2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS) | 978-1-6654-9199-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSCDS56580.2023.10104722

Enhancing Environmental Sounds Classification through Deep Learning Techniques

Siva Krishna Dasari[1]
Department of Computer Science & Engineering
[1]GMR Institute of Technology,
Rajam, Vizianagaram
Sivakrishna.d@gmrit.edu.in

Sarath Kumar Kella[2]
Department of Computer Science & Engineering
[2]GMR Institute of Technology,
Rajam, Vizianagaram
20341A0591@gmrit.edu.in

Raghava Manda[3]
Department of Computer Science & Engineering
[3]GMR Institute of Technology,
Rajam, Vizianagaram
20341A05B6@gmrit.edu.in

Abstract— **Environmental sound classification has become an important application for the research process in recent years, as it has many applications in several fields such as urban noise management, wildlife monitoring, and intelligent sound system. Sound classification involves mainly classifying the different sounds and predicting the class of that sound by using deep learning techniques. These techniques have proven to be very effective in many sound analysis and classification areas. This study uses neural networks to learn high-level features from audio clips, then build some layers, and finally make a fully connected layer to know the final classification. Generally, Mel-Frequency Cepstral Coefficients (MFCCs) have been taken as a feature representation for audio signals. These features can be generated from the spectrograms, which are inputs to the organized model. The model can be built using networks like artificial neural networks and convolution neural networks. The architecture of the neural network is designed to effectively capture similar features of environmental sounds by reducing the competition of complexity. To implement the sound classification, a UrbanSoundDataset8k is used to predict the individual classes.**

Keywords— Mel-Frequency Cepstral Coefficients (MFCC), Feature representation, Neural networks, Spectrogram, Raw audio signals, Audio processing.

I. INTRODUCTION

Over the years, the rapid improvement of smart cities and the raise in demand for monitoring and management of urban environments have led to the increasing importance of environmental sound classification. Audio classification is a complicated task in the field of audio processing with many applications widely ranging from recognition and classification to monitoring the sounds. The main aim of audio classification is to categorize the audio signals into several existing classes such as music, speech, etc. Many traditional techniques in machine learning have been used for audio classification over the years.

In this work, we proposed a deep-learning neural network for audio classification. The generalized model is based on ANN and CNN, which will give more accuracy that can be suggestable for the classification of environmental sounds. The proposed model is evaluated based on its performance and compared to the predefined state-of-the-art methods.

In addition to this, we mainly focus on the challenges that can be faced while proposing a deep learning model for audio classification that influence the recent technology in deep learning. The approach is intended to know the high-level of noise and distortion that can be useful for reducing complexity levels for the model. Another key challenge is feature representation in the audio signal. This can be solved by extracting the features from MFCCs that have proven very effective and not optimal in all cases but can deal with audio signals significantly.

In this paper, the proposed model has shown the ability to learn more effectively in the representation of features from the audio signals. Moreover, these approaches classically require a large amount of labeled data to train the proposed model which can be difficult and consume more time. The direct attraction of the MFCC features can reduce dimensionality to work with a certain model and eliminates the need for hand-crafted features to get the successful model at the time of prediction.

This project mainly consists of audio processing for the feature extraction from MFCCs and that can be represented in arrays in which ANN can be taken in single-dimensional arrays for prediction purposes. In addition to this, multi-dimensional arrays can be used for feature representations that can be represented as image processing for classification. This can be taken into different customized CNN models to predict the correct class for the audio signal.

II. RELATED WORKS

Over many years, several approaches have been proposed for classifying different environmental sounds. The approaches used for audio classification can be mainly divided into two categories.

B Wu and XP Zhang proposed an approach to environmental sound classification via time-frequency attention and frame-wise self-attention-based neural networks. It uses a combination of two attention mechanisms to improve the accuracy of ESC. TFA is used for time-frequency information and FSA is used to build a relationship between frames of an audio clip. UrbanSound dataset which is challenging because of numerous clutter interference[1]. Changsong and Barsim proposed a multi-level attention model for weakly supervised audio classification. In this, the model can take a small amount of label information for training as a class label instead of frame-level clips. It performs the existence of a state-of-the-art model on different audio clips [2].

