Performance Evaluation of Environmental Sound Classification: A Machine Learning Stacking and Multi-Criteria Metrics Based Approach

Kashif Iqbal, Amir Ul Din, Affan Alim*, Umer Bin Sadiq, Saleem Ahmed

College of Computing and Information Sciences, KIET, Karachi, Pakistan
*Corresponding author: affanalim@kiet.edu.pk

Abstract

This study proposes an Environment Sound Classification Task (ESC) model that includes numerous element channels given as a contribution to Machine learning with an Attention instrument. ESC is a significant testing issue. The interest in the paper lies in utilizing different part channels involving the MFCCs-Mel Frequency Cepstral Coefficients a mutual module in speaker detection and artificial speech systems. LPCs-Linear Prediction Coefficients and Linear Prediction Cepstral Coefficients were the most commonly used types in ASR- Automated speech recognition. The paper also discusses some basic features of MFCCs and how to put them into practice. The techniques used for this project are Background Gaussian Noise and Time Shifting, by observing that every technique is implemented with a provided probability, however, when at the time of generation of a new sample the same spectrogram input may have several combinations. We go through the blend information expansion method to additional lift execution. Our model can accomplish cutting-edge execution on three benchmark climate sound characterization datasets, for example, the UrbanSound8K.

Keywords—Environmental sound classification, images of the spectrogram, machine learning model, deep learning model, features of sound

1 Introduction

In One of the most important applications of Machine Learning is Environment Sound Classification which is concerned with detecting sounds in their surroundings. It's a difficult assignment that entails categorizing different sounds into one of several categories, such as police vehicle or ambulance sirens, animal sounds, aircraft, conversation sounds, and so on. The effective ESC models entail more than one standard audio feature extraction procedure and DNN (deep neural networks). We discuss extraction techniques with features different like the MFCC-Mel-frequency Cepstral Coefficients [1], GFCC-Gammatone Frequency Cepstral Coefficients, CQT-Constant Q-Transform and Chromagram. The first stage of an automatic speech recognition system is feature extraction by the identification of different parts of an audio signal. The system helps us in filtering linguistic content from useless sounds such as background noise. The feature extraction stage is followed by the classification stage.

The sound produced by human beings and animals is determined by the shape of their vocal tract. Once we know the exact shape of the vocal region, we will be able to represent the sound signals being produced from the vocal region. A short-time power spectrum is used to represent the shape of the vocal tract using MFCC. MFCCs-Mel Frequency Cepstral Coefficients are often used nowadays to automatically detect speech. Before MFCCs became the norm, Linear Prediction Coefficients (LPC's) and Linear Prediction Cepstral Coefficients (LPC's)were the two dominant features used for automated speech recognition [2]. In this paper, we will discuss some main features of MFCCs and their significance for automated speech recognition systems.

The Max-polling domains were used only one at a time. After the convolutional layer, they combine time and frequency, independently combining the future stage and enabling the processing time and fre-

ISSN: 2523-0379 (Online), ISSN: 1605-8607 (Print)
DOI: https://doi.org/10.52584/QRJ.2101.10

quency. We design a module that enables both channel attention and spatial similar modules, requiring an attention-weight-matrix through dimensions equal to the block output of DCNN, therefore each feature map of output in the individual channel has its attention weight. In this module, we used the depth-wise separable convolution to attain a nominal increase in the number of parameters [3]. These elements were demonstrated with some normal Machine Learning calculations. The performance gives a satisfying introduction reason provides traditional and feature extraction ability. The rest of the paper is distributed as follows: The related works are described in section 2 followed by the detailed methodology along with the proposed model is presented in section 3. An experimental analysis is discussed in 4, the result and discussion part is defined in 5, future enhancement of this research is discussed in 6 and finally, the paper is concluded in section 7.

2 Related Work

Multiple approaches have been devised and proposed for sound classification, in [4], the author proposed a deep learning (state-of-the-art) solution by using techniques for audio feature extraction in order to develop multiple channel input for deep learning (CNN), i.e MFCC- Mel-frequency-cepstral-coefficients, GFCC-gammatone-frequency-cepstral-coefficients, constant chromagram and Q-Transform, and chromagram. These feature extraction techniques help CNN classifier by creating multi-channel input with an accuracy of approximately 90%. However, CNN is widely used for image classification purposes, but it can be applied to sound classification [5] with Tensor Deep Stacking Network (TDSN). The study shows that it can be applied in critical areas. This approach is evaluated with promising results. TDSN is another extension of DSN - Deep Stacking Network. Some other [6] methods have also been applied on the datasets(audio files) to get the different acoustic patterns, for example, mel-spectrogram (MS), harmonic-part (HC) based spectrogram, and percussive-part (PC) based spectrogram.This approach used sub-spectral net for the bird sound classification and also used class-based late fusion.

