Environmental Sound Classification

Moinak Chakraborty (techmoinak@Knights.ucf.edu)
Sayma Sultana (sayma@Knights.ucf.edu)
University of Central Florida

***Abstract*- Sound classification is a common problem in machine learning in order to contribute to robotics and assist deaf people. Deep learning has improved performance of classification models. However, extracting acoustic features from raw soundwaves for neural network model is challenging till now. In this paper, we have presented preliminary work for classifying environmental sounds using neural network models. Accuracy of models have been calculated on average accuracy over five leave-one-fold-out evaluations and accuracy (34.36%) for neural network is higher than baseline implementation (k-NN- 32.20%,) for this dataset.**

I. Introduction

Classification is one of the classic supervised learning procedures in machine learning where machine learns from the data provided to the machine and classifies the data into predefined number of classes. The same classification algorithm could perform on structured or unstructured data. The main moto of the classification algorithm is to classify the structured or unstructured data input into different categories which were predefined. The input data would be mapped with the correct categories in output. Some of the usage of the classification algorithms are mentioned below. The tumor in our body is malignant or not, it could be one of the biggest usages of the classification algorithm. We have to take help of the computer vision in such cases though. It can identify the risk factor of the diseases. It can classify some letters, some images of animal kingdom. It can predict weather. It can predict the winner of the voting pole. So, these are few applications of classic classification algorithm in machine learning.

Here, in our paper, we are going to illustrate one of the biggest usages of classification algorithm where we can classify our environmental sounds. As we know we are surrounded by many of environmental sounds every time. Sometimes, we enjoy the sounds, sometimes we do not because of the noises created in the environment. However, we found significance to classify the environmental sound. We will explain what significance we really found in the environmental sound classification. We will explain later what phenomenon in which field really inspired us to classify the environmental sound. We should mention here, sounds are not easy dataset. Getting environmental sounds are not too hard nowadays. But, how we will get in which format and what could be the duration of those sounds and how many sounds we have to collect for which particular objects, those things matters a lot because we know for letting machine learn by any classification algorithm requires very good amount of significant data where we could do train our machine and test on the same data set as well. So, getting the proper dataset is one of the challenges we faced at first. Then, we got to know about Freesound from our professor which is one of the biggest sources of sounds. We gave a huge trust on the Freesound and started doing our learning procedure on the environmental dataset found in Freedsound and verified the same with our professor at the beginning of the project.

It should be mentioned that during our project we found several research papers already available on the same kind of classification mechanism. We have gone through some of the projects papers which we will mention at last as our references. We tried to implement in different approach sometimes with different folding techniques, feature selection techniques We should mention here that data like sound is one of the biggest challenges. As, we know the machine learning algorithms are dependent on the feature selection. Feature engineering is one of the crucial parts of the machine learning algorithm. For the data like sounds where anything is not present in structured way, feature selection is one of the biggest challenges. In classical classification algorithm sometimes, we do feature selection manually and then we start doing the learning algorithm for machine. But, here for the dataset like sound, we cannot manually select features. That is why we used neural network techniques to select the feature and to learn the machine as well. We searched and experienced with our data and found that neural network is one the easiest method here to do the feature engineering here for the dataset like sound.

II. Neural Network

The term itself comes from neurons. Neurons are there in our human body which connects each nervous connection to others. Sometimes it helps to take decision because it learns from experience, sense, past records and so on. The concept of neural network also comes from human neurons. As human brain can recognize something after seeing it by several interactions happened in human brain. Similarly, one machine can also recognize patterns, images, any other predefined classes of objects with color or with some different patterns or some different time frames and so on.