Hershey and Chaudhuri proposed a CNN architecture for large-scale audio classification. The authors compare and analyze the difference CNN model and evaluate and large dataset. It provides an effective CNN model and highlights of the model to get good performance [3]. Aksoy and Uygar proposed a CNN model that contains multiple convolution and pooling layers then followed by a fully connected layer for the output classification. This model outperforms predefined state-of-the-art models on different classifications [4]. Anam Bansal and Naresh Kumar Garg proposed ML approaches such as SVM and decision trees and deep learning approaches such as CNN and RNN. The author describes a comprehensive analysis of the performance of different datasets and provides valuable insights into the current state-of-the-art method [5]. Zhang and Qiao proposed an attention-based convolutional recurrent neural network (CRNN). It is a connection between CNN and RNN-based attention mechanisms. This model is created to extract important features from the audio clip to get both local and global information that can be used for training the model to outperform the existing predefined models and demonstrates effectiveness in the combination of CNN and RNN [6].

The authors Baljinder Kaur and Jaskirat Singh mainly focused on the task of environmental sound classification which insists on classifying several audios in an acoustic environment into different classes. The authors also review pre-existing models including traditional ML models and deep neural networks and provide an analysis of their model performances. The authors also provide challenges involved in their research and describe the overcomes of the models [7]. Wenjie and Xianqing proposed a temporal-frequency attention-based convolution neural network mechanism. It mainly focuses on the time and frequency components of audio signals. This model can be evaluated on a large dataset to get better performance. It demonstrates the importance of choosing both the temporal and frequency domains as input for this model [8]. Walden and Dasgupta are mainly focused on how to improve the environmental perception of autonomous vehicles. The author proposed environmental sound classification could enhance the perception capabilities of vehicles. They work on CNN and RNN and evaluate the proposed model on a huge dataset. The authors also describe how to detect and respond to various audio signals in the environment and provide highlight the potential of classification as a tool for enhancing capabilities [9]. The authors Bahmei and Birmingham proposed a model CNN-RNN using data argumentation and deep convolutional generative adversarial networks for environmental sound classification. The combination of CNN and RNN with data argumentation utilizes a DCGAN to generate the extra training samples for enhancing the representation power of the model. This can be trained on big datasets to give better performance. The model size is too large so it’s consuming a large space [10]. Fady Medhat, David Chesmore, and John Robinson proposed a masked conditional neural network for sound classification. This approach utilizes masks to selectively extracted important features from the important data. This model demonstrates the effectiveness of MCNNs for sound classification. The authors also provided Long Short-Term Memory from the recurrent neural network to take the output as an input from the previous state to classify the output [11]. Mang, Canadas-Quesada proposed a Cochleogram-based Adventitious Sounds Classification mechanism. These adventitious sounds are unwanted and not a part of the intended audio signal such as noise. The authors proposed the CNN model taking input as Cochleograms which are spectrograms [12].

Burak Uzkent, Hakan, and Buket are the authors who proposed some of the traditional machine learning approaches such as support vector machine, decision tree, random forest, logistic regression, naïve Bayes, and Gaussian mixture models. These have been mostly used for the audio classification of environmental sounds. This approach considers most features from both frequency and time domains as inputs given to the model. The extracted features that can be most effective to get good model performance [13]. Aditya Khamparia and Deepak Gupta presented a CNN and Tensor Deep Stacking Networks (TDSNs) as deep learning approach. The author states that CNN is a feature extractor and TDSN is an output classifier. The model which can be trained on several datasets provides better performance. The authors also highlight the potential of the combination of connections between different neural network architectures [14]. Leo Cances and Etienne Labbe demonstrate the comparison of different Semi-supervised deep learning algorithms. Semi-supervised learning is a machine learning type that uses both labeled and unlabelled data to improve the accuracy of that model. The authors also compare the models based on their performance they came to provide a model which gave high accuracy [15].