The research has also proved that CNN with a mixup technique on the dataset has outperformed the VGG10 Net [7]. Mixup was used with CNN for more accuracy on the datasets UrbanSound8k, ESC50 and ESC10. However, mixup and feature space distribution were implemented explored, and found that mixup is much better for the accuracy improvement. Due to various sound classes in universal sound separation, the uncertainty to identify what types of sources are available is

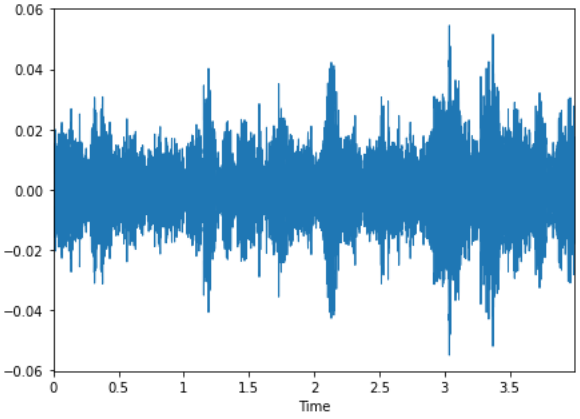

Fig. 1: Audio File of Data

quite high. It can be reduced [8] by the implementation of conditioning on the prediction of sound classifiers. In this approach, the sources are repetitively separated, categorized, and sent to the next separation stage. In the classification, feature extraction plays a vital role accuracy of sound events [9] deep features are extracted with the help of the deep learning (CNN) model on UrbanSound8k and DCASE-2017 datasets. Deep feature vectors are produced for robust environmental sound classification in spite of using soft-max and layers of classification.

Studies have proposed that Mel-frequency spectral coefficients provide better results in heart disease diagnosis [10] such as associated to time domain and Short-Time-Fourier-Transform based Features. CNN with parallel pooling structure [11] can be used to diagnose the pulmonary disorder by using lung sound classification. 1D CNN trained directly from the audio waveform [12] provides more promising results than 2D CNN, however, if both 1D and 2D CNN are combined then better and more accurate results can be achieved. CNN with Mel Spectrogram [13] is used on UrbanSound8k and ESC-10 datasets with an accuracy of 99.49%.

3 Methodology

3.1 Process Overview

3.1.1 Feature Engineering

In this section we do some feature engineering before building any model, we not only need a dataset, but we want a usable representation for better results. An Audio File is represented as a time series with the dependent axis being the amplitude of the audio waveform. The audio files (waveform) contain all information in which we create features to train our model. However, the shape of a waveform does not

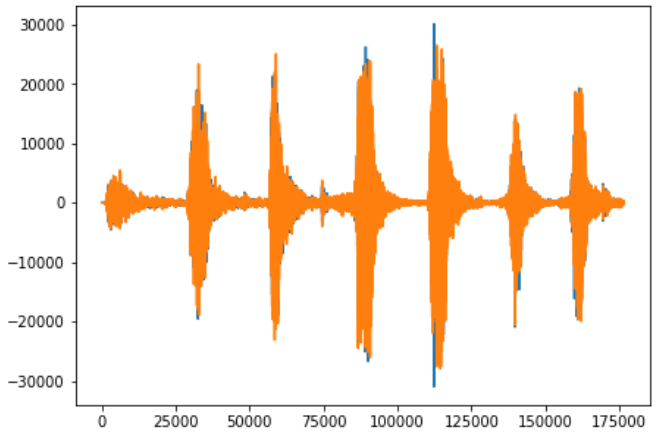

Fig. 2: Audio File with 2 Channels

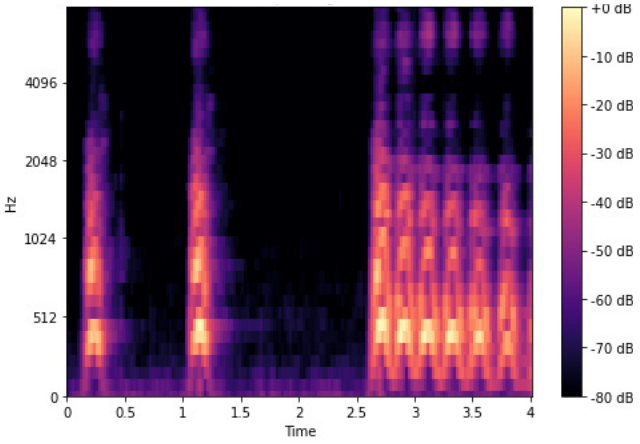

Fig. 4: Mel Spectrogram

carry enough discriminating information, so we need to transform the waves into a more usable form. Though sometimes the data-set consists of enough information to train the model to produce more appropriate outcomes, even then input data should be scrutinized and transformed - this allows us to extract the best model for our features. This is usually much better and different than a model that works on our raw input data. We constantly clean the dataset before feature engineering, focusing on carefully removing outliers, removing extraneous input data in accordance with business logic, or removing noise. For audio, data cleaning may consist of making audio samples equal in length and padded with silence at either end.

3.1.2 *Data Augmentation*

It is a technique that is used for the compensation purpose for the relatively smaller dataset(s), different types of Augmentation techniques have been used in the proposed model. It is accomplished online by applying the techniques to the info-spectrograms prior to them being placed in the respective model [13]. To guarantee that the proposed model is prepared on similar models these strategies are applied with given probabilities.

The two important techniques utilized for this project are explained as follows. i- Background Gaussian-Noise (G-N): By this technique, background white noise is an add-on to the spectrogram. ii- Time Shifting: The provided Spectrogram is shifted to the right by applying this technique - a section of the spectrogram is shifted out of the fixed length of the edge.

Taking into account that every single strategy is carried out within a provided probability, at the time of generation of another sample it is conceivable that the same spectrogram may contain multiple combinations.