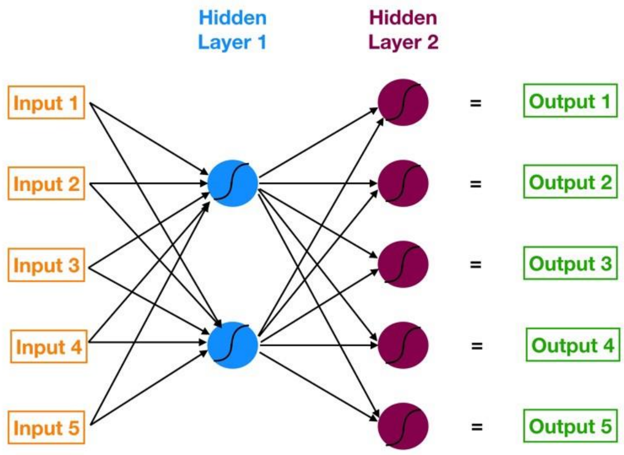


Figure 1: Neural Network with two hidden layers

We should start discussing neural network with *Figure 1*. Neurons are the basic units of one neural network. As illustrated in *Figure 1*, all the nodes involved in the architecture is called neurons. The above architecture is 3 layers neural network where two are hidden layers and one is output layer. We do not consider input layer in number of layers. The arrows in *Figure 1* illustrates that how each node is interconnected with other nodes in the complete architecture. As we can see in the first hidden layers, we only have two neurons and from the input layers, each input node has two different connections with different weights for these two nodes in the first hidden layers. So, total 10 different connections exist from input layer to first hidden layer. We are assuming the first input node which is Input 1 and the connections from first Input 1 o two nodes in first hidden layer. We are assuming two distinct connections and their weightage like *w1,1* and *w1,2* where *w1,1* means the connection between Input 1 and Neuron 1 which has the weight of *w1* and *w1,2* denotes the connection from Input 1 to 2nd neuron with weight *w2*. Now if we will calculate output of each neuron in hidden layer 1 as

$$Z1 = W1*In1 + W2*In2 + W3*In3 + W4*In4 + W5*In5 + Bias\_Neuron1$$

So, activation of neuron would be Sigmoid(Z1) where sigmoid is one of the functions applied in that hidden layer which will let our output in between 0 and 1. We can use one matrix multiplication to get the calculation received in the first hidden layer.

$$\begin{bmatrix} W1,1 & W2,1 & W3,1 & W4,1 & W5,1 \\ W1,2 & W2,2 & W3,2 & W4,2 & W5,2 \end{bmatrix} \times \begin{bmatrix} X1 \\ X2 \\ X3 \\ X4 \\ X5 \end{bmatrix} + \begin{bmatrix} Bias1 \\ Bias2 \end{bmatrix} = \begin{bmatrix} Z1 \\ Z2 \end{bmatrix}$$

Figure 2: First hidden layer calculation

We should pay attention to one thing that for the neural network, the weighted calculated value in each hidden layer are going to forward to check the value calculated will be similar or not with the output layers. That is the main moto of the neural network. If the value is not correct it could come back again toward the input to change the weight and again it can check the value is correct or not. That is how one machine can learn through neural network. So, here two important properties we should consider illustrating. One is forward propagation and the other one is backward propagation.

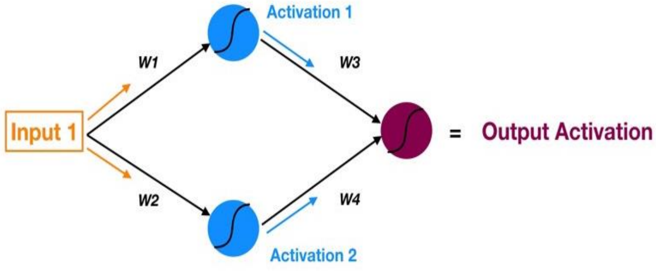


Figure 3: Forward Propagation

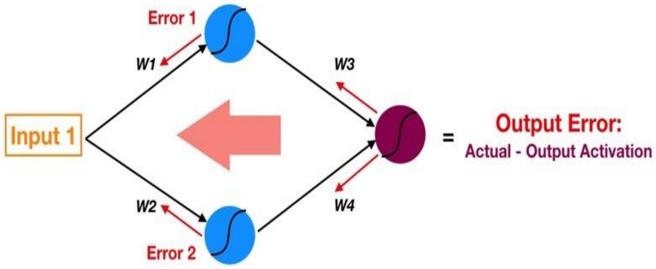


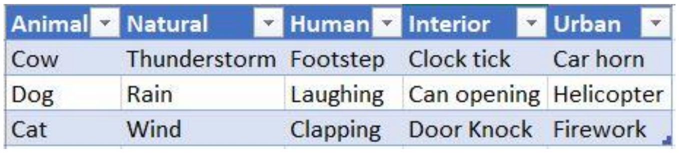
Figure 4: Backward Propagation

As we can see we are quite bother of error calculation while we are doing the backward propagation. It happens because the same is checking with the training dataset and if any significant amount of difference would be found, it could backtrack with an error value and reconstruct the weights involved.

