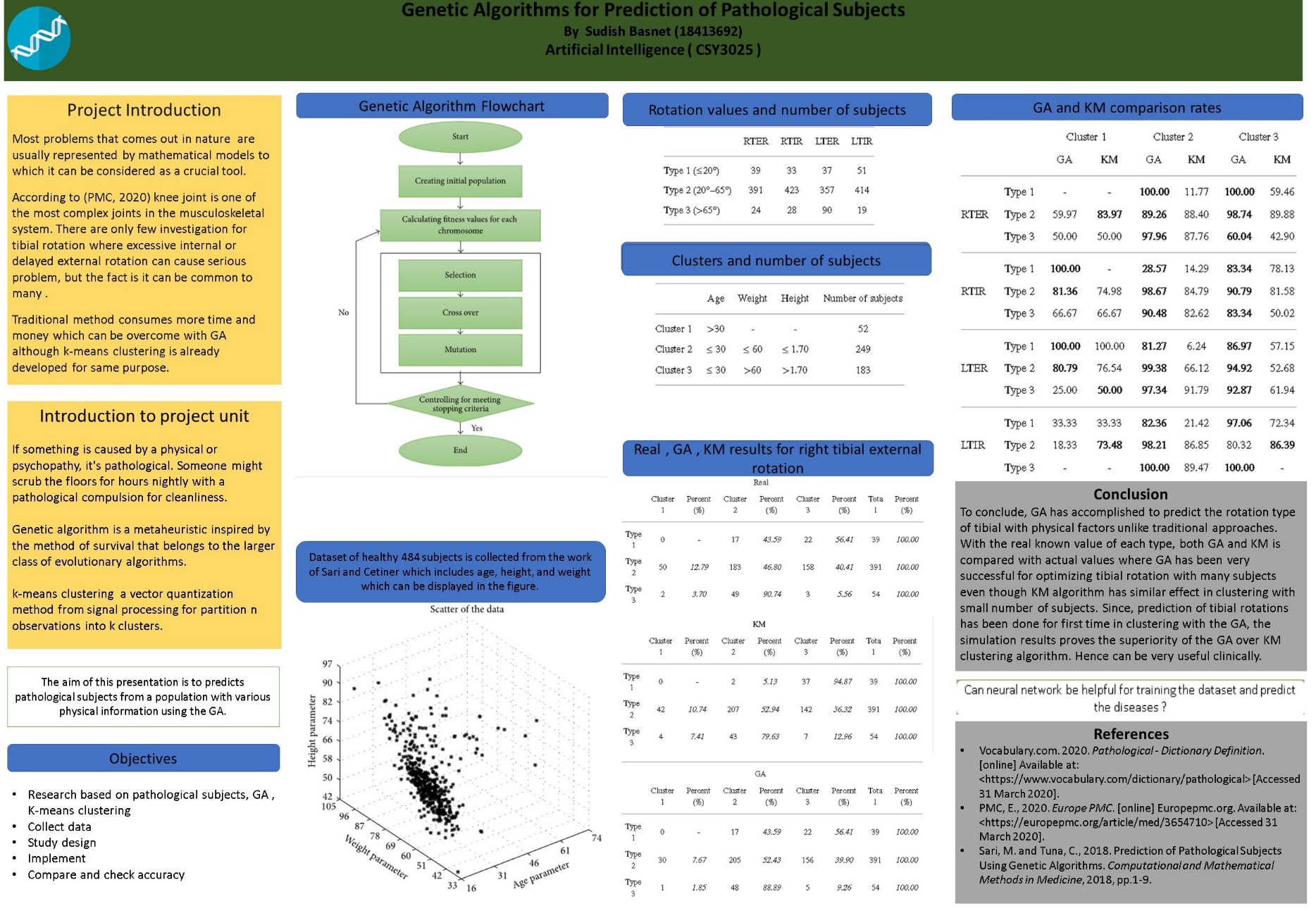
**Sound Recognition with Neural Network**

|  |  |  |  |
| --- | --- | --- | --- |
| **CSY3025 : ARTIFICIAL INTELLIGENCE** | | | |
| Submission Date: | **3rd May 2020** | Module Tutor: | Mamta Bhattarai |
| Student Name: | Sudish Basnet | | |
| Student ID: | 18413692 | | |
| Tutor Comments |  | | |