III. METHODOLOGY

3.1 DATASET

We used a UrbanSound8K dataset which is used for audio classification consisting of 8732 audio clips which are collected from various locations in New York and recorded at a sample rate of 44.1kHz with each length of 4sec. The audio clips are labeled with 10 classes such as air_conditioner, car_horn, children playing, dog_bark, drilling, engine_idling, gunshot, jackhammer, siren, and street_music. This dataset is widely useful in research on environmental sound classification.

3.2 LIBRARIES

Some of the important libraries that can be used for the implementation of audio classification are as follows:

3.2.1 TensorFlow:

It is an open-source software library for deep learning that can be helpful for audio classification tasks. It provides a wide range of tools for building and training a deep neural network and has built-in support for working with raw audio data.

3.2.2 Scikit-learn:

This is a popular machine learning and deep learning library in python. It provides tools for data pre-processing, feature selection, feature extraction, training, and testing the model. This can be used in conjunction with many libraries such as TensorFlow and librosa to construct an end-to-end audio classification model.

3.2.3 Librosa:

This is a python library for storing and analyzing audio signal data. This provides different audio processing tools which include the ability to compute Mel-spectrogram and MFCCs. This can be widely used in the field of speech and music analysis for classification.

3.3 AUDIO DATA PRE-PROCESSING

In this stage data preparation and feature extraction are involved. The audio signals can be transformed into numerical representation as Mel-Frequency Cepstral Coefficients derived from spectrograms. These features provide more compaction and the representation of audio in an understandable way that can be taken into a deep learning model. Some of the processes that can be used for feature extraction are as follows:

3.3.1 MFCCs:

MFCCs are commonly used for audio signal representation for speech and music recognition. These are a set of extracted features from the audio signals which capture the most important characteristics in the signal's spectral information i.e., the distribution of energy across various frequency bands. Generally, these are derived from a Mel-scaled spectrogram for an audio signal. Once the MFCCs are extracted in the form of a single-dimensional array and multi-dimensional array as feature computation from the images.

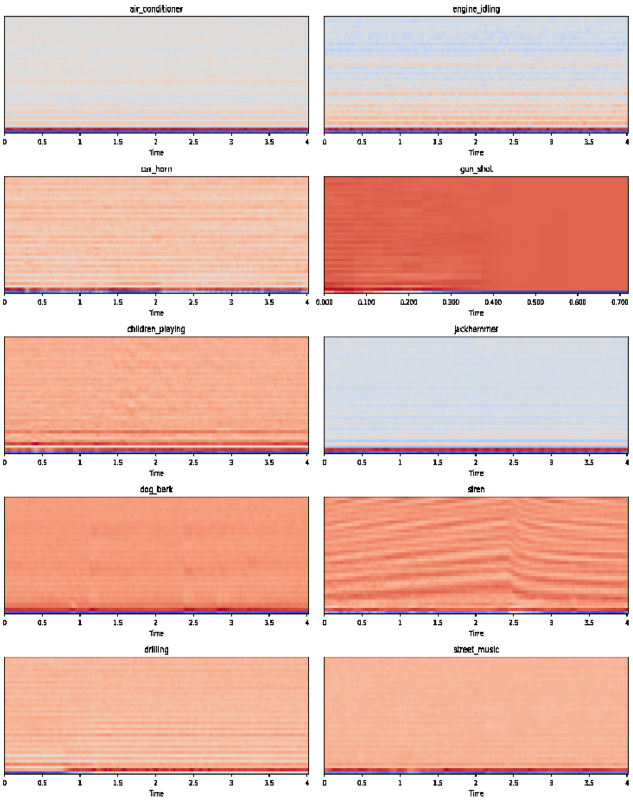


Fig. 1 MFCCs Representation

3.3.2 Mel-spectrogram:

This can be used for audio signal representation which can be helpful for feature extraction to classify the audio. Mel-spectrogram is based on a time-frequency representation of an audio signal that shows the energy distribution across various frequencies at each time step. The advantage of the Mel-spectrogram is that it can help in reducing the dimensionality of the occurrence of data. This can capture the most important spectral characteristics from the audio signals while reducing the many features required to signal representation. This can be mainly helpful in audio classifications that can make it easier to train the model to get classify the different types of audio signals accurately based on spectral information. This can be used as input to deep neural networks such as CNN and RNN for audio classification tasks. This can be corresponding to identifying patterns and relationships that allow it to make correct predictions of the label about a new audio sample.

