep neural network that is powerful enough to learn the underlying features from the waveform directly, allowing the classification task to be completed; however, the model's complexity and computing cost are enormous, and a significant number of input samples are expected. As a result, time-frequency spectrum properties are more promising for classification, as will be detailed in the next subsections.



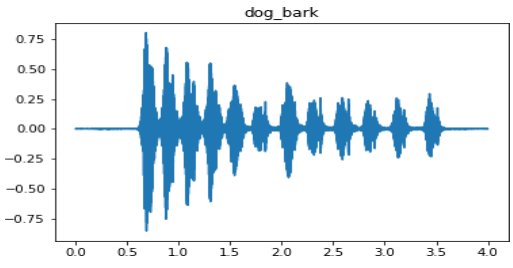
(a) Wave form of air conditioner



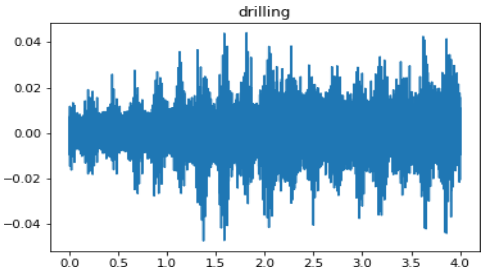
(b)Wave form of carhorn



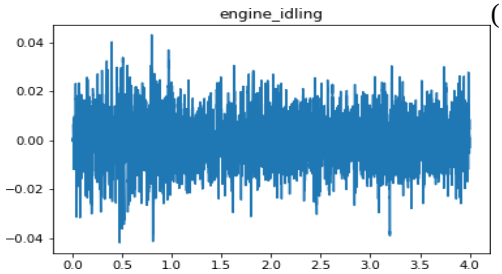
(c)Wave form of children playing



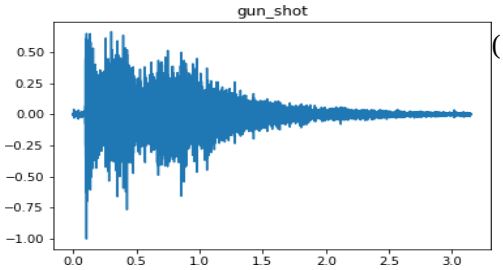
(d)Wave form of dog bark



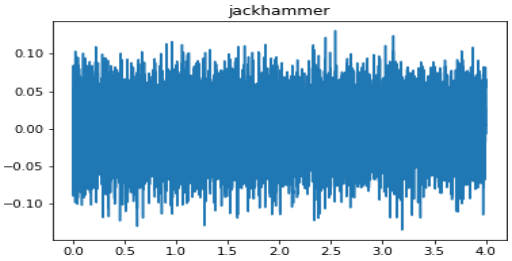
(e)Wave form of drilling



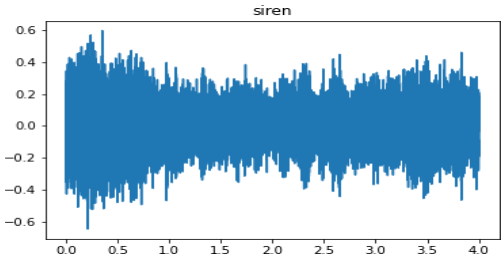
f)Wave form of engineidling



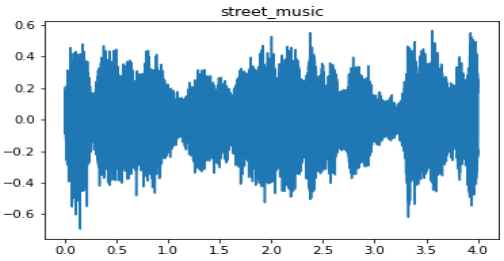
g)Wave form of gunshot



(h)Wave form of jack hammer



(i)Wave form of siren



(j)Wave form of street music

Spectrogram