So, here we illustrated the overview of neural network. That is how one machine can learn by neural network concept. We can see here that machine is learning here from its mistake. It can change the value of initial weights to get the perfect output as the training data. Neural network has different types as well. One is ANN named like Artificial Neural Network, CNN named like Convolutional Neural Network, RNN named like Recurrent Neural Network. We tried to implement those three different algorithms in our model and tried to compare which could give us better result.

III. Our Sound Dataset

We already mentioned that we choose such subject are where getting the dataset and using the same is little challenging. As we mentioned that as our project name environmental sound classification is something which says that the dataset itself is about environmental sound. Our dataset is full of 2000 environmental sounds specially distributed in five major categories. Here is a snapshot attached to get better view of our dataset.

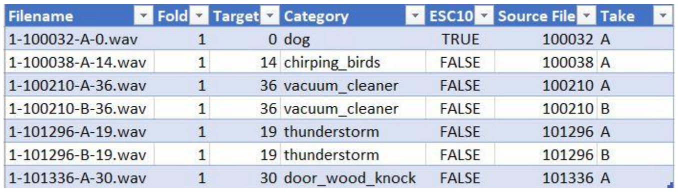


Animal	Natural	Human	Interior	Urban
Cow	Thunderstorm	Footstep	Clock tick	Car horn
Dog	Rain	Laughing	Can opening	Helicopter
Cat	Wind	Clapping	Door Knock	Firework

Figure 5: Dataset Snippet

As illustrated in the above Figure 5, we can get a brief snippet of our dataset where we specifically have five different categories like Animal, Natural, Human, Interior, Urban and for those five categories, we have different kinds of sounds spreaded around. For an example, for animal category we have cow, dog, cat, hen, goat etc. and for Human category we have sounds like Footsteps, Laughing, Clapping, Snoring etc. and so on. In total, we have 2000 sounds around that we got from the Freesound. Let's talk about the data. The environmental sound data (shortly we call the same as ESC dataset) is the collection of short environmental sounds extracted raw from the natural environment which has length of 5 second each and each has 44.1 kHz frequency. Alongside with dataset we got a nice extract of labeled data from Freesound as well to map each sound with each category. The dataset has ESC-50 dataset apart from the 2000 sound clips. One other metadata information is there whose name is ESC-10, one other dataset where we have 400 environmental recordings which is nothing but the subset of the ESC-50 dataset. ESC-10 has 10 classes and 40 categories per classes. We would like to illustrate the metadata information (ESC-50) below.

Figure 6 has seven essential metadata information or columns named Filename, Fold, Target, Category, ESC10, Source File, Take.



Filename	Fold	Target	Category	ESC10	Source File	Take
1-100032-A-0.wav	1	0	dog	TRUE	100032	A
1-100038-A-14.wav	1	14	chirping_birds	FALSE	100038	A
1-100210-A-36.wav	1	36	vacuum_cleaner	FALSE	100210	A
1-100210-B-36.wav	1	36	vacuum_cleaner	FALSE	100210	B
1-101296-A-19.wav	1	19	thunderstorm	FALSE	101296	A
1-101296-B-19.wav	1	19	thunderstorm	FALSE	101296	B
1-101336-A-30.wav	1	30	door_wood_knock	FALSE	101336	A