3.1.3 *Librosa*

For this experiment, we used Librosa for data preprocessing and feature extraction. Sometimes pre-

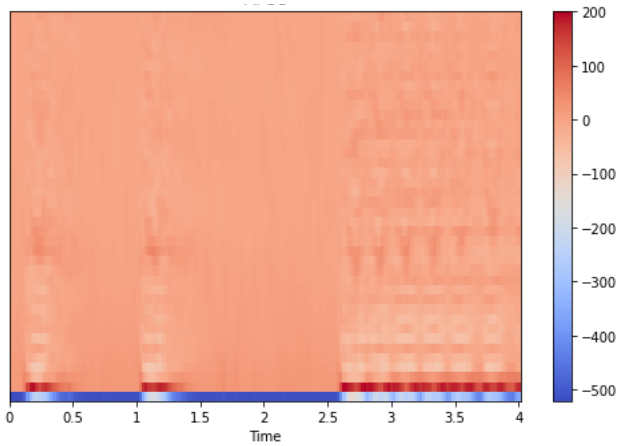

Fig. 3: Mel Frequency Cepstrum Coefficients

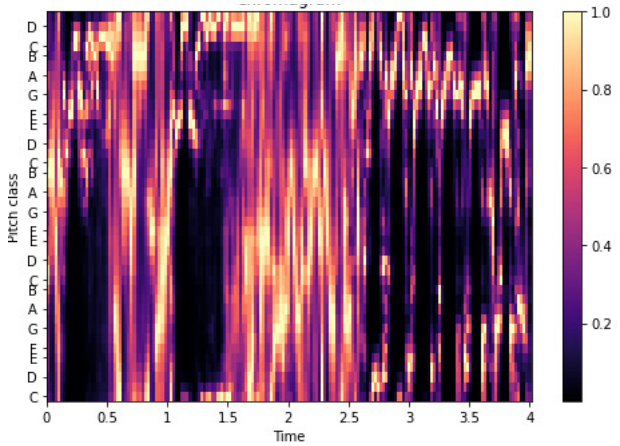

Fig. 5: Chromagram Feature

processing become necessary to increase the accuracy, so Librosa library can read audio files which can be shown in Figure 1 furthermore, convert them to their sufficiency values for each example of audio. Let us say there is a sound document of 4s and the inspecting pace of the sound record is 22050 Hz. It shows that a single audio file is made by using samples of amplitudes such that in each second 22050 samples of amplitudes are recorded and can be converted into original audio with 2 channels which could be shown in Figure **??**. Therefore in audio files (4 Sec) which contain 22050 is sampling rate could be expressed in the form of an array size of $4 \times 22050 = 88200$.

MEL Features

i. MFCC

In SRS (Speech Recognition System), the initial phase is to extract features that can recognize the components of the audio signal that are worthwhile to identify the content of linguistics and also the removal of completely other unnecessary stuff which brings information similar to noise and motion of background etc.

One of the essential elements to comprehend speech is that the sounds produced by humans are cleared by the form of uttered (Vocal) zone which contains the teeth, tongue, etc. This form of shape defines what sound comes out. We could achieve an accurate depiction of the phoneme that is being produced once we could accurately extract the shape. The formation of the vocal region displays itself in the packet of the little (short-time) power- spectrum, this envelope is represented accurately by M.F.C.Cs which is the main job of MFCCs. [14].

MFCCs (Mel-Frequency-Cepstral-Coefficients) is a characteristic that is vastly utilized in automatic speech and speaker identification as shown in Figure 3. In deriving our M.F.C.Cs, we have also produced an additional feature we can make use of. When we mapped the frequencies of a power spectrogram to the Mel scale, we produced a Mel Frequency Spectrogram - a simple analog of the power spectrogram with the frequency scale in Mel. We're going to use the Mel Specotrgrogram as a feature of its own.

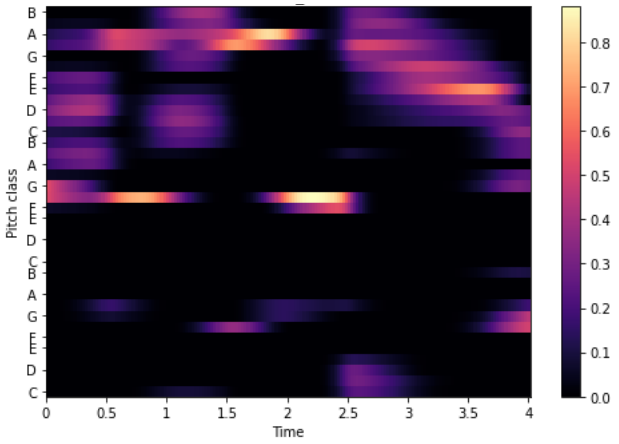

Fig. 7: Chromagram CENS

ii. Mel Spectrograms and Mel-Frequency Cepstrums

In deriving our M.F.C.Cs, we have also produced an additional feature we can make use of. When we mapped the frequencies of a power spectrogram to the mel scale, we produced a Mel Frequency Spectrogram - a simple analog of the power spectrogram with the frequency scale in mels [15]. We're going to use the Mel Specotrgrogram as a feature of its own which could be depicted in Figure **??**.