Contents

[Abstract 3](#_Toc39263673)

[1 Introduction 3](#_Toc39263674)

[1.1 Report Introduction 3](#_Toc39263675)

[1.2 Project Introduction 3](#_Toc39263676)

[1.3 Project Aims and Objectives 4](#_Toc39263677)

[2 Background and literature review 4](#_Toc39263678)

[2.1 Problem Domain 4](#_Toc39263679)

[2.2 Sound Processing 5](#_Toc39263680)

[2.3 Introduction to Neural Network 5](#_Toc39263681)

[2.3.1 Artificial Neural Network 5](#_Toc39263682)

[2.3.2 Convolutional neural network 6](#_Toc39263683)

[2.4 Data Augmentation 7](#_Toc39263684)

[3 Solution Implementation 8](#_Toc39263685)

[3.1 Convolutional neural network for environmental sound classification 9](#_Toc39263686)

[3.2 Artificial neural network for voice and digits recognition 14](#_Toc39263687)

[4 Testing 16](#_Toc39263688)

[4.1 Accuracy Testing 16](#_Toc39263689)

[4.1.1 Accuracy testing for CNN 16](#_Toc39263690)

[4.1.2 Accuracy testing for ANN 18](#_Toc39263691)

[5 Evaluation 21](#_Toc39263692)

[6 Conclusions 21](#_Toc39263693)

[7 References 22](#_Toc39263694)

[7.1 Articles, Journals, Blogs , Research references 22](#_Toc39263695)

[7.2 References for codes 23](#_Toc39263696)

[Figure 1 : Sound Processing 6](#_Toc39263697)

[Figure 2 : Artificial Neural Network 7](#_Toc39263698)

[Figure 3 : Convolutional Neural Network 8](#_Toc39263699)

[Figure 4 : Sound distortion 9](#_Toc39263700)

[Figure 5 : Spectrogram 11](#_Toc39263701)

[Table 1 : Overall Testing Report 21](#_Toc39263702)

# Abstract

This paper justifies various aspects of deep learning applying artificial neural network (ANN) and convolutional neural network (CNN) for sound processing. With the first stage of collecting raw samples with as much information as possible where for environmental sound recognition 10 samples are taken and for voice few samples are taken with spoken digits too. Analysis of prediction with highest possibilities is done in order to know which neural network approach suits most for the case like voice recognition, digits recognition or simply sound recognition. With the accuracy on output which is given by output layer of neural network after providing input to input layers with as much data as possible and computed in hidden layer to determine and train the system with the sample, the highest accuracy approach will be selected out for the purpose. In case of limitation to sound samples, data augmentation will be done and carryout the process accordingly.

# 1 Introduction

## 1.1 Report Introduction

This report covers theoretical solution with some practical work reflection for the sound processing by recognizing voice as well as sound in the background according to need with the help of different neural network approaches as because of its flexibility. Practical solution that have been carried out or works that have been done already is explained detailly in this report with proper references to related works and solution testing. With the introduction to the project, report begins with every step that have been carried out throughout the system solution from background research to implementation and testing whether the solution gives the accuracy high as much as possible with huge raw augmented data set.

## 1.2 Project Introduction

In this busy world full of crowds everyday everyone especially workers related to fields like Banks, Legal Authorities, Telecommunication etc. comes at a point that they need to interact with several peoples and the problem is that with the help of calls or online communication they can’t confirm if the person opposite to them are really them. For the fact that it always the case when people call us to confirm our identity with several questions regarding personal details like contact numbers, mail id, citizen id , addresses etc. as per the organization confirmation data. And the problem is that not either of us can remember our personal information if it comes to citizen id or passport id .And neither the officials who tries to contact us can confirm.

VOICE RECOGNITION BASED ON NEURAL NETWORKS (Zaatri, Azzizi and Rahmani, 2015) also discuss about the use and importance of sound recognition in different aspects.

In such a situation neural network is one of the ways which can help to confirm either the person that officials are talking are the one with some recognition technique which adversely can contribute to another problem which is mainly focused for legal authorities like police to know the background noises which can actually lead to some clue towards their case .