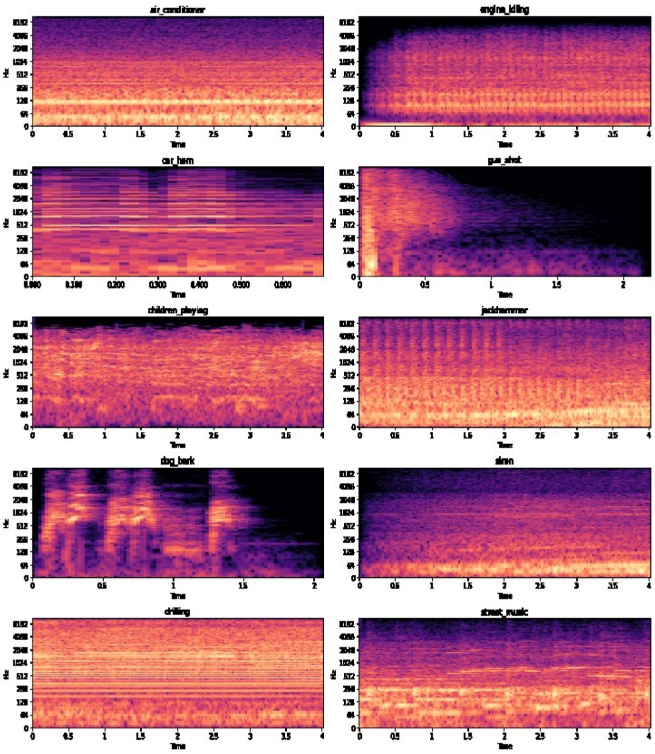


Fig. 2 Mel-spectrogram Representation

3.3.3 Working process of Mel-spectrogram to extract MFCCs:

The working steps are as follows:

1. First, compute the spectrogram as a time-frequency representation by using a Short-Time Fourier Transform (STFT).
2. Next, the spectrogram is transformed into a Mel-scaled spectrogram. This can be done by mapping a linear frequency scale to a non-linear frequency scale that matches the frequency resolution of the human auditory system. This can be useful to represent how human listeners can receive the signal.
3. Then it can be logarithmically compressed to decrease the dynamic range of data. It helps to balance the main frequency components and reduces the large fluctuations and noise in the signal.
4. Finally, a discrete cosine transform (DCT) is applied to a logarithmically compressed Mel-scaled spectrogram for the extraction of MFCCs. These are the numerical coefficients of the spectral content in

the audio signal while reducing dimensionality in the data.

3.4 PROPOSED MODEL

The model is based on convolution neural networks to enhance the environmental sound classification. Our proposed model is compared to the CNN-RNN method which was described by Bahmei, B., Birmingham (2022) [10].

3.4.1 CNN architecture:

A CNN is a deep-learning model mainly used for image classification. However, similar principles can be applied to audio classifications, as audio signals can be represented as either spectrograms or time-frequency representations that have a grid-like structure like images. There are mainly five layers in CNN. The mathematical model of a CNN involves a set of weighted inputs, an activation function, and a bias.

$$z = b + w1x1+w2x2+\ldots+wn* \ xn \qquad (1)$$

$$a = f(z) \qquad (2)$$

where z is the weighted sum of inputs plus bias. w1,w2,…wn are the weights allocated to inputs x1,x2,..xn respectively, f is the activation function and a is the output of the neuron. The activation function used here is ReLU. This RelU can able to reduce overfitting and produces better generalization performance on unseen data of the network.