What about cepstrum, when we took the log of the spectral amplitudes in a Mel-scaled power

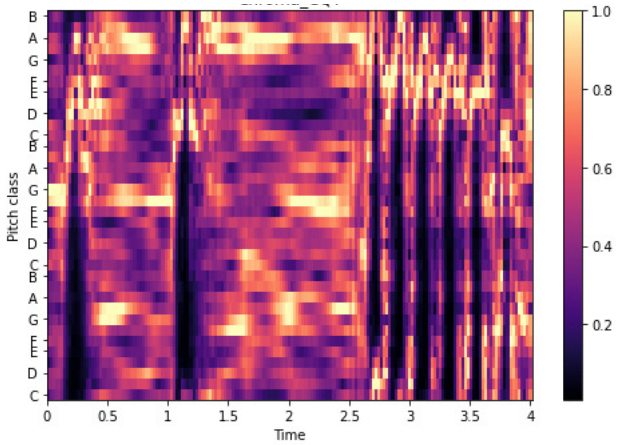

Fig. 6: CQT Chromagram

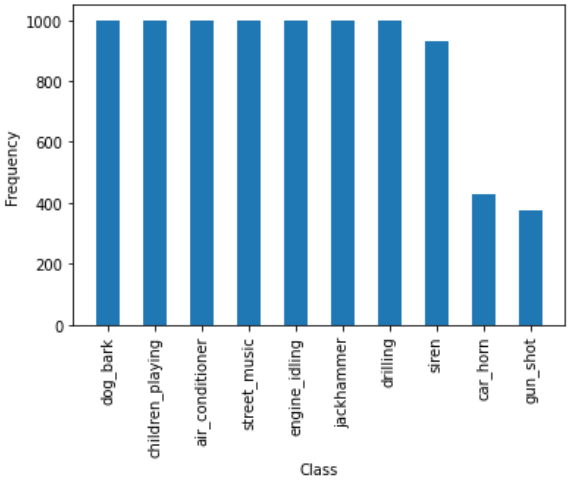

Fig. 8: Frequency of Class Distribution

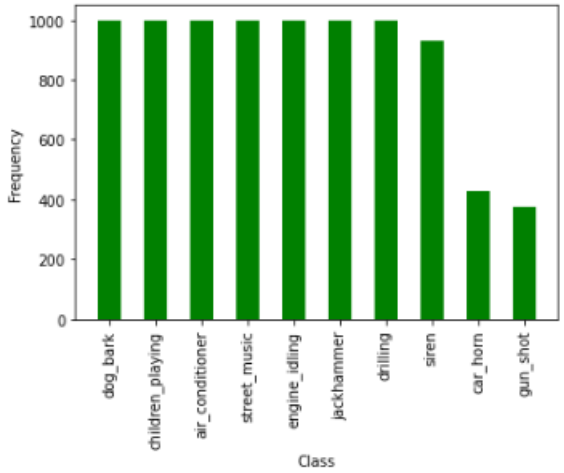

Fig. 9: Class Frequency Distribution

spectrogram, we could have plotted that result as a time series and produced what is known as a 'Mel-frequency cepstrum', now in the 'quefrency' domain, so confusingly named because we have applied a transformation to an audio signal in its frequency domain, but have plotted it as a time series - not quite in the frequency domain. In the cepstrum, we observe that wherever there is a periodic element in the original audio signal. We've already used the Mel-frequency cepstrum (MFC) as a feature since that is precisely the source of our MFC coefficients.

iii. Chroma Features

If we consider the voice or music, the terminology Chroma feature or chromagram highly relates to the twelve different classes of the pitch. Features based on Chroma, termed as 'pitch-class-profiles', a mighty tool for the analysis of music whose pitches can be meaningfully classified (mostly into 12 classes) and tuning of them approximates to the equally-tempered scale which could be shown in Figure 5. One major property of chroma features is, they capture the harmonic and the melodic properties of music, due to robustness against changes in timbre as well as instrumentation

iv. Chroma Stft

Either from a power spectrogram or a waveform. The chroma variations have three types, i.e., chroma-stft, chroma-cqt, and chroma-cens. The very first two of these (chroma stft and chroma cqt) are the two identified approaches that are better to draw chroma. This technique executes the audio sample as input and plots each S.T.F.T bin to chroma using chroma-stft [16].

v. Chroma CQT

Constant-Q-chromagram. On this side, it transforms and portrays each cq-Bin to chroma using Chroma CQT and constant-q which can be visible in Figure 6.

vi. Chroma Cens

The chroma variant 'Chroma Energy Normalized' (CENS) is shown in Figure 7. Chroma has three variants applied in Librosa: chroma-s.t.f.t, chroma-c.q.t, and chroma-cens. Chroma-s.t.f.t and chroma-c.q.t are two substitute methods to plot/draw chroma. A short-time Fourier transform is applied by Chroma-stft of the audio input and plots each s.t.f.t bin to chroma, on the other hand, chroma-c.q.t utilizes constant-Q transform and portrays each cq-bin to chroma.