Figure 6: ESC-50 dataset snippet

Let's illustrate each of the columns of ESC-50 dataset which is really important to make our dataset into a structured format. First column name is Filename which is unique in that table for sure. File name is nothing but the file name of each sound clips in .wav format with a length of five seconds. If we closely look, file name has one pattern which it follows every time. We will talk about that later. Moving forward to the next column which is Fold. We know whenever we train machine something with the data it is always better to use fold structured dataset. It means that our dataset consists five folds altogether. We will train our data with five folds every time. Means, for a homogenous propagation, we use four folds every time as training data and the fifth one as test data. It means that of we have five folds, we can use first, second, third, fourth fold data as training data and the fifth one as test data for the first time. For the second time, we can use first, second, third, fifth fold as training data and fourth one as test data and so on. It gives the machine a smoothing way to learn and folds architecture of dataset increases the perfection of prediction more. Moving forward, third column of ESC-50 dataset is Target. One of the essential columns of the metadata information. As we said earlier, that we have five different categories of sounds, we should mention here that we have fifty different classes in those five categories. More specifically, we have ten different classes for each category which makes the data like fifty different classes. In the ESC-50 dataset, we have zero to forty-nine classes. Those classes represent target in the ESC-50 dataset. Fourth column is Category in the dataset itself which is nothing but one to one map of the third column of target. For an example, Target 0 means dog in the category. Like that, 1 means rooster, 2 means pig, 3 means cow, 4 and so on. Moving forward, fifth column is ESC10. That one is interesting column where we will get to know that the same file is present in ESC-10 dataset or not. Earlier, we discussed a bit of ESC-10 dataset as well where we told that the same is about 400 sound clips of 10 classes and 40 categories. If the sound clip is present in ESC-10 then it would be TRUE otherwise it falls. Here, we should mention that the same column could be the optional too for taring our machine learning model. The sixth column is Source File which is really one of the important columns in ESC-50 dataset which is nothing but a name of the source of the clips. It has some distinct numbers. For an example, if file name is 1-18757-A-4.wav, the column source file will have the data like 18757. For that example, this sound has extracted one particular source which is frog here. But, if we see another example, 1-31836-A-4.wav is a different file name and for that source file is nothing but 32836 which again a frog. So, same kind of sounds, same class of sounds might be extracted from two or more than two different sources. The last but not the least column is Take. Take has data like A, B, C, D. In which take we extracted the sound from the source, the column signifies that. It is possible that we take same sources in different takes to make our data smoother to learn. Now, the time came for briefing the file name pattern. For an example, we are just picking up one file name like 1-50060-A-10.wav. The first '1' here represents the fold number. As we discussed earlier, we have five folds data. '1' indicates the first fold we are considering here. The next numbers like '50060' is the source file from where we extracted the sound. Here, '50060' is one of the environmental sounds which is rain. The third letter is 'A' which says it is nothing, but in which take we extracted the sound. For extracting sound, we used different takes and 'A' is one of them. The fourth number '10' means in which category the file name belongs. '10' indicates target 10 which is rain here. The last '.wav' denotes the format of the sound which is constant here. We extracted the sounds from environment as .wav file format. So, that is how our file name got arranged to identify one particular sound. That is all about our sound data and its metadata information in ESC-50 file.

IV. Overview of Proposed Approach

We should briefly describe what our problem statement is. We have 2000 environmental sounds which we have to classify into 50 classes or sometimes 5 categories. So, the method we have use the classification. As we discussed earlier, we are using neural network to extract the data to train our machine learning models. We introduced our data and its properties in the previous section. The data is sound which is the biggest challenge here to extract the feature. To get the idea of the sound and how sound waves are changing with time format, we plotted some graphs from the environmental sounds. We specially extracted two types of graphs from the environmental sounds to get the overall idea how amplitude, frequency is changing with time changes. It really helped us to understand the characteristics of the sounds itself and the difference of sounds involved, extracted with each other. Two plots specifically we have extracted from the sounds provided by the Freesound dataset. One is Wave Plot; other one is Log Power Spectrogram. Wave Plot is kind of sine graph which changes according to time. The spectrogram here is kind of a new graphs which is specially built for sound waves extraction. It shows the distribution of energy in both time and frequency. Log Power Spectrogram is a bit of update here on the spectrogram. Here are the snapshots attached below to illustrate how those plots looks like. We extracted the spectrogram as well from the urban sound clips to get the overall idea of the sounds features. Here are some snapshots attached. We are attaching the wave plots, spectrogram, log power spectrogram for snoring, clock alarm, frog, helicopter and cricket successively.