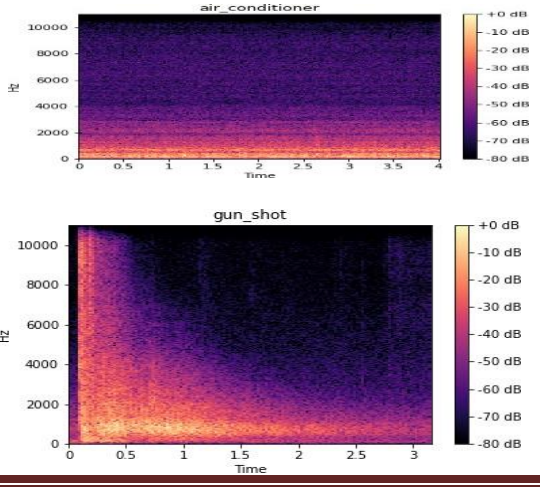
The spectrogram is a signal representation that shows signal strength with time at different frequencies. Spectrograms can be two-dimensional with colour representing a third variable. On the horizontal axis, one dimension represents time, while on the vertical axis, the other dimension represents frequency. The third variable depicts the amplitude or power of a particular frequency at a given moment, which is represented by the colour intensity or brightness of each pixel point in the two-dimensional graphs. As a result, a spectrogram depicts the frequency spectrum of sound as a function of time. Not only can we see where the energy is divided over frequency in a spectrogram, but we can also see how energy levels change over time.

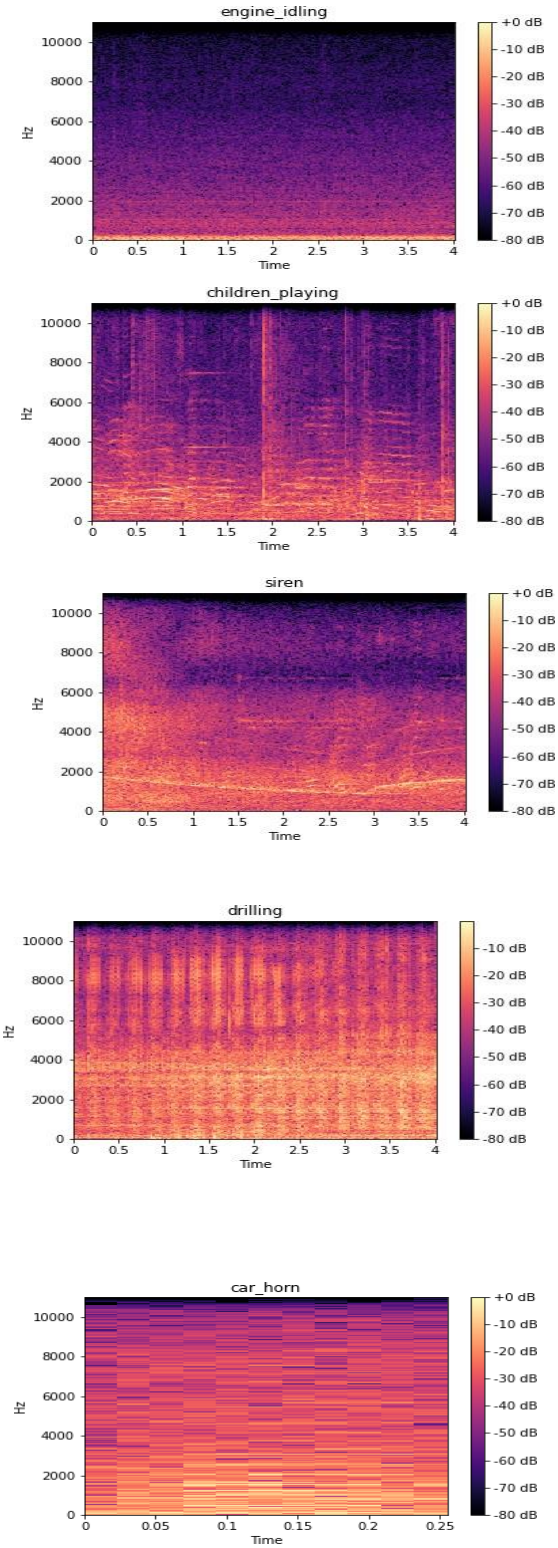
Given an audio signal, which is a set of sampled points denoted as x(n), to Each frame is applied with a window function $w(n)$, such as Hann window, in order to over comes pectral leaks through sidelobes.