3.2 Experimental Work

3.2.1 Dataset

In this section popular datasets (UrbanSound8K) [17] are used to evaluate sound classification. It is very important to understand the dataset structure which you are using to train the model that how to excerpt features from them. Urban-Sound-8K-datase is categorized following 10 classes and class frequency distribution is shown in Figure 8.

i. Jackhammer, ii. Gun Shot, iii. Car Horn, iv. Dog bark, v. Drilling, vi. Engine idling, vii. Air Conditioner, viii. Siren, ix. Street Music , and Children Playing

This dataset contains 8732 labeled sound sections which are up to 4sec of record duration and the audio files are recorded with a 22.05 kHz sampled frequency. All sections have to be taken from field recordings which would be uploaded on free sound. The classification of these files is sorted into multiple folds (10 classes) which include (folder name from fold1 to fold10), it helps in the comparison and training with the auto-classification results. Moreover,

TABLE 1: Comparative model with ML and DL classifier accuracy and loss

sno. no.	Classifier Name	Accuracy Score	Loss
1	KNN	84.04%	1.8194
2	MLP	83.34%	0.6865
3	CNN	79.51%	0.6908
4	QDA	77.33%	3.165
5	Decision Tree	70.29%	10.39
6	SVC (Linear)	65.66%	1.072
7	SVC (RBF)	61.19%	1.1448
8	Gaussian NB	48.08%	3.165
9	Adaboost	45.62%	2.1899
10	Proposed Model (RF)	88.26%	0.6295

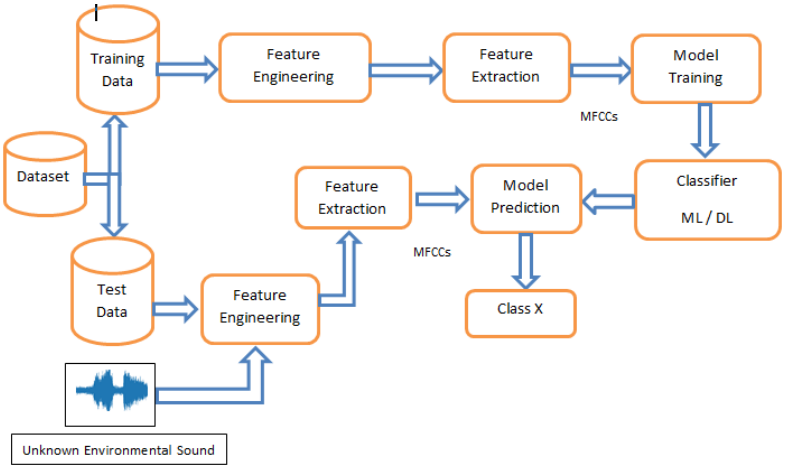

Fig. 10: A detailed process diagram of this study

to the sound segments, a comma-separated values (CSV) file comprises metadata regarding every single section and this file contains various columns such as filled, label, classid corresponding to label, salience, etc, and the Audio files are in ".wav format" and the sampling-rate, bit-depth, and as of the original files uploaded on the Free sound the count of channels is similar.

3.2.2 Hardware and software Used

In order to perform this experiment, we implied the Lenovo-Legion-81-F.V-Laptop. All the specifications of the mentioned system are as follows; Intel Core-i7 with 6 cores, and 12 logical processors were used with the Graphics card of Nvidia-GeForce-GTX-1080 consists 8-GB GDDR 5-X V-RAM, and the total memory of this system is 16GB.

3.2.3 Software Implemented

Multiple software (python, pandas, Keras, TensorFlow, seaborn, NumPy, Libros, scifi etc) were used for the completion of this experiment to develop and teach ML & ANN model.

3.2.4 Google CoLab

Mainly we used CoLab for training the model to train the dataset for extracting the result. This has been done by utilizing 'spectrogram ()', the built-in function of the CoLab to generate a spectrogram of the audio signal. These generated spectrograms were saved using the save as (GCF, name. format) function over a loop iteration equal to the number of samples present in the dataset.

3.2.5 KERAS

It is an API [18] provided by TensorFlow which is exactly devised to strengthen deep- neural-network architectures. Keras (TensorFlow) is implied in the research to develop the proposed model. Keras covers several optimizers and activations which could be utilized quite effectively in the proposed model.

3.3 Comparative Models

3.3.1 KNN

The K-nearest Neighbour abbreviated as K-NN Algorithm is proposed to seek the amount of k training samples that are closest to a mark object within the provided training set. Additionally, KNN identifies the dominant class from the k training samples; then, assigns this dominant class to the mark object, where k is the number of coaching samples. Therefore, the overall mechanism of the K-NN algorithm is that every sample must have identical attributes while they're classified within the exact category in a very feature space, in which the type contains the k most neighboring samples [19].

3.3.2 Adaboost

Boosting algorithms are not classifiers but they can be used with pair of any classifier i.e. Random Forest or SVN. AdaBoost provides a given feeble or base learning calculation multiple times in a progression of rounds $t = 1...T$. One of the primary ideas of the calculation is to save a circulation or set of loads over the preparation set. The heaviness of this dissemination [20].

3.3.3 SVM Linear

The SVM algorithm's prime goal is to find out a hyperplane in an n-dimensional space (where n represents the total number of features) that segregates data points. However, countless hyperplanes are selected to split the two groups of data points. Finding a plane with the greatest margin is our aim or the longest distance between data points from both classes. Some reinforcement is obtained by maximizing the margin distance, which makes it more manageable to classify the future data points. [21].

3.3.4 SVM RBF

SVM with RBF Kernel is an AI calculation which is fit to group information focuses isolated with outspread based shapes like this: And that capacity in AI is astounding because it would be able "embrace" the information focuses intently, definitively isolating them out [22].