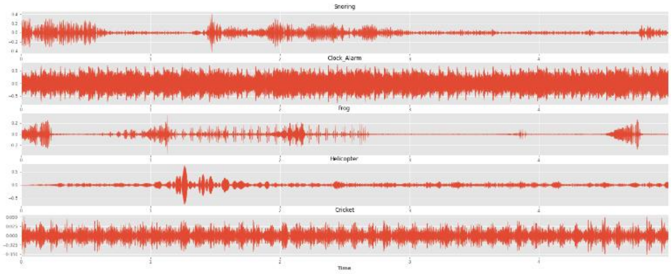


Figure 7: Wave Plot

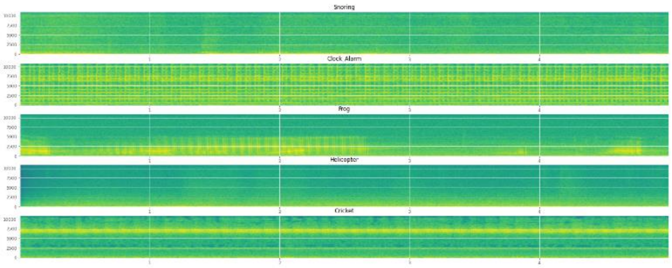


Figure 8: Spectrogram

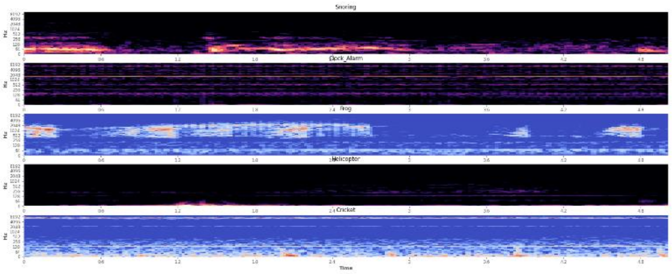


Figure 9: Log Power Spectrograms

Figure 7,8,9 illustrates the wave plot, spectrogram and log power spectrogram respectively. So, these are overall feature extraction procedures we have gone through to check which feature takes important role to classify the sound by the neural network. We would like to mention some important feature extraction methods we use for feature engineering. We got the overall pictorial idea how the frequency looks like for every sound. Some feature extraction methods are MFCC, Chromagram, Melspectrogram, Tonnetz, Spectral Contrast and many more. In our project, we used some of those above-mentioned feature extraction procedures. We would like to describe briefly on each of those feature extraction techniques. MFCC stands for Mel frequency cepstral coefficient, Chromagram describes tonal content of an audio or image signal in a condensed form. Melspectrogram means acoustic time frequency representation of a sound or image. Tonnetz means traditional harmonic relationship. Spectral Contrast shows spectral peaks, the spectral valley and their difference in each frequency. Then we train our model with the extracted features that we got. Some important libraries are Librosa, Matplotlib, Sklearn, Tensorflow and the algorithm we have through first is Artificial Neural Network. Overall, we used 3 layers architecture for our machine learning model. Generally, we have 50 classes. We train our model for 2000-6000 times. We use learning rate in the range of 0.001 to 0.00001 and we used sigmoid in the first hidden layer and the tanh function in the second hidden layer. For application of ANN, we got accuracy of 35.4 with the training epoch like 6000 and learning rate of 0.00001 which is the best among our experiments through ANN. We would like to mention here that cost of the function gets reduced with the increase of training epoch. Average runtime was 20 min around, accuracy was 31.4 percent and precision were 31.37 percent.

V. METHODS

Neural Network Model: For training neural network model, we have extracted following features from audio clips:

Mfcc: Mel-frequency cepstral coefficients,

Melspectrogram: Compute a Mel-scaled power spectrogram

Chorma-stft: Compute a chromagram from a waveform or power spectrogram

Spectral_contrast: Compute spectral contrast, using method defined in [1]

Tonnetz: Computes the tonal centroid features (tonnetz) [3]

We have used Librosa library to extract these useful characteristics from audio files and fed into the neural network model. We have configured our model using these parameters: training epoch = 6000, learning rate = 0.0001, number of hidden units in layer one = 280, number of hidden units in layer two = 300, number of classes = 50, number of dimensions = training features. shape[1], sd = 1/np.sqrt(number of dimensions). We have used sigmoid and tanh functions in first hidden layer and second hidden layer respectively for non-linearity.

Here we are attaching the graph for cost vs iteration.