Then, the Fourier Transform is applied to each short-timeframe.

The power spectrum density is obtain by taking the square of the magnitude of thefrequency spectrum. In each short-time frame, its power spectrum density is denoted as $P_\alpha(k)$, which is given as follows

$$P_\alpha(k)=|X_\alpha(k)|^2$$





3. Finally, the spectrogram is obtained by transferring the power into the decibel scale,which is a unit measuring the intensity of a sound.This decibel scale is inspired byhuman's perception of loudness.

II. CONCLUSION

Although there are numerous ways for categorising city sounds, the study clearly shows that the right combination of Machine Learning Technique with the sound archive device created by myself gives great results. Making the system distinguish sound that has travelled through the environment will be valuable for both research and surveillance purposes. It's also referred to as schooling a method for categorising a specific sound and demonstrating that it's far as good as humans at predicting the surroundings.

ACKNOWLEDGEMENT

We would love to thank our Management, Advisor and Principal of Arya Institute of Engineering and Technology College for continuously motivating us to do studies in our paintings surroundings and encourages us to post extra studies papers.

III. REFERENCES

[1] Li David, Tam Jason and Toub Derek (2013): Auditory Scene Classification using Machine Learning Techniques in IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events.

[2] Shekar Melati Ayu Indah (Ayu) and Karyono Kanisius (Karyono) (2014): (AudiTion)Android based Sound Detection application for Hearing Impaired using AdaBoostM1 Classifier with RepTree WeakLearner in APCASE- Asia Pacific Conference on Computer Aided System Engineering .

[3] Ye Jiaxing, Kobayashi Takumi and Murakawa Masahiro (2016): Urban Sound Event Classification based on Local and Global features Aggregation in Applied Acoustics.

[4] Piczak Karol J. (2015): Environmental Sound Classification with Convolutinal Neural Networks in IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP). (pp. 1–6).

[5] Tokozume Yuji and Harada Tatsuya (2017): Learning Environmental Sounds with end-to-end Convolutional Neural Network in ICASSP- IEEE International Conference on Acoustics, Speech and Signal Processing.

[6] R.K Pooja, Shetty Srishti, M Suhani and Mr. D.R. Janardhana (2014) Sound Classification using Machine Learning and Neural Networks in IJIRT-International Journal of Innovative Research in Technology.

[7] Minkyu Lim, Donghyun Lee, Hosung Park, Yoseb Kang, Junseok Oh, Jeong-Sik Park, Gil-Jin Jang and Ji Hwan Kim (2018): Convolutional Neural Network based audio Event Classification in KSII (Korean Society for Internet Information) Transactions on Internet and Information System.

[8] Mendoza Jose Marie, Tan Vanessa, Fuentes Jr. Vivencio, Perez Gabriel and Tiglao Nestor Michael (2019): Audio Event Detection using Wireless Sensor Network based on Deep Learning in International Wireless Internet Conference.

[9] Silva Bruno da, Happi Axel W., Bracken An and Touhafi Abdellah (2019): Evaluation of classical Machine learning Techniques towards Urban Sound Recognition on Embedded Systems in Applied Sciences.

[10] Khamparia Aditya, Gupta Deepak, GiaNhu Nguyen and Krishna Ashish (2019):Sound Classification using Convolutional Neural Network and Tensor Deep StackingNetwork in IEEE Access.

[11] Green Marc and Murphy Damian Thomas (2019): Environmental Sound Monitoring using Machine Learning on Mobile devices in Applied Acoustics 159.

[12] Afshankaleem and Prabha I. Santi (2019): Enhancement of Urban Sound Classification using various Feature Extraction Techniques in IJRTE-International Journal of Recent Technology and Engineering.

[13] Lidy T., Schindler, A (2016): CQT-based convolutional neural networks for audio scene classification and domestic audio tagging in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016), DCASE2016 Challenge, vol. 90.

[14] Nanni, L., Costa, Y.M.G., Lucio, D.R., Silla, C.N. Jrand and Brahnam, S (2017): Combining visual and acoustic features for audio classification tasks in: Pattern Recognition Letters.