3.3.5 MLP

The multi-layer perceptron (MLP) works on feed-forward augmentation, it is a part of a neural network. It contains three layers, 1) the input layer, 2) the output layer, and 3) concealed layer. The input layer is responsible to pass and receive the input signal, however, forecast and categorization tasks are the responsibility of the output layer. The random number of hidden layers performs the role of the real computational engine of the MLP, which is inserted between the input and output layers. Likewise, in a feed-forward network, the data flow between the input and the output layers is in the forwarding direction. The neurons in the MLP are trained by the backpropagation learning algorithm. Besides comparing any continuous function the MLPs are meant to handle non-linear issues [23]

3.3.6 Decision Tree

The supervised ML (machine learning) methods contain the DTA (decision tree algorithm). It has the ability to resolve individual regression and classification problems. The primary purpose of this methodology is to predict the value of a target class variable by improved ML model, for which the DT (decision tree) consumes the representation of the tree to resolve the problems; the class label can be represented through the leaf node and the attributes are represented by interior node [24].

3.3.7 Gaussian NB

GNB (Gaussian Naive Bayes) only takes continuous data-valued features and then models them entirely as GND (Gaussian Normal Distributions). For building a simple GNB model, it assumes that the data features with no co-variance (Independent Dimensions) could be characterized among the parameters [25].

3.3.8 Quadratic Discriminant Analysis

In Quadratic and Linear discriminant analysis, the only difference is that we comfortably assumed that all the mean and covariance of classes were entirely equal. Through this, we would independently have to calculate a result [26].

3.4 Proposed Model

Our proposed model is defined in this section which is used for sound classification, mainly we used UrbanSound8K dataset for data preprocessing and feature extraction The MS (Mel-spectrograms), MSRF (Mel-spectrogram-related feature) sets and the sound literature classification have been largely applied in machine learning and deep learning (ANN, MLP) models which show more accuracy for sound classification, we used MFCC (Mel-Frequency Cepstral Coefficients) from the audio samples. The frequency distribution of the MFCC summarizes throughout the window size, therefore it is likely to analyze individually time features of the sound and the frequency.

For extracting the features we have applied Librosa to extract the MS (Mel- spectrograms), which is named Feature 1, with 128 Mel Filters banks which are ranging from 0 to $80000Hz$. After extracting the features we create the data frame which holds multiple classes and then do a class distribution which could be shown in above Figure 9. After that, we split the dataset into an independent and dependent dataset and then pass for training with a ratio of 0.8, and the process of evaluation is performed through the remaining part of the complete dataset, we used multiple classifiers of Machine Learning and deep learning which are listed below and process of the proposed model could be shown in Figure 10.

TABLE 2: The Impact of Cross Validation on the Performance of the Proposed Model

Dataset Name	CV=5	CV=10	CV=15
Urbansound8k	88%	89%	90%

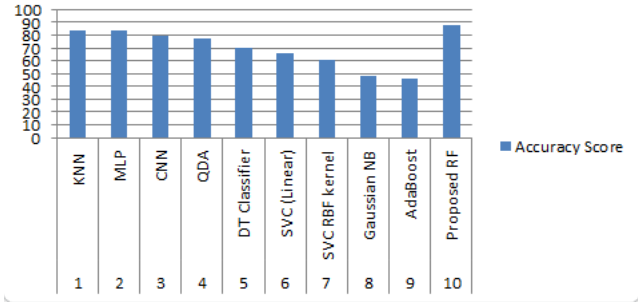

Fig. 11: Accuracy Score Comparison

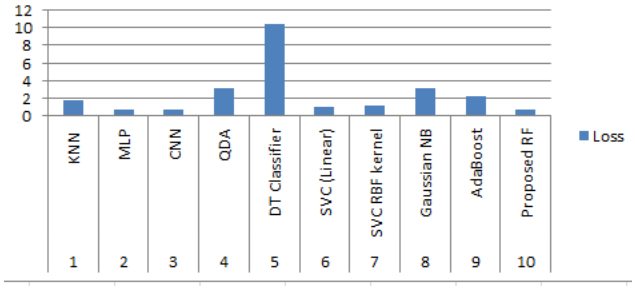

Fig. 12: Loss Comparison

4 Experimental Analysis

4.1 Evaluation Method

UrbanSound8K dataset is to be used for the proposed Machine Learning (Random Forest) and comparatively, with Deep Learning (ANN and CNN) training and the processes of evaluation, the selected dataset with a ratio of 0.8 is divided randomly for the proposed Machine Learning and Deep Learning (ANN and CNN) for the process of training, in addition, the evaluation process has to be performed on the remaining part of the dataset. On the dataset (Urbansound8k), the performance of classification is to be tested. In this paper, the main criteria of evaluation comprise Accuracy, Specificity, Precision, Sensitivity, Log-Loss, Hamming-Score, and Jaccard-Score. These mentioned criteria could be calculated through confusion matrix values and with a classification report as given below the next equations.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$F1score = \frac{2 \times Pre \times Sen}{Pre + Sen} \tag{5}$$

where
Pre = Precision and Sen = Sensitivity

$$HLoss = average(ytrue \times (1 - ypred) + (1 - ytrue) \times ypred) \tag{6}$$

$$LogLoss = y \times (log(ypred) + (1 - y) + log(1 - ypred)) \tag{7}$$

$$JccardIndex = \frac{|ynypred|}{|y| + |ypred| - |ynypred|} \tag{8}$$

4.2 Convolutional Neural Network

In this paper, we applied the CNN model for comparative analysis by using the Keras library, i.e. executed over TensorFlow. The CNN work by maintaining a 2-layered deep architecture with a fully connected layer while one for the output prediction layer. The first CNN layer contains 32 filters of 3×3 size with ReLU activation that shows 32 attribute maps which are generated by a middle layer [27].