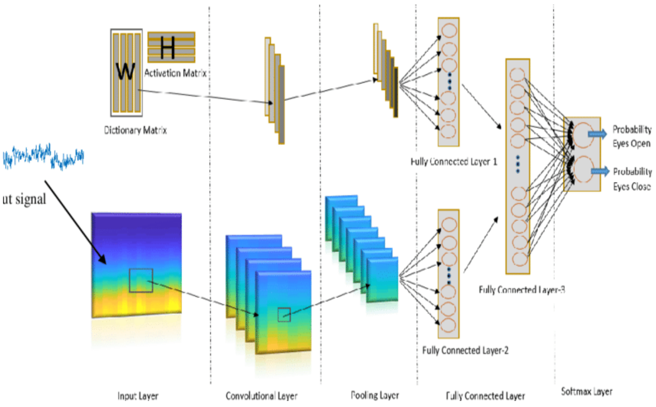


Fig. 3 Gurve, Dharmendra & Krishnan, Sridhar.(2018). Deep Learning of EEG Time-Frequency Representations for Identifying Eye States. Advances in Data Science and Adaptive Analysis accessed 10 January 2023, <https://www.researchgate.net/publication/325126184>

1. Input layer: This layer receives either time-frequency or spectrogram representation of the audio signal. This layer is reshaped to have a 3D structure in which the first dimension corresponds to a number of samples, the second dimension corresponds to a number of frequency bins and the third dimension corresponds to a number of channels.

2. Convolutional layer: The set of layers that can be represented after the input layer, in which each layer different filters are applied to input data to extract important features and produce feature maps. The filters or kernels can slide over the input data based on the stride by computing the dot product between filter weights and input data, if requires use padding also. Then the outputs can be stacked to form a feature map. The proposed model is based on Conv1D and Conv2D. Conv1D is used to process the data by taking input as a 1D array in the form of time series signal features whereas Conv2D is used to process grid-like structures like images. The input can be taken as a 2D matrix in which each element corresponds to a pixel from the obtained image.

3. Pooling layer: These layers are used to reduce spatial dimension from the obtained feature map while retaining the most important characteristics. Max pooling is mostly used in CNN for getting maximum pixel features by using stride from the activation layer mainly used is the ReLU layer.

4. ReLU Layer: It is used as an activation function for neural networks. This layer brings non-linearity in the neural networks, for which it is important to learn relationships between input layers and output layers. This function states that If the input x is greater than 0, then the output will be x otherwise 0. It can be mathematically defined as follows:

$$F(x)= \max(0, \ x) \qquad (3)$$

5. Fully connected layer: After that, the feature maps are flattened and passed through one or more fully connected layers. In these layers, we use some of the activation layers like ReLU. From that, we can connect to the output layer.

6. Output layer: This layer provides the prediction for the audio classification task. In this, we use the ReLU activation function for the classification of multiple classes.

3.4.2 Model Deployment:

We will choose a deep neural network model based on the model optimization and testing which involves the existing hyperparameters of the models and the performance of the model based on accuracy then make a necessary adjustment to the model based on the results of the testing then the model can be saved in a format and deployed into cloud platform or some other for the real-time classification of environmental sounds. This mainly involves choosing a model architecture, fine-tuning hyperparameters of the model to the specific task then testing and saving the optimized model. Finally, we developed an API that allows users to make predictions with respect to the deployed model.