## 1.3 Project Aims and Objectives

Throughout this project three major problem will be solved which are already justified and pointed out below:

* Speaker voice recognition(Biometric authentication)
* Classification of environmental sound (Urban areas sounds especially for police cases)
* Digits recognition after spoken(Numbers from 0-9 for pin codes recognition)

Process that will be followed throughout the system implementation till completion by the user and system itself include:

* Analog to digital conversion (Collect audio data) by user
* Data Augmentation (Due to need of large dataset) by system
* Train system with the dataset either ANN or CNN

After the system train itself with data the further process will be:

* Input dataset to neural network (CNN or ANN) input layer by user
* System process at hidden layer for every computation by neural network
* Outputs layers gives the predicted output by system with neural network
* Accuracy check and evaluation of prediction by user
* Do data augmentation as per the accuracy requirement and re-train as needed

# 2 Background and literature review

## 2.1 Problem Domain

Solving cases for police with traditional move regarding kidnapping , criminal escaping won’t be always a right move if we have lack of time or victim is in critical condition so there need to be something systematic machines to know the background noises or culprit voice as soon as interact with the suspect.

Many of the time there will be some fraud calls to police in order to give them wrong pattern during some case so that voice recognition in association with government to track each individual record can really help.

I f we take some daily examples then for BANK, Insurance companies etc. needs to confirm customers by talking or confirming pin code and the traditional method with some data collected to confirm will not always be useful. There will always be interference of digits when customer say by phone.

The use of voice authentication for a financial institution provides a second, user-friendly way for the customer to authenticate. The bank had been using a 5-digit telephone identification code, and it offered one factor authentication. The problem was that customers keep forgetting them and they were expensive to reissue, Kadar said. (Voice Authentication in the Future for Online Banking, 2020)

According to the journal Voice Identification by Man and Machine: A Review of Research (Bull, 1981), it justifies how the voice recognition in kidnap cases will be useful without interacting with victim.

## 2.2 Sound Processing

The processing of sound signals is a subfield of signal processing dealing with the electronic amplification of audio signals. Generally audio signals are electronic sound wave which are the representations longitudinal waves moving through the space, consisting of rarefactions and compressions . The energy found in audio signals are usually measured in the decibels.

For the purpose to interpret the echo, system need to translate it into binary for this reason sound is usually recorded by a microphone which is then converted into a digital signal and computer will then convert the samples to binary.

Since audio signals can be expressed either in digital or analog format, processing can take place in either domain. Analog processors operate directly on the electrical signal, while digital processors operate mathematically on its digital representation. (Audio signal processing, 2020)

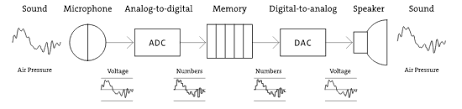


Figure : Sound Processing

## 2.3 Introduction to Neural Network

### 2.3.1 Artificial Neural Network

Neural networks in Artificial Intelligence which stand for ANN are computing frameworks that are ambiguously propelled by the natural or simply biological neural networks which constitute creature brains. Such frameworks learn to execute itself by considering illustrations without being totally programmed, for the most part without being modified with task-specific rules. For illustration, in sound acknowledgment, they might learn to recognize the converted analog to digital signal and carry each single sound form at equivalent frequencies and track the changes with a set of data and learn the word or recognize accordingly.

(Artificial neural network, 2020) In [image recognition](https://en.wikipedia.org/wiki/Image_recognition), they might learn to identify images that contain cats by analyzing example images that have been manually [labeled](https://en.wikipedia.org/wiki/Labeled_data) as "cat" or "no cat" and using the results to identify cats in other images. They do this without any prior knowledge of cats, for example, that they have fur, tails, whiskers and cat-like faces. Instead, they automatically generate identifying characteristics from the examples that they process.



Figure 2 : Artificial Neural Network

## 2.3.2 Convolutional neural network

(Convolutional neural network, 2020) The name “convolutional neural network” indicates that the network employs a mathematical operation called [convolution](https://en.wikipedia.org/wiki/Convolution). Convolution is a specialized kind of linear operation. Convolutional networks are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers.

As compared with ANN it also consists of same layers, but the hidden layers are made of series of convolutional layers, combined with dot products or multiplication which has consequence for indices in the matrix which affects how weight will be set at an index point.

It will be useful for this project because of its nature of discriminating Spectro-temporal patterns, distinction of sound when masked with another sound in time per frequencies.