5 Experimental Results and Discussion

Table 1 depicts the comparison of different methods' accuracy scores on the dataset including ML and DL. Undoubtedly, Random Forest is the winner with the highest accuracy of 88.26%. Figure 10 is the graphical representation of Table 1. Table 1 displays a comparison of loss against each classifier and the proposed method, i.e., Random Forest, has the lowest loss - 0.6295, while Figure 11 is the graphical comparison of all the classifiers of ML and DL with the proposed method.

Table 2 shows the contribution in the proposed model, we increased the cross-validation up to 15 and got the distinct results, CV with value 5 the performance is 88%, CV with value 10 the performance is 89%, and CV with value 15 the performance is 90%. However, Table 3 shows Hamming Loss comparison with other ML & DL models and again Random Forest is an obvious winner with the lest value of 0.1110, the same comparison is shown in Figure 12.

Table 4 is the presentation of the Jaccard Score comparison of the other ML & DL models with the proposed model, Random Forest has the highest value of 0.8001. In Table 5, the other performance criteria for the dataset are shown and the proposed model has a better score in precision, recall, and f-1 score

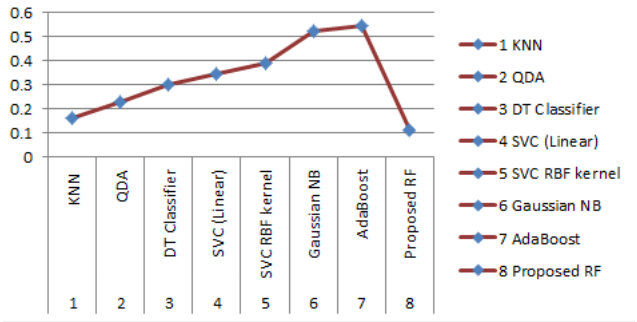

Fig. 13: Hamming Loss Comparison

as compared with the rest of the models. A pie chart is depicted in Figure 13 to show the graphical form of Table 4.

6 Future work

Future work aims to enhance the existing work in a more efficient manner to use it in health sciences. Different diseases such as Tuberculosis, Heart, COVID, and so on, can be diagnosed by implementing the proposed techniques. However, bird species can be identified with the proposed techniques, if we add more bird sounds then this will be very helpful. The same model can be used efficiently and effectively in pulmonary disorders.

7 Conclusion

In this paper, we evaluate the use of CNN and multiple ML models to classify the sound signal using spectrograms of the sound spectrum. This paper shows that we achieve more accuracy in Random Forest with the help of boosting algorithm instead of deep neural architecture. The approach we are using is slightly different instead of direct sound classification first we do the data pre-processing and augmentation using spectrogram which uses feature extraction using librosa which reduced the trainable parameters. In the

TABLE 3: Comparison of Hamming Loss with Proposed Model with other ML & DL Model

sno no	Classifier Name	Hamming Loss
1	KNN	0.1597
2	QDA	0.2266
3	Decision Tree	0.301
4	SVC (Linear)	0.3434
5	SVC (RBF)	0.388
6	Gaussian NB	0.5191
7	Ada boost	0.5437
8	Proposed Model (RF)	0.1110

TABLE 4: Comparison of Jaccard Score with Proposed Model with other ML & DL Model

sno no	Classifier Name	Jaccard Score
1	KNN	0.7245
2	QDA	0.6304
3	Decission Tree	0.5371
4	SVC (Linear)	0.4887
5	SVC (RBF)	0.4408
6	Gaussian NB	0.3165
7	Adaboost	0.2955
8	Proposed Model (RF)	0.8001

experimental tests using CNN and ML models, we obtained the best-balanced accuracy, weighted specificity, weighted sensitivity, weighted precision, and weighted F1-score for classifying 43 bird species are $86.31\%, 93.49\%, 99.63\%, 93.76\%, and 93.31\%$. Performing this experiment, we can evaluate that this approach shows promising results for the development of sound classification. Moreover, the UrbanSound8K datasets were used for the experimental process, in addition, the classification accuracies and f1-score were calculated for performance estimation. In our first phase experiment, we achieved prominent results with 87% accuracy 0.78 F1-score. In our second phase experiment with the help of boosting method achieved 88% accuracy with a 0.78 F1-score, finally, we accomplished a promising result with 90% accuracy by applying cross validation with 5-Fold, 10-Fold, and 15-Fold. The results of the proposed model and its performance also equated with the results of the state-of-the-art technique. The results of the proposed model over all other compared models exhibited that the proposed model outclassed in entirely compared models.

References

[1] A. Chowdhury and A. Ross, "Fusing mfcc and lpc features using 1d triplet cnn for speaker recognition in severely de-

TABLE 5: Other Performance Criteria for the Urbansound8k to the proposed Model

Class Index	Precision	Recall	F1-Score
0	0.97	0.96	0.97
1	0.97	0.76	0.85
2	0.79	0.87	0.83
3	0.84	0.81	0.82
4	0.89	0.89	0.89
5	0.96	0.96	0.96
6	0.97	0.78	0.87
7	0.90	0.95	0.92
8	0.95	0.95	0.95
9	0.76	0.81	0.78

graded audio signals," *IEEE transactions on information forensics and security*, vol. 15, pp. 1616–1629, 2019.