3.5 WORKING STEPS FOLLOWED BY THE PROPOSED SYSTEM

1. The entire model can be saved and created in the Flask environment to run the server and allow access to be deployed.

2. Select an audio clip as input to the model which is compatible with the deployed model.

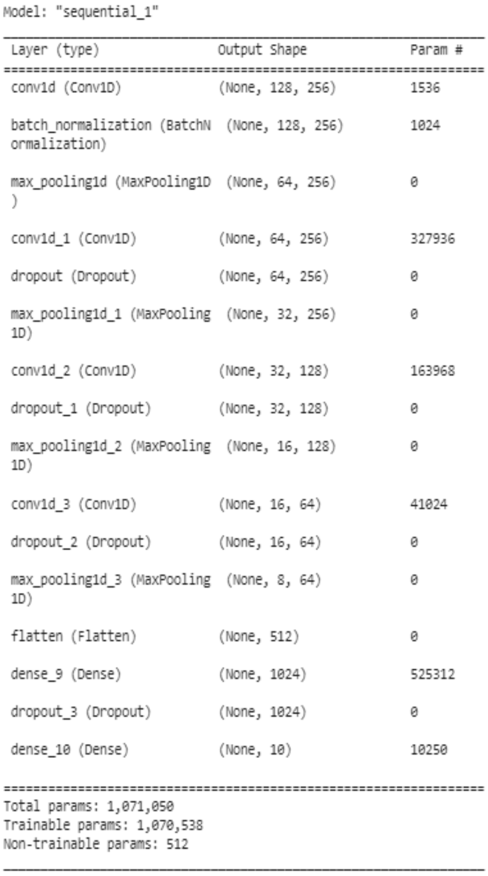
3. Generate an API request to deployed model that request can be made using a web client for accessing the request.

4. The model can receive the input and perform inference using the trained model to make a correct prediction about the class of the sound. The prediction class will be labeled class.
5. The API will return the response as the predicted output to the request sent by the user.

IV. EXPERIMENTS AND RESULTS

4.1 EXPERIMENT SETUP

We evaluate the performance of our model by training the different deep neural network models and trained them by the dataset discussed earlier UrbanSound8K. After training each model then evaluate and compare the models and select the model which gives good accuracy among them and satisfies all constraints according to the requirements.



```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d (Conv1D)             (None, 128, 256)          1536

 batch_normalization (BatchN  (None, 128, 256)         1024
 ormalization)

 max_pooling1d (MaxPooling1D  (None, 64, 256)          0
 )

 conv1d_1 (Conv1D)           (None, 64, 256)           327936

 dropout (Dropout)           (None, 64, 256)           0

 max_pooling1d_1 (MaxPooling  (None, 32, 256)          0
 1D)

 conv1d_2 (Conv1D)           (None, 32, 128)           163968

 dropout_1 (Dropout)         (None, 32, 128)           0

 max_pooling1d_2 (MaxPooling  (None, 16, 128)          0
 1D)

 conv1d_3 (Conv1D)           (None, 16, 64)            41024

 dropout_2 (Dropout)         (None, 16, 64)            0

 max_pooling1d_3 (MaxPooling  (None, 8, 64)            0
 1D)

 flatten (Flatten)           (None, 512)               0

 dense_9 (Dense)             (None, 1024)              525312

 dropout_3 (Dropout)         (None, 1024)              0

 dense_10 (Dense)            (None, 10)                10250

=================================================================
Total params: 1,071,050
Trainable params: 1,070,538
Non-trainable params: 512
_________________________________________________________________
```

Fig. 4 Implementation Section

The above result is a summary of a Sequential neural network with multiple layers for a 1D convolutional neural network (CNN). The model takes input as a sequence of data, by applying a series of convolutions, pooling, and dropout layers, and finally gives a prediction. Here the trainable parameters in the model are equal to 1,070,538.

4.2 EXPERIMENT RESULTS

This was implemented on the three models: ANN, Conv1D, and Conv2d. The selection of ANN is mainly because in audio classification extracting features is complex and also the dataset (UrbanSoundDataset8k) used for the model gives good accuracy about 93.07% based on the comparison analysis which takes MFCCs features as inputs in a 1-D array and makes some hidden layers then connect to the output layer for the prediction of the class but it did not satisfy certain dimensions constraints allowing that model to produce some drawbacks. Then, building the model using CNN with certain layers and accepting the multi-dimensional arrays then gives 90.73% accuracy and also overcomes the drawbacks faced at the time of implementation of ANN and emerged as the best working model for the classification of environmental sounds. There are several simulation tools used for audio classification. The proposed model uses TensorFlow, Keras, and Librosa.