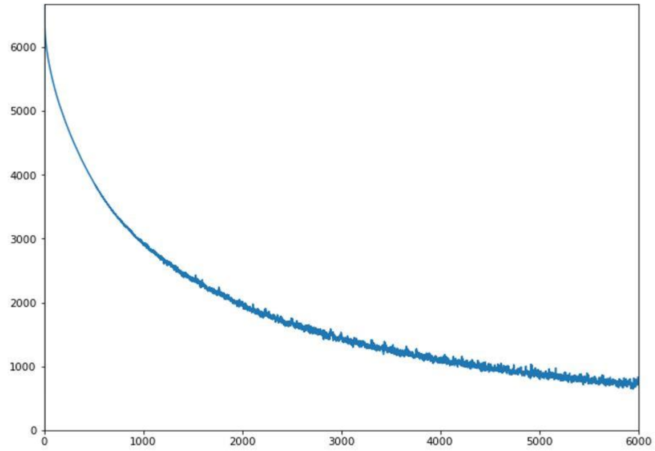


Figure 10: Cost history against number of iterations

Convolutional Neural Network Model: We borrowed the idea of extracting features from audio clips for CNN model from this paper [2] We know that CNN performs better on image dataset. From this paper, we got the idea to extract features from audio dataset to feed as different channels (like values of RGB for an image) into the CNN network.

All sound clips are of 5s, we divided each audio clip into a collection of segments of 60 rows and 41 columns. Then we have calculated log scaled mel spectrogram and corresponding delta values for each segment. Log scaled mel spectrogram and delta will be two channels for our CNN model. Other sound features can be calculated in this way and fed into the model as separate channel. We have calculated this for all clips in a fold, saved them into an array and applied one hot encoding.

Configuration parameters for our model are frames = 41, bands = 60, feature size = 2460(60x41), number of labels = 50, number of channels = 2, kernel size = 30, depth = 20, number hidden = 200, learning rate = 0.001, number of training iterations = 2000.

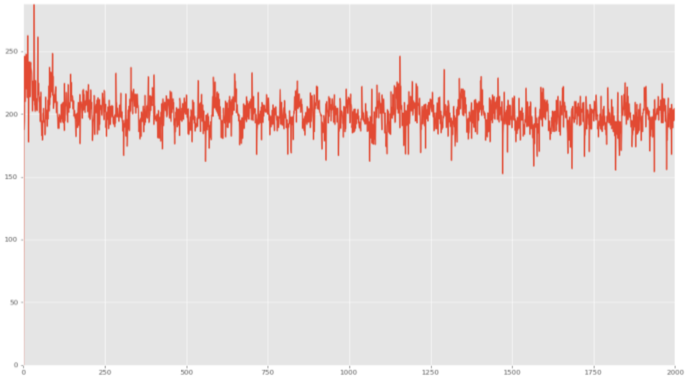
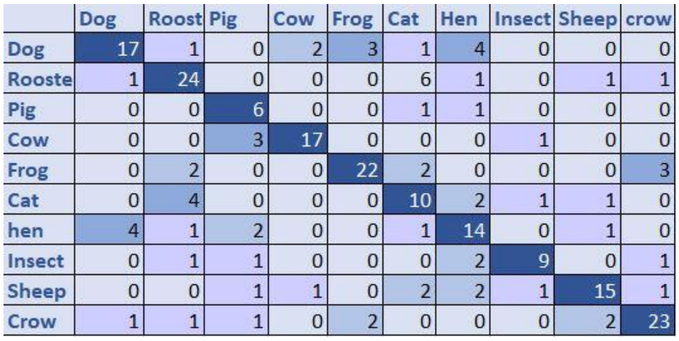


Figure 11: Cost history against number of iterations

Recurrent Neural Network Model: For this model, we have used MFCC feature of sound clips and applied one hot encoding, Configuration for our model is: learning rate = 0.001, number of training iteration = 2000, batch size = 50, display step = 200, number of steps = 41, number of hidden layers= 300, number of classes = 50.

After training the model, we built our confusion matrix. We build our confusion matrix for all of the classes we have. However, we are providing one snapshot of our confusion matrix related to animal class.



	Dog	Roost	Pig	Cow	Frog	Cat	Hen	Insect	Sheep	crow
Dog	17	1	0	2	3	1	4	0	0	0
Rooste	1	24	0	0	0	6	1	0	1	1
Pig	0	0	6	0	0	1	1	0	0	0
Cow	0	0	3	17	0	0	0	1	0	0
Frog	0	2	0	0	22	2	0	0	0	3
Cat	0	4	0	0	0	10	2	1	1	0
hen	4	1	2	0	0	1	14	0	1	0
Insect	0	1	1	0	0	0	2	9	0	1
Sheep	0	0	1	1	0	2	2	1	15	1
Crow	1	1	1	0	2	0	0	0	2	23

Figure 12: Confusion Matrix

From the above-mentioned confusion matrix in Figure 12, we get to know how much accuracy we get for the animal class in our dataset. We are uploading the whole confusion matrix in the link mentioned at the end of the paper.