[2] S. A. Alim and N. K. A. Rashid, *Some commonly used speech feature extraction algorithms*. IntechOpen London, UK:, 2018.

[3] J. Sharma, O.-C. Granmo, and M. Goodwin, "Environment sound classification using multiple feature channels and attention based deep convolutional neural network." in *Interspeech*, 2020, pp. 1186–1190.

[4] S. Qiu, H. Zhao, N. Jiang, Z. Wang, L. Liu, Y. An, H. Zhao, X. Miao, R. Liu, and G. Fortino, "Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges," *Information Fusion*, vol. 80, pp. 241–265, 2022.

[5] A. Khamparia, D. Gupta, N. G. Nguyen, A. Khanna, B. Pandey, and P. Tiwari, "Sound classification using convolutional neural network and tensor deep stacking network," *IEEE Access*, vol. 7, pp. 7717–7727, 2019.

[6] J. Xie, K. Hu, M. Zhu, J. Yu, and Q. Zhu, "Investigation of different cnn-based models for improved bird sound classification," *IEEE Access*, vol. 7, pp. 175 353–175 361, 2019.

[7] Z. Zhang, S. Xu, S. Cao, and S. Zhang, "Deep convolutional neural network with mixup for environmental sound classification," in *Chinese conference on pattern recognition and computer vision (prcv)*. Springer, 2018, pp. 356–367.

[8] E. Tzinis, S. Wisdom, J. R. Hershey, A. Jansen, and D. P. Ellis, "Improving universal sound separation using sound classification," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 96–100.

[9] F. Demir, D. A. Abdullah, and A. Sengur, "A new deep cnn model for environmental sound classification," *IEEE Access*, vol. 8, pp. 66 529–66 537, 2020.

[10] A. K. Dwivedi, S. A. Imtiaz, and E. Rodriguez-Villegas, "Algorithms for automatic analysis and classification of heart sounds–a systematic review," *IEEE Access*, vol. 7, pp. 8316–8345, 2018.

[11] F. Demir, A. M. Ismael, and A. Sengur, "Classification of lung sounds with cnn model using parallel pooling structure," *IEEE Access*, vol. 8, pp. 105 376–105 383, 2020.

[12] S. Abdoli, P. Cardinal, and A. L. Koerich, "End-to-end environmental sound classification using a 1d convolutional neural network," *Expert Systems with Applications*, vol. 136, pp. 252–263, 2019.

[13] Z. Mushtaq, S.-F. Su, and Q.-V. Tran, "Spectral images based environmental sound classification using cnn with meaningful data augmentation," *Applied Acoustics*, vol. 172, p. 107581, 2021.

[14] Y. Su, K. Zhang, J. Wang, and K. Madani, "Environment sound classification using a two-stream cnn based on decision-level fusion," *Sensors*, vol. 19, no. 7, p. 1733, 2019.

[15] Z. Mushtaq and S.-F. Su, "Environmental sound classification using a regularized deep convolutional neural network with data augmentation," *Applied Acoustics*, vol. 167, p. 107389, 2020.

[16] L. Roberts, "Chromagram ," https://librosa.org/doc/main/auto_examples/plot_chroma.html, March-06-2013, [Online; accessed 19-May-2022].

[17] JustinSalamon, "UrbanSound8k Dataset ," https://urbansounddataset.weebly.com/urbansound8k.html, 2008, [Online; accessed 19-July-2008].

[18] F. Chollet, "KERAS API ," https://keras.io/, 2015, [Online; accessed 21-May-2022].

[19] G.-F. Fan, Y.-H. Guo, J.-M. Zheng, and W.-C. Hong, "Application of the weighted k-nearest neighbor algorithm for short-term load forecasting," *Energies*, vol. 12, no. 5, p. 916, 2019.

[20] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999.

[21] S. Ahmad, S. Agrawal, S. Joshi, S. Taran, V. Bajaj, F. Demir, and A. Sengur, "Environmental sound classification using optimum allocation sampling based empirical mode decomposition," *Physica A: Statistical Mechanics and its Applications*, vol. 537, p. 122613, 2020.

[22] V. Rainardi, "SVM - RBF Kernel ," https://dwbi1.wordpress.com/2021/05/24/svm-with-rbf-kernel/ , 2021, [Online; accessed 19-May-2022].

[23] P. C. S. Abirami, "Multi Layer Perceptron ," https://www.sciencedirect.com/topics/computer-science/multilayer-perceptron, 2020, [Online; accessed 19-May-2022].

[24] A. Sharma, "Decision Tree Model ," https://www.analyticsvidhya.com/blog/2021/02/machine-learning-101-decision-tree-algorithm-for-classification/, 2021, [Online; accessed 19-May-2022].

[25] Opengenus, "Guassian NB ," https://iq.opengenus.org/gaussian-naive-bayes/, 2021, [Online; accessed 19-May-2022].

[26] GeeksforGeeks, "Quadratic Discriminant Analysis ," https://www.geeksforgeeks.org/quadratic-discriminant-analysis/, 2021, [Online; accessed 19-May-2022].

[27] K. Iqbal, S. A Khan, S. Anisa, A. Tasneem, and N. Mohammad, "A preliminary study on personalized spam e-mail filtering using bidirectional encoder representations from transformers (bert) and tensorflow 2.0," *International Journal of Computing and Digital Systems*, vol. 11, no. 1, pp. 893–903, 2022.