Table 1: Accuracy for each model

MODELS	ACCURACY
ANN	93.07
Conv1D	90.58
Conv2D	90.73

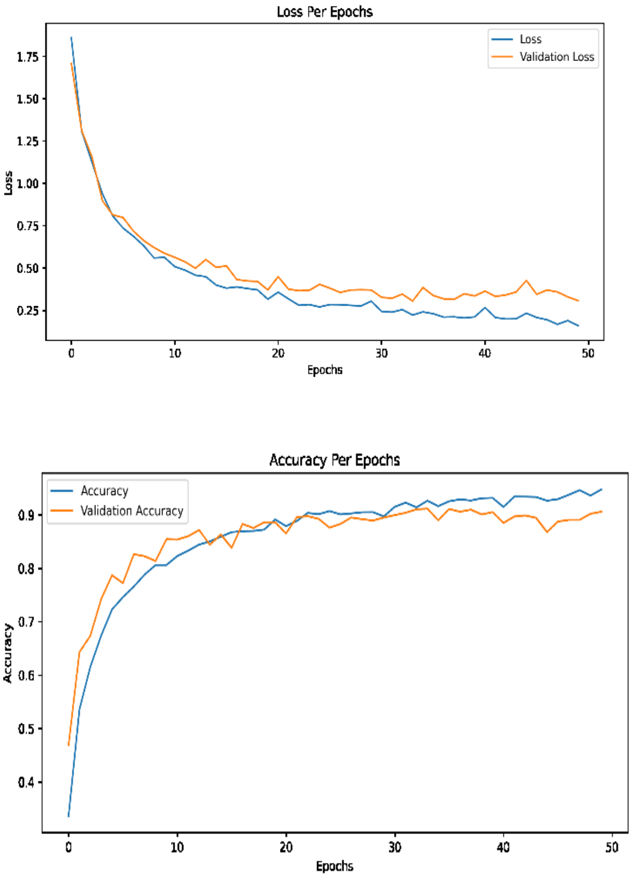


Fig. 5 Graph Analysis

The target labels in the dataset can be labeled as 0 to 9 for each class to be represented by using Label Encoder and that can be used at the time of comparison analysis between the actual and predicted classes. With this, we can easily measure how much loss can be occurred so that, it reduces in the result. The dataset is huge so high accuracy was achieved because of extracting more MFCC features and adding more layers for training the model. By using confusion metrics, we can also measure the F1-score, precision, sensitivity, and specificity for selecting the good model.

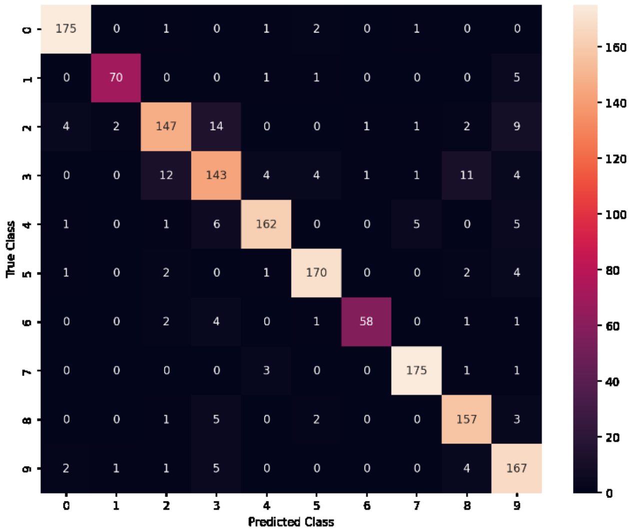


Fig. 5 Confusion Metrics

While training the neural networks one needs to specify the number of neurons in the hidden layers, and epochs to train the model all of those are called hyperparameters. The hyperparameters used in this model are learning rate which is taken as 0.001, number of epochs considered as 50, batch size as 32, total layers considered as 10, and used the Adam optimizer.