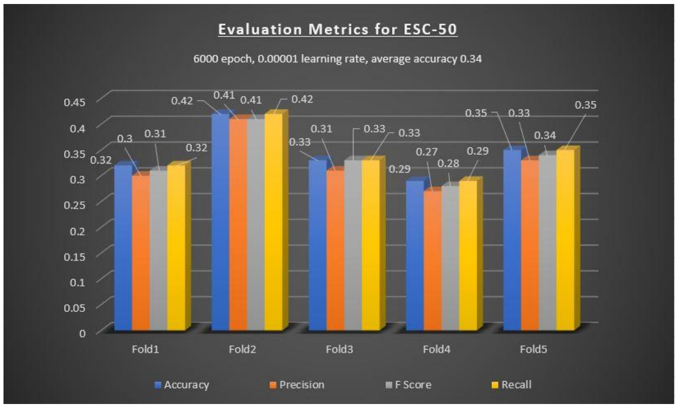


Figure 13: Evaluation Metrics

In Figure 12, we see our evaluation metric for ESC-50 dataset. As discussed, we used fivefold break up of our dataset. For five folds, we plotted Accuracy, Precision, F Score, Recall and the average accuracy is 34.36 percent altogether. We attached link of our directories where learning models codes in python will be there.

VI. Baseline Methods for comparison

Many methods are there to classify the environmental sounds. We would like to talk about Baseline – SVM and Baseline -k NN more specifically. Before that we will start discussing about human classification accuracy. In the reference [16], they made a tally with human sound classification challenge as well. From the paper itself, we got to know that for ESC-50 dataset, human was 81.3 percent accurate to classify 50 class sounds. But the writer said there even though it is accurate enough, the question is still there that for how much dataset it could be so accurate. For the large dataset, it might be possible human could not be such accurate anymore. So, here, it depends on the volume of dataset. It might be possible that for large dataset, one trained machine can predict better than the human. From that human based sound classification, it got proven that human can score more on the small dataset but not for large dataset. Here in that paper [16], we see that they used SVM (Scaler Vector Machine) for feature extraction and taring a machine learning model. They got the accuracy around 39.6 percent. So, here for 2000 environmental sounds, it took like 39.6 percent. Here, for our model for using ANN, we got around 34.36 percent for 6000 epoch and 0.00001 learning rate. For baseline K-NN as well, they got the accuracy of 32.20 percent where our ANN model gives around 34.36 percent. So, from the paper [16], we can say that trained and attentive listeners could score flawlessly on the smaller dataset. But when the dataset is large then machine learning model will take place on the human ability and score more.

VII. Related Work

Before starting the project, we went through many other papers which are more likely on the neural network architecture and some of those are classifications of sounds, images. Specially, we want to recall one of the papers which is based on the Urban Sound Classification. There on the paper, the author gave the idea to use neural network to classify the urban sounds which is a bit of similar working areas mentioned. The name of the paper was “Enhancement of Urban Sound Classification Using Various Feature Extraction Technique” written by Afshankaleem, I. Santi Prabha published in IJRTE, 2019. The paper was quite recent and that is why after getting interest on that specific domain, we considered the paper as reference and tried to establish the neural network machine learning techniques for Environmental Sounds where the writers gave concept on the Urban Sounds. We would like to describe in brief that what they exactly mean to say. They used MFCC (Mel Frequencies Cepstral

Coefficients). It has better continuous process to derive better feature process. We have gone through one other paper named "Environmental Sound Classification based on Multi Temporal Resolution Convolutional Neural Network combining with multi-level features" written by Boqing Zhu, Kele Xu, Dezhi Wang, Lilun Zhang, Bo Li and Yuxing Peng. They have extracted environmental sounds from environment and gave input like raw waveforms as input. On the other hand, the proposed architecture also aggregates hierarchical features from multi-level CNN layers for classification using direct connections between convolutional layers, which is beyond the typical single-level CNN features employed by the majority of previous studies. We compared our ANN proposed architecture with these two above papers. We got accuracy sometimes better than the previous paper or sometimes we got the less accuracy because of number of epochs, learning rates and all.

Different types of approaches have been introduced for extracting features from sounds and classifying those for years. MFCC and zero crossing rate features have been widely used for training general classifiers like Support Vector Machine, Random Forest model, Gaussian Mixture Model for sound classification [2], [4], [5].