Figure 3 : Convolutional Neural Network

## 2.4 Data Augmentation

To summarize in simple sentence, data augmentation is the process to create more data sample from single data sample. In general, it is the strategy which enables practitioners in order to increase diversity of data significantly which was for the purpose of training which can be done without sampling new data by the techniques like cropping, flipping or padding depending upon the purpose of need. Data augmentation are always in need when our purpose to train neural network which can be accurate with high numbers of dataset.

(Xiaodong Cui et al.) states that two data augmentation approaches, vocal tract length perturbation (VTLP) and stochastic feature mapping (SFM), are investigated for both deep neural networks (DNNs) and convolutional neural networks (CNNs). In this journal, the main focused approach is increasing speech and speaker variation with limited training data which can be applied to this project too where we are training dataset with ANN and CNN, respectively.

(“INTERSPEECH 2015 Abstract: Ko et Al.”) this paper also specifies the need of data augmentation for projects like ours related to neural networks so that it will help to avoid overfitting as well as improve robustness of the model. The common techniques stated in this paper is to change the speed of the audio data where 4 different LVCSR tasks were given with the data volume available and with the help of data augmentation, average of 4.3% improvement was seen with each phase.

# 3 Solution Implementation

Throughout the implementation to cover the problem domain artificial neural network will be used for voice and digits recognition whereas convolution neural network for environmental sound classification with the sample’s dataset of few peoples and urban areas sounds like children playing, dog barking etc.

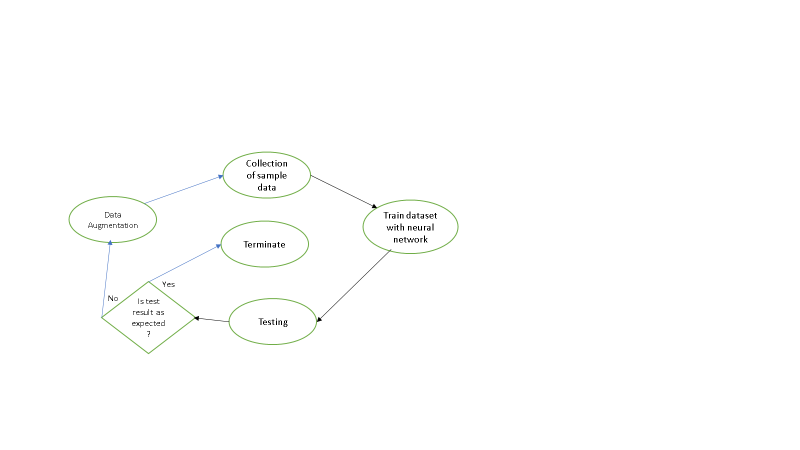
Though either ANN or CNN can be used individually but to test the accuracy for the problem we can train the system with both neural network and use accordingly. (Cowling and Sitte) also states the use of ANN for environmental sound recognition which is compared in this paper with dynamic time warping classification and learning vector quantization techniques. In case of CNN (Su et al.), this paper justifies it is appropriate for environment sound taxonomic problems and dramatically outperform other conventional methods with experimental result data. So that this project focuses on ANN and CNN to decide the performance for sound recognition depending upon various aspects.

At first data augmentation is done to create many data due to the requirement of neural network to train the system where the simple diagram can be shown below for data augmentation where slight distortion is made to original sound.



Figure 4 : Sound distortion

System process that will be followed during the implementation to testing is shown with below system cycle.



## 3.1 Convolutional neural network for environmental sound classification

In this section convolution neural network will be used because of it feature of discriminating Spectro-temporal patterns and dataset of sound will be used from UrbanSound8k with 10 categories which are as below:

* Playing Children
* Engine running
* Street music
* Horns
* Drilling machines
* Siren
* Gun shots
* Air conditioner
* Jackhammer
* Dog barking

At first in order to visualize audio or simply sound signals we must convert it into spectrogram, which is 3d visualization of frequencies over time which can be visualize in below figure:

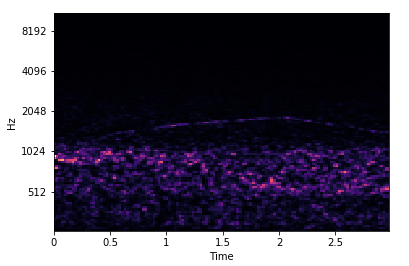