V. CONCLUSION

In this paper, after completion of the implementation on different models, CNN is the proposed model which can satisfy the requirements and allowed the users to classify any audio-based application as like environment sound classification. This CNN can be allowed to build with many attention-based models according to the application of the user. This model can give the prediction for the given audio clip and show it, output class, to the user after the deployment of the optimized model. In this way, audio classification for the environmental sounds can vary the different audios based on the time and frequency domains.

VI FUTURE SCOPE

We want to expand our work by creating a hybrid model by using another deep learning model probably the LSTM model in addition to the previously suggested model information which can possibly give more accuracy by extracting some other important features that we previously missed extracting from there. This can possibly high chances to give correct predictions more frequently by doing this method in future experiments.

VII REFERENCES

[1] Wu, B., & Zhang, X. P. (2021). Environmental sound classification via time–frequency attention and framewise self-attention-based deep neural networks. *IEEE Internet of Things Journal*, *9*(5), 3416-3428.

[2] Yu, C., Barsim, K. S., Kong, Q., & Yang, B. (2018). Multi-level attention model for weakly supervised audio classification. *arXiv preprint arXiv:1803.02353*.

[3] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., ... & Wilson, K. (2017, March). CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 131-135). IEEE.

[4] Aksoy, B., Uygar, U. S. T. A., Karadağ, G., Kaya, A. R., & Melek, Ö. M. Ü. R. (2022). Classification of Environmental Sounds with Deep Learning. *Advances in Artificial Intelligence Research*, *2*(1), 20-28.

[5] Bansal, A., & Garg, N. K. (2022). Environmental Sound Classification: A descriptive review of the literature. *Intelligent Systems with Applications*, 200115.

[6] Zhang, Z., Xu, S., Zhang, S., Qiao, T., & Cao, S. (2021). Attention based convolutional recurrent neural network for environmental sound classification. *Neurocomputing*, *453*, 896-903.

[7] Kaur, B., & Singh, J. (2021). Audio Classification: Environmental sounds classification.

[8] Mu, W., Yin, B., Huang, X., Xu, J., & Du, Z. (2021). Environmental sound classification using temporal-frequency attention based convolutional neural network. *Scientific Reports*, *11*(1), 21552.

[9] Walden, F., Dasgupta, S., Rahman, M., & Islam, M. (2022). Improving the Environmental Perception of Autonomous Vehicles using Deep Learning-based Audio Classification. *arXiv preprint arXiv:2209.04075*.

[10] Bahmei, B., Birmingham, E., & Arzanpour, S. (2022). CNN-RNN and data augmentation using deep convolutional generative adversarial network for environmental sound classification. *IEEE Signal Processing Letters*, *29*, 682-686.

[11] Medhat, F., Chesmore, D., & Robinson, J. (2020). Masked conditional neural networks for sound classification. *Applied Soft Computing*, *90*, 106073.

[12] Mang, L. D., Canadas-Quesada, F. J., Carabias-Orti, J. J., Combarro, E. F., & Ranilla, J. (2023). Cochleogram-based adventitious sounds classification using convolutional neural networks. *Biomedical Signal Processing and Control*, *82*, 104555.

[13] Uzkent, B., Barkana, B. D., & Cevikalp, H. (2012). Non-speech environmental sound classification using SVMs with a new set of features. *International Journal of Innovative Computing, Information and Control*, *8*(5), 3511-3524.

[14] Khamparia, A., Gupta, D., Nguyen, N. G., Khanna, A., Pandey, B., & Tiwari, P. (2019). Sound classification using convolutional neural network and tensor deep stacking network. *IEEE Access*, *7*, 7717-7727.

[15] Cances, L., Labbé, E., & Pellegrini, T. (2022). Comparison of semi-supervised deep learning algorithms for audio classification. *EURASIP Journal on Audio, Speech, and Music Processing*, *2022*(1), 23.