Nowadays, neural network models are being used for sound classifications. It has been seen that Convolutional Neural Network (CNN) performs better than others [6] for sound classification. Features which are based on spectral components of sounds like MFCC [7], GTCC [8] are commonly used for training CNN models.

Sound Net [9] has leveraged a deep convolutional architecture model for sound recognition. They have used unlabeled videos and images in their CNN model and applied the discriminative visual knowledge for extracting features for sound classification.

Dictionary based classification has also been introduced [10] for sound dataset. Sound clips are modeled with mixture models and organized in dictionary form. Then each element of the dictionary is compared against each clip of the testing dataset and the best match is considered as the sounds' class.

For image [11] classification and music analysis [12], end to end learning approach is being used. Tycho et al. has applied this method in environmental sound dataset too [13].

VIII. CONCLUSION

As we all know nowadays neural network is one of the biggest research areas, we thought to use the same for our classification problem throughout the environmental sounds. As we discussed earlier, training data with sounds are little challenging. Getting a data is the biggest challenge and then making that unstructured data to the structured one is one of the other challenges. Thanks again to Freesound again for giving the dataset handy and make it structured through the ESC-50 dataset. We used many feature extraction techniques alongside with MFCC feature which has acceptable average accuracy. As of now we applied ANN and CNN to our machine learning models and got the accuracy, precision with different epochs and learning rates. We have many things to do in future as well. We should apply RNN (Recurrent Neural Network) again in our machine learning model and will compare the accuracy, precision values alongside with the other paper works. Other thing is we will try to collect more data and will train our model with more data to find more accuracy, more precision, less error. We will try to handle other feature extraction procedure as well. We have to find any other feature extraction technique which is more accurate to machine learning model.

Please find our link for our entire projects for learning machine to classify environmental sounds. https://drive.google.com/drive/folders/1bCVjcvv5sGtIeLft6mvvJnYRgfgihyyH

IX. REFERENCE

[1] **Music type classification by spectral contrast feature. Dan-Ning Jiang ; Lie Lu ; Hong-Jiang Zhang ; Jian-Hua Tao ; Lian-Hong Cai**

[2] **Environmental sound classification with convolutional neural networks by Karol J. Piczak.**

[3] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using gmm supervectors for speaker verification," ***IEEE signal processing letters*****, vol. 13, no. 5, pp. 308–311, 2006.**

[4]"Enhancement of Urban Sound Classification Using Various Feature Extraction Techniques ", Afshankaleem, I. Santi Prabha, IJRTE, 2019

[5] Y. Tokozume and T. Harada, "Learning environmental sounds with end to-end convolutional neural network," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2017, pp. 2721–2725

[6] Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in ACM International Conference on Multimedia, 2014, pp. 1041–1044

[7] D. M. Agrawal, H. B. Sailor, M. H. Soni, and H. A. Patil, "Novel teobased gammatone features for environmental sound classification," in ***European Signal Processing Conference*****, 2017, pp. 1809–1813**

[8] SoundNet: Learning Sound Representations from Unlabeled Video, Yusuf Aytar∗ MIT yusuf@csail.mit.edu Carl Vondrick∗ MIT vondrick@mit.edu Antonio Torralba

[9] A MIXTURE MODEL-BASED REAL-TIME AUDIO SOURCES CLASSIFICATION METHOD, Maxime Baelde1,2, Christophe Biernacki1, Raphael Greff ¨ 2

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in ***International Conference on Neural Information Processing Systems*****, 2012, pp. 1097–1105.**

[11] J. Lee, J. Park, K. L. Kim, and J. Nam, "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms," ***arXiv preprint arXiv:1703.01789*****, 2017.**

[12] Utilizing Domain Knowledge in End-to-End Audio Processing Tycho Max Sylvester Tax Corti, Jose Luis Diez Antich Hendrik Purwins, Lars Maaløe Corti
[13] https://cs231n.github.io/optimization-2/
[14]https://towardsdatascience.com/understanding-neural-networks-19020b758230
[15]https://pathmind.com/wiki/neural-network, https://pathmind.com/wiki/use-cases
[16] ESC: Dataset for Environmental Sound Classification, Karol J. Piczak Institute of Electronic Systems Warsaw University of Technology Warsaw, Poland K.Piczak@stud.elka.pw.edu.pl