Retraction

Retraction: Urban Sound Classification Using Convolutional Neural Network Model (IOP Conf. Ser.: Mater. Sci. Eng. 1099 012001)

Published 20 September 2022

This article has been retracted by IOP Publishing following an allegation that the work contains tortured phrases [1].

IOP Publishing has investigated in line with the COPE guidelines, and agree the article contains a number of tortured phrases, and bears similarities to an unreferenced earlier work by different authors [2].

IOP Publishing wishes to credit PubPeer commenters [3] for bringing the issue to our attention.

The authors have neither agreed nor disagreed to this retraction.

[1] Cabanac G, Labbe C, Magazinov A, 2021, arXiv:2107.06751v1
[2] C Mike Smales, (2019), ‘Sound Classification using Deep Learning’, https://mikesmales.medium.com/sound-classification-using-deep-learning-8bc2aa1990b7
[3] https://pubpeer.com/publications/02AE4E307B9D410A1220AF8781EBC

Retraction published: 20 September 2022
Urban Sound Classification Using Convolutional Neural Network Model

Srishti Garg1, Tanishq Sehgal, Aakriti Jain1, Yash Garg1, Preeti Nagarath1 and Rachna Jain1

1Bharati Vidyapeeth's College of Engineering, Paschim Vihar, New Delhi-110063

E-mail: garg.srishti39@gmail.com

Abstract. The programmed content-based order of urban sound classes is a significant part of different developing methods and applications, for example, observation, urban soundscape comprehension and commotion source distinguishing proof, along these lines the exploration subject has increased a great deal of consideration lately. The objective of this paper is to create a proficient AI based plan for urban sound classification. Ongoing fruitful utilizations of convolutional neural systems (CNNs) to sound order and discourse acknowledgment have spurred the quest for better information portrayals for progressively proficient preparation. Visual presentations of a sound signal, through different time-recurrence portrayals, for example, spectrograms offer a very good representation of the worldly picture of the original signal. Utilizing a spectrogram picture of the sound and afterward changing over the equivalent to information focuses (As is accomplished for pictures). This is effortlessly done utilizing mel_spectogram a function of Librosa. At the approval stage, we lead tests on Urban Sound 8K database which comprises 10 classes of urban sound happenings with 8732 real-world sound clips. As a result, we see how convolutional neural network (CNN) frameworks with raw sound waveforms improve the exactness in urban sound classification and clearly shows the structure concerning the number of parameters.

1. Introduction
Sounds in the built-up areas have been a big job in understanding the part of interactive media. With such significance including a developing interest, describing ecological sound is the productive method to utilize it. Dissimilar to discourse and sound signs, urban sounds are for the most part unstructured sounds. They incorporate different genuine clamors produced by human exercises, running from transportation to recreation exercises. Programmed urban sound classification trained models could distinguish the noise disturbance source, profiting urban liveability, like loud noise control, sound observation, soundscape evaluation furthermore, acoustic condition arranging. Be that because it may, sounds in an urban condition are generally made out of various sound sources, which makes it challenging to classify them into categories.

1.1. Objective
Truly, contrasting with many research chips away at programmed discourse acknowledgment (ASC, for example, [1]) or music data recovery (MIR, for example, [2]), the investigations towards urban sound classification investigation are relatively restricted. To distinguish a prevailing sound occasion from its disordered blends, most past examinations center around finding the relevant proficient sound element
that is illustrative of the objective sound class, regardless of whether the sound is in part veiled by other sound sources. In this pattern, early takes a shot at urban sound arrangement depend on physically designed highlights that have been broadly utilized in sound/discourse characterization, for example, (MFCCs), MPEG 7 sound highlights, or other redid fleeting and ghastly acoustic highlights. Joined with some broad classifiers (like SVM, k-NN, GMM, arbitrary backwoods, and so forth.), these highlights are then arranged into their related sound classes. In music data recovery (MIR), researchers have generally changed over crude waveforms of sound signal. 2D portrayals have been viewed as a viable type of sound information by breaking down the sign with parts (e.g., STFT) and utilizing portrayals in recurrence. During the recent environmental sound classification timework, handling of sign and artificial intelligence approaches including factorization of grid [2][3][4][5], word reference learning [6][7], wavelet filter banks [1][8][9][10] are regularly utilized. This procedure alleged "built highlights" requires noteworthy building exertion and significant earlier information about the issue. Likewise, highlight building is regularly heuristically structured and probably won't be ideal for the errand.

Further in this project GMM, RNN, CRNN, MFCC and ASR for sound classification are being compared. CNN model which has proved to be extremely efficient in learning deep architectures is used. There are two processes of converting sound clips to images, spectrograms and Mel Frequency Cepstral Coefficients (MFCC). The spectrogram conversion has been used here. The significant differentiation (figure 1) is that a spectrogram utilizes a direct dispersed recurrence scale (so every recurrence container is separated an equivalent number of Hertz separated), though a MFCC depends on a semi logarithmic divided recurrence scale, which is vaguer to how the manual hear-able framework measures sound.

![Figure 1. Different Waveforms Showing the Scale. Spectrogram Uses Linear Spaced Frequency Scale and MFCC Uses a Quasi-Logarithmic Spaced Frequency Scale](image_url)

1.2. Contribution

Late advances in profound neural systems have empowered elements realizing which takes crude sound signs, in this manner limiting the exertion of preprocessing the information. Highlight learning with crude information was endeavored to explain a music labeling job through a CNN simulation [11]. Here, the system along the waveforms of time vs space has been applied to perceive a mixture of different
areas [12][13][14]. For the urban sound classification task, start to finish urban sound grouping frameworks with a CNN have proposed [15]. The presented method uses deep learning and has achieved a very high accuracy of 78% with UrbanSound8K dataset for 10 different classes. They demonstrate that a start to finish framework is fit for separating highlights from the waveforms. The dataset contains 8732 sound excerpts which are of time length less than or equal to 4 seconds, they are from ten different classes, which are: Children_playing, air_conditioner, dog_barking, siren, drilling, car_horn, jackhammer, gun_shot, street_music, engine_idling. Our contribution is evaluating the Urban Sounds also to suggest an efficient methodology to categorise sounds types from different auditory sensations, which achieves a better result as compared to state-of-the-art methods.

1.3. Uniqueness of Paper
A user defined function named the ‘wave plotter’ function was used to study the different characteristics of the sound samples. Using this function, we got to know about the following properties of some sound files chosen at random:
1. Sampling rate
2. Bit depth
3. Number of channels (Mono and Stereo nature)
4. Duration
5. Number of samples
6. Class
7. Plot of the wave sample

This gave a clear picture of what exactly the behaviour of a particular sound file is.

Following are the characteristics of Dog_bark sound clip chosen at random from the folders:
- Sampling rate: 44100Hz
- Bit depth: 16
- Number of channels: 2
- Duration: 0.31755102040816324 second
- Number of samples: 14004
- Class: dog_bark

Following are the characteristics of Dog_bark sound clip chosen at random from the folders:
- Sampling rate: 44100 Hz
- Bit depth: 16
- Number of channels: 2
- Duration: 4.0 second
- Number of samples: 176400
- Class: children_playing

Histograms are being used as the pictorial representation technique for the sound clips to helps to reach the results. It is known that gradient descent with adaptive learning rate can help speed up convergence in neural networks. There is python code and visual illustration of three widely used optimizers — AdaGrad, RMS-Prop, and Adam. Adam combines the best properties of RMS-Prop and AdaGrad to work well even with noisy or sparse datasets.

1.4. Organization
This paper contains 6 sections. First section includes an introduction where the motivation around the problem statement, contribution and the uniqueness of our paper are covered. Second section contains related work which has brief information about the work done previously around this subject, it also includes what is the objective of this paper. The third section has the Experimental setup, which gives the detailed information about the dataset, approach used, comparison of CNN model with other...
techniques and system architecture. The next section explains the algorithm used to counter the problem statement. Fifth section records the result, and the last section concludes our project.

2. Related Work
During our research and learning phase, we scrutinized various research papers published on Urban Sound Classification. Here is the review of some of the papers we examined:

2.1. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification [16]
The above research paper deep CNN specification is proposed wherein a mixture with a group of enhanced diversity of knowledge, which produces varied results for environmental sound classification. Better results are fetched when coupled with data augmentation: this mix surpassed the proposed CNN structure without augmentation. At a result the influence of every variation on the sound class accuracy was determined.

2.2. Environmental Sound Classification with Convolutional Recurrent Neural Networks [17]
The intention of this paper was to apply and analyze convolutional neural networks models on sound datasets. Three datasets were taken: a 5-fold (ESC-10 and ESC-50) and 10-fold (UrbanSound8K).

2.3. Using Deep Convolutional Neural Network to Classify Urban Sounds [18]
The purpose of this paper was to research categorizing urban sound issues using CNN. It suggests a way-out, highlighting the effect on the sound categorization performance based on the input spectrograms. Boxplots along with confusion matrices were plotted for various inputs. Results obtained, suggest that, the time resolution index has worthy results on the categorization performance, and better of these are habitually fetched for input with middling time resolution. Diving ahead in research, it might throw light on the inter-relation of the sound properties.

2.4. Urban Sound Classification using Convolutional Neural Networks Model
The goal of our project is to differentiate between various sound excerpts accurately, using convolutional neural networks. We've transmogrified sound files into spectrograms- an intensity vs frequency plot using Librosa. We have tested UrbanSound8K database which comprises 10 classes of urban sound occasions with 8732 real-world clips. The results convey the manner in which convolutional neural framework with waveforms improves the exactness in urban sound arrangement and outfit practicality of its formation concerning the count of parameters. The model meshwork is effectuated in Python with Keras [18]. While treating the training data, the Neural network hones cross entropy loss through small batch ADAM [19] gradient descent optimization algorithm. Each small batch constitutes of random inputs of the training data collection. The training of the model is carried out using 90 epochs.

3. Experimental Setup
3.1. Dataset
For the analysis of the problem statement, we have use UrbanSound8K dataset [20] which comprehends 10 day to day sounds taken from urban territory, namely, forced air system, vehicle horns, kids playing, barking dogs, penetrating, motor lingering, gun shot, jackhammer, alarm and music from the streets. It contains 8732 marked sound example passages of about 4 seconds stretch or less, 9.7 hours in total. The dataset is metamorphosed into 10 folds. It is not necessary that each fold contains the same sound type and hence ensures that the distribution of training and testing of a lot of sounds are indiscriminate and proper. Analyzing every time, the default overlay 10 is used for our test set, and the remaining for
preparing that is the train set [19][18]. The sound signs are examined to 8 kHz and normalized to zero mean and fluctuation one for calculation and assessment speed.

3.1.1 Segment Names
Slice_file_name: It is the name of the sound file. It has the following arguments: [fsID]-[classID]-[occurrenceID][sliceID].wav, where:
1.1. fsID: The Freedsound ID of the profile from which this excerpt
1.2. classID: It is the numeric representation to all the class names
1.3. occurrenceID: a numerical differentiator to recognize various events of the sound inside the first chronicle
1.4. sliceID: a numeric identifier to recognize various excerpts taken from a similar event.
start: It is the beginning time of the sound clip from the free stable chronicle.
end: It signifies the time where the sound clip ends in the first free stable chronicle.
salience: It is the abstract significant rating of the sound wherein 1 = frontal area and 2 = foundation.
fold: It is the fold number in which the corresponding sound clip has been placed.
classID: It is the numeric representation to all the class names which are as followed:
  • 0: Air_Conditioner
  • 1: Car_Horn
  • 2: Children_Playing
  • 3: Dog_Bark
  • 4: Boring
  • 5: Engine_Idling
  • 6: Gun_Shot
  • 7: Jack_Hammer
  • 8: Alarm and
  • 9: Street_Music
Class: it signifies the type of sound clip i.e., Children_playing, air_conditioner, dog_barking, siren,drilling, car_horn, jack_hammer, gun_shot, street_music, engine_idling.

3.2. Proposed Approach
Our way of working is applying a CNN technique and training our model by splitting it into 75-25 train-test portions of data. The dataset used is Urban Sound8K. The sound clip is extracted from free sound. The extracted sound clips are of less than or equal to 4 second. The sound clips are then plotted in a time frequency graph and the pictorial representation that is spectrograms. Feature extraction of the sound clip has been done by extracting the MFCC’s. Comparing the spectrograms and the sound clip characteristics, the accuracy is calculated. Refer table 1.

Table 1. Technique comparison between RNN, CNN, CRNN and LSTM

| Technique | Dataset            | Advantage                                           | Disadvantage                                                   |
|-----------|--------------------|-----------------------------------------------------|---------------------------------------------------------------|
| RNN       | Urban Sound8k      | Efficient in learning the long-term related context in the auditory signals | Does not easily capture the unvarying frequency in the domain, providing a high-level modeling of the data making it more difficult. |
| Model | Dataset | Feature Description | Shortcoming Addressed |
|-------|---------|---------------------|-----------------------|
| CNN   | Urban Sound8k | It capable of fetching higher level features which are immutable to local spectral and time related variations | Addresses the former shortcoming by learning filters that are displaced in both time and frequency, lacking however longer time related context information |
| CRNN  | Urban Sound8k | In CRNN, CNN's capacity to learn neighborhood interpretation stable channels and RNN's bit of leeway of demonstrating short- and long-haul transient conditions are gathered in one item. | It is difficult to detect small segments which is due to the limited discriminative -ness of the global visual features. It only predicts a label, not a segmentation box. |
| LSTM  | Urban Sound8k | Less Data Loss than CNN, better performance and accuracy, more generalized approach, and proves to be a reliable classifier for the networks that take magnitude mel-spectrograms. | LSTM is more computationally intensive and prone to overfitting, although has less trainable parameters than CNN. |

![Figure 2](Retracted). Accuracy and Loss Evaluated on Train Data
The accuracy and loss on train and test data have been evaluated and plotted by the model trained using CRNN and LSTM. CNN accuracy was recorded 80% on the train data and LSTM accuracy was recorded 85% on the train data (figure 2) whereas on the test data CNN accuracy was recorded 77% as and LSTM accuracy was recorded as 82% (figure 3).

3.3. Preprocessing
The raw sound sign isn't reasonable as an immediate contribution to a morpheme because of its extraordinary high spatial property and the manner in which it would be unlikely for comparable sounds to be situated parallel in vector space. Along these lines, a famous methodology for specifying gaining from the sounds is to convert over the sign into a periodic recurring depiction, a decision being the mel-spectrogram. We disengage log-scaled mel spectrograms with 10 portions covering the discernible recurrence run with jump size of time length less or equal to 4 seconds. We appraised with more groups, but this didn’t give the desired result and so we reached a conclusion of dwelling to 10 groups. While we could bring in usage the subsequent logarithmic-mel-spectrograms originally as a part for the constituent learning, it has been shown that the scholarly foreground can be efficiently improved by withdrawing the info measurements utilization for example ZCA. Then this is passed to the element learning square. Element adapting can be implied to single edges of the logarithmic mel-spectrograms, or on the other hand to a few sequential edges bringing about 2D patches. Gathering a few continuous casings (by connecting them into a solitary bigger vector before PCA brightening), otherwise called shingling, permits us to learn highlights that consider worldly elements. This choice is especially intriguing for urban clamor like sounds, for example, lingering motors or jackhammers, where the transient elements might increase our capacity to recognize sounds whose brief pinnacles can be fundamentally the same as.

3.4. Convolution Neural Network Model
It is a neural structure with several layers, where each one comprising of a majority of individual neurons. With more sagacious structures, CNN has ended up being incredibly prolific in learning theoretical highlights. Weight sharing, spatial examination, and nearby association are three weighty hallmarks of CNN. CNN gives the information of all integrands for devising up the model just as order,
yet with incredible measure of calculation. The edifice of the CNN is delineated in figure 4. It shows 3 convolution layers, 2 max pool layers, and 1 completely associated layer.

**Figure 4. Layer Architecture of CNN**

3.5. **System Architecture**
A Convolutional Neural Network System is fabricated from several layers stockpiled in a sagacious design: (figure 5) an information slab, a gathering of convolutional and pooling slabs (which can be consolidated through different methods), a predetermined number of completely associated concealed layers, and a yield (misfortune) layer.

**Figure 5. Method of Approach**

The real contrast, when compared with the multilayer perceptron, is observed in the presentation of a mix of convolution and pooling tasks. A convolutional layer presents an outstanding technique for figuring out covered units which intends to misuse the local design present in the 2-D data (generally, anyway not obliged to, pictures). Each covered unit, as opposed to being related with all the information sources brought from the past layer, is confined to planning simply a little piece of the whole data space (for instance a 3x3 window of picture parts (picture pixel)), called its responsive field. The heaps of a particularly enclosed information assortment are made out of a convolutional segment (channel) which is applied on the whole information space, achieving a component map. Thusly, one parcel of burdens can be utilized again for the whole data set. This relies on the explanation that locally significant features
would be similarly, useful in various spots of the data blocks - a framework which not simply incomprehensibly diminishes the amount of boundaries to evaluate, anyway improves capacity to key developments of the data. A run of the mill convolutional layer will comprise of various channels (including maps). Further reduction in measurements can be achieved by the pooling layers, which stir up the nearby cells of a component map. The most generally recognized pooling assignments transporter out are claiming the greatest or mean of the data cells. This further improves invariance to numerous translations [21].

3.5.1. Loading the Necessary Libraries and Analyzing the Dataset
The main packages and libraries used here include numpy, pandas, sklearn, seaborn, R skimage, librosa, tensorflow.

3.5.2. Feature Extraction using Librosa and Building the Database
There are 3 methods of feature extraction from audio file:
- To use the mffc data of the audio clips.
- Making use of a spectrogram image of the audio clip and then converting it to data points and storing them into a numpy array. This can be done using the mel_spectrogram function of Librosa library.
- Combining both the features to build a more efficient model.

Spectrogram method for the audio clip feature extraction has been chosen:

In this method, initially the data is analysed by plotting the plots of the sound files and also by plotting the spectrogram images of the sound files. In order to obtain the path of the file a generalised code has been written that fetches the path of the sound files, iterates through each one of them and the labels have been converted to categorical data for classification. CNN has been used as the primary layer to classify data [22][23].

Librosa has been used to extract features. Each fold is visited and the data for each file is extracted. In order to obtain the data of the spectrogram image, the mel_spectrogram function of librosa that fetches the data of the spectrogram is used. This data is directly corresponded into a numpy array. The Kaiser Best extraction technique is used in this process which ensures a faster extraction of the features. After treating the data, (cleaning and reshaping), a 75-25 split has been made. Classes (Y) have been converted to Categorically Encoded Data using Keras.utils and further the CNN model is created.

3.5.3. Creating the CNN Model and Obtaining the Results
At first the 2D convolution layer of 64 units is made utilizing the tanh actuation work which makes a convolution part that is convolved with the layer contribution to supply a tensor of yields. 2D CNN layers take a 3-dimensional info, normally an image with three shading channels. They pass a channel, additionally called a convolution part, over the picture, investigating somewhat set of pixels straightforwardly, and clearing the window until they have examined the total picture. The convolution activity computes the scalar result of the pixel esteems inside the current arrangement of pixels with the substance conveyed inside the channel window.

Presently Max pooling activity for 2D spatial information is done which downsamples the information portrayal by taking the greatest incentive over the window characterized by pool_size for each measurement along the component's hub. At that point a convolution 2D layer of 128 units is made utilizing tanh initiation work. Here, the dropout layer is included request to forestall overfitting. In convolutional neural frameworks, where the quantity of boundaries is diminished through weight sharing, the amount of evaluated regards is by far most of the events more prominent than the amount of getting ready cases by a huge degree. This can achieve poor out-of-test theory.
Tanh activation function: The Tanh work is an actuation work which re-scales the qualities between -1 and 1 by applying a limit simply like a sigmoid capacity.

Softmax activation function: The Softmax relapse is a type of strategic relapse that standardizes an information esteem into a vector of qualities that follows a likelihood conveyance who's all-out summarizes to 1. The yield esteems are between the range [0,1] which is pleasant on the grounds that we can stay away from twofold characterization and suit the same number of classes or measurements in our neural system model. This is the reason softmax is once in a while alluded to as a multinomial strategic relapse. Condition 2 speaks to the softmax work. Another name for Softmax Regression is Maximum Entropy (MaxEnt) Classifier. The capacity is generally used to process loss that can be normal when preparing an informational collection. Known use-instances of softmax relapse are in discriminative models, for example, Cross-Entropy and Noise Contrastive Estimation. These are just two among different methods that endeavor to upgrade the present preparing set to improve the probability of anticipating the right word or sentence.

The organization is actualized in Python with the Keras bundle. During the preparation cycle, the ConvNet model enhances cross-entropy misfortune through little group ADAM inclination plunge advancement calculation. Every small clump comprises of irregular contributions of the preparation dataset. Model is prepared with 90 epochs.

3.5.4. Obtaining Results and Storing them into a csv File
Since the model is trained using 90 epochs so the accuracy for each iteration was obtained which had a range of 47% to 98% as iterations proceed further. After this, the accuracy is obtained for the test data which roughly comes out to be 86% roughly. After this the predicted values are now stored into a csv file.

3.5.5. Dropout Learning
Overwhelming neural constructions have a disposition of tendency towards overfitting. In fact, even in Convolution frameworks, where the quantity of boundaries is diminished through weight division, the amount of evaluated regards is by far most of the events more noteworthy than the amount of planning cases by a huge degree. This can achieve less than impressive out-of-test hypothesis. One point of view is to deal with this issue with dropout learning. The thought is extremely clear, yet uncommonly convincing. In every planning accentuation each covered unit is subjectively emptied with a predefined probability (at first half), and the learning technique continues consistently. These self-assertive annoys feasibly shield the framework from picking up deceiving conditions and making complex coadaptation between hid units. Thusly colossal social occasions of neurons become strong not simply with respect to various neurons. Configuration averaging familiar by dropout endeavors with ensure that each hidden unit learns feature depictions that are generally sure in making the correct portrayal answer.

3.5.6. Optimization
The design is worked with Keras [19] and Tensorflow [20]. The sound sign is narrowed down with inspecting rate 8 kilo Hz using the library Librosa [24]. The detail shape vector is of 32000 vector length. The CNN models utilizing Adam [25] and straight out Categorical Cross Entropy as a loss function is prepared. To use all out cross-entropy loss, the class marks have been differentiated into an all-out organization which suggests that each class is a vector of 10-dimensions i.e., each one of the zeros aside from a 1 in the file corresponding to the class.

4. Algorithm
   Input: Train Sound Clips
   Output: Test sound Clip class Label
   1. If (trained model is not available)
   2. Load the necessary libraries
   3. Load the dataset
   4. Fetch the sound clip characteristics
   5. Classification of sound clip class in the folders
6. Fetch the spectrogram plots using mel_spectrogram function and plot them.
7. Perform reshaping and data cleaning.
8. Split the dataset into train and test data groups.
9. Treat the data through 2 convolution layers.
10. Study the accuracy of train and test data using 90 epochs.
11. Else:
12. Load the trained model.
13. Test the model with a test data group.
14. Calculate the accuracy.

According to the Algorithm, experiment starts with loading the necessary libraries and packages which include Pandas, NumPy, Matplotlib, SK Learn, GridSearchCV, Tensorflow, Keras, and Librosa. The algorithm supports a condition check whether a trained model is present or not. If present, it is loaded or else a wav file containing it is loaded. This wav file is then converted to spectrogram plots which represent a comparison of frequency and the intensity of sound file chosen. Along with the spectrogram plots different characteristics of the sound files like Sampling rate, Bit depth, Number of channels (Mono and Stereo nature), Duration, Number of samples, Class, Plot of the wave sample are also obtained. To avoid passing the path repeatedly to fetch each sound file from the dataset individually, a generalised code was written that fetches us the path of the sound files so that the data can be studied easily. For the feature extraction, iterate through all the 8732 sound pieces. The data entries of the sound have been resampled using the Kaiser_fast/ Kaiser_best method. Mel S spectrum function is used to extract the spectrogram plot points and put them in a numpy array. Reshaping and data cleaning is followed after fetching the plot points in a numpy array which is followed by a train and test split of the entire dataset. Class type from these sound clasps is then recognized utilizing the CNN structure. The CNN model is then prepared and approved utilizing preparing and approval sets, separately and the prepared model is spared. For working up the CNN model, the information is passed through two convolution layers utilizing the tanh enactment work which acquires the qualities obtained through the mel_spectrogram work between -1 and 1. Max pooling layer underpins a window size of 2*2 that emphasizes through the information to get the tensor of yield. The dropout layer helps in lessening the overfitting in the information which is generally found in the datasets utilized in the profound learning models. The preparation information is gone through the thick layer having a clump size of 50 which implies that in every age emphasis, 50 information passages are served to the model. Trained model is tested with testing datasets and spectrograms, accuracy epoch and loss-epoch are studied.

5. Results
The CNN model fetches the best results for image classification. In this model 90 epochs have been used with a batch file size of 50 which means that each iteration will have 50 sound clips to be treated. The accuracy of the epochs varies between 45% to 98% roughly, while epochs are fed with training data constituting 75% of the total data involved. The predictions have been made on testing data which constitutes 25% of the total data. The accuracy thus obtained after the predictions are made is roughly about 86.44%

The prime objective of our project was to assess whether the CNN (Convolution Neural Network) can be effectively applied [26] to Urban Sound Classification [27] considering and keeping in mind the constrained dataset that is available. Epochs Accuracy (as shown in figure 6) of the 90 epochs performed on the training data.
CNN model accuracy calculated with the dataset Urban Sound8K was calculated as 0.8644. KNN model accuracy was recorded less than the CNN model accuracy. For improving the score, CRNN [28] can be used further. In KNN, yield totally depends on closest neighbors, which may not be a decent decision for our dataset. Additionally, it is touchy to remove measurements. Then again, CNN separates the highlights from the information, which is exceptionally useful for making investigation. What's more, late accomplishment in the CNN uncommonly wavelet for the sound application, we liked to go with CNN.

Figure 6. Comparison between Techniques (RNN vs DeepNN vs CNN)

Figure 7. Spectrograms of Jackhammer, dog_bark and Drilling.
The input sound signal .wav file was converted into the .jpg that is in the spectrograms (figure 7) for obtaining the results for the model.

MFCC highlights are one of the great highlights which have worthy normal characterization precision with various AI arrangements. Relative investigation of various classifiers with MFCC highlights has sensible grouping precision to understand the troubles in picking best classifiers for USC. The second measurement of MFCC highlights called MFCC-second likewise has satisfactory arrangement exactness for USC. For future work, various highlights will be examined for a lot of creative highlights for best characterization exactness in USC. Although both MFCC and MFCC-second highlights have adequate grouping correctnesses with AI classifiers, we have to attempt to improve the arrangement precision of USC by thinking about other sign preparing strategies and classifiers.

Earlier, it was observed that it is cumbersome to differentiate among some of the classes. Infact, the following sub-groups are close enough in shape:

- Mundane sounds for Air.Conditioner, Drilling, Engine.Idling and Jack_Hammer.
- Sharp peaks for Dog.Barking and Gunshots.
- Similar pattern for Children.Playing and Street.Music.

Confusion matrix (figure 8, figure 9) has been used to find out if the final model also struggles to categorise among different types of sound.

![Confusion matrix](image)

**Figure 8. Urban Sound 8k Confusion matrix using CRNN**
6. Conclusion
Conducted experiments show that a convolutional model beats basic methodologies dependent on physically built highlights and accomplishes a comparative level as other element learning strategies. It is observed that CNN model can be adequately implemented to sound cluster assignments even with limited datasets. Furthermore, all things considered, an extensive increment in the size of the available information would immensely optimize the presentation of prepared models and accuracy of the same. As the opening to human exactness is at this point critical. Aside from past investigations done, the objective is to completely consider input spectrograms [29] on the grouping execution of the sound class information by the model. Further examination will unfurl more data about the connection between hearable sensations and the time goal of its range portrayal utilizing the CRNN and a lot more models. CNN is a CRNN having no intermittent layers, and a RNN is a CRNN with no convolutional layers.

7. References
[1]. M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo, “Audio keyword generation for sports video analysis,” ACM TOMCCAP, vol. 4, no. 2, pp. 1–23, 2008.
[2]. B. Ghoraani and S. Krishnan, “Time–frequency matrix feature extraction and classification of environmental audio signals,” IEEE Trans. on Audio, Speech, and Lang. Process., vol. 19, no. 7, pp. 2197–2209, 2011.
[3]. A. Mesaros, T. Heittola, O. Dikmen, and T. Virtanen, “Sound event detection in real recordings using coupled matrix factorization of spectral representations and class activity annotations,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, Apr. 2015, pp. 151–155.
[4]. E. Benetos, G. Lafay, M. Lagrange, and M. D. Plumbley, “Detection of overlapping acoustic events using a temporally constrained probabilistic model,” in IEEE International
Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, Mar. 2016, pp. 6450–6454.

[5]. V. Bisot, R. Serizel, S. Essid, and G. Richard, “Acoustic scene classification with matrix factorization for unsupervised feature learning,” in Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016, pp. 6445–6449.

[6]. J. Salamon and J. P. Bello, “Unsupervised feature learning for urban sound classification,” in IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, Apr. 2015, pp. 171–175.

[7]. P. Foggia, N. Petkov, A. Saggiase, N. Strisciuglio, and M. Vento, “Audio surveillance of roads: a system for detecting anomalous sounds,” IEEE Trans. on Intell. Transp. Syst., vol. 17, no. 1, pp. 279–288, 2016.

[8]. S. Chu, S. Narayanan, and C.-C. J. Kuo, “Environmental sound recognition with time frequency audio features,” IEEE Trans. on Audio, Speech, and Lang. Process., vol. 17, no. 6, pp. 1142–1158, 2009.

[9]. J. T. Geiger and K. Helwani, “Improving event detection for audio surveillance using gabor filterbank features,” in 23rd European Signal Processing Conference (EUSIPCO), Nice, France, Aug. 2015, pp. 714–718.

[10]. D. M. Agrawal, H. B. Sailor, M. H. Soni, and H. A. Patil, “Novel TEO-based gammatone features for environmental sound classification,” is submitted in European Signal Processing Conf. (EUSIPCO), Kos island, Greece: August 28 – 2 September 2017.

[11]. S. Dieleman and B. Schrauwen, “End-to-end learning for music audio,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 6964–6968.

[12]. Zoltan Tüske, Pavel Golik, Ralf Schlüter, and Hermann Ney, “Acoustic modeling with deep neural networks using raw time signal for lvcsr.” in INTERSPEECH, 2014, pp. 890 894.

[13]. Pavel Golik, Zoltan Tüske, Ralf Schlüter, and Hermann Ney, “Convolutional neural networks for acoustic modeling of raw time signal in lvcsr,” in Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[14]. Dhaka, V. S., R. C. Poonia, and L. Raja, “A literature review on dedicated short range communication for intelligent transport,” International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE) 3.9 (2013): 1066-1071.

[15]. Y. Tokozume and T. Harada, “Learning environmental sound with end-to-end convolutional neural network,” in IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP), New Orleans, USA, 2017, pp. 2721–2725.

[16]. J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," in IEEE Signal Processing Letters, vol. 24, no. 3, pp. 279-283, March 2017, doi: 10.1109/LSP.2017.2657381.

[17]. K. J. Piczak, "Environmental sound classification with convolutional neural networks," 2015. IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), Boston, MA, 2015, pp. 1-6, doi: 10.1109/MLSP.2015.7324337

[18]. H. Zhou, Y. Song and H. Shu, "Using deep convolutional neural network to classify urban sounds," TENCON 2017 - 2017 IEEE Region 10 Conference, Penang, 2017, pp. 3089-3092, doi: 10.1109/TENCON.2017.8228392.

[19]. Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das. Very deep convolutional neural networks for raw waveforms. In IEEE ICASSP, pages 421–425 2017.
[20]. Justin Salamon, Christopher Jacoby, and Juan Pablo Bello, “A dataset and taxonomy for urban sound research,” in Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014, pp. 1041–1044.

[21]. J. Salamon and J. P. Bello, “Feature learning with deep scattering for urban sound analysis,” in the 2015 23rd European Signal Processing Conference (EUSIPCO). IEEE, 2015, pp. 724–728.

[22]. François Chollet, “Keras,” GitHub repository: https://github.com/fchollet/keras, 2015.

[23]. Abadi, Martín, Agarwal, Ashish, Barham, Paul et al., “TensorFlow: Large-scale machine learning on heterogeneous systems”, 2015.

[24]. Karol J. Piczak. "Environmental sound classification with convolutional neural networks", 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), 2015

[25]. Boqing Zhu, Changjian Wang, Feng Liu, Jin Lei, Zhen Huang, Yuxing Peng, Fei Li. "Learning Environmental Sounds with Multi-scale Convolutional Neural Network", 2018 International Joint Conference on Neural Networks (IJCNN), 2018.

[26]. Jonghee Sang, Soomyung Park, Junwoo Lee. "Convolutional Recurrent Neural Networks for Urban Sound Classification Using Raw Waveforms", 2018 26th European Signal Processing Conference (EUSIPCO), 2018

[27]. J. Sang, S. Park, and J. Lee, “Convolutional recurrent neural networks for urban sound classification using raw waveforms,” in the 2018 26th European Signal Processing Conference (EUSIPCO). IEEE, 2018, pp. 2444–2448.

[28]. S. Chachada and C.-C. J. Kuo, “Environmental sound recognition: A survey,” vol. 3, 10 2013, pp. 1–9.

[29]. V. Boddapati, A. Petef, J. Rasmussen, and L. Lundberg, “Classifying environmental sounds using image recognition networks,” Procedia computer science, vol. 112, pp. 2048–2056, 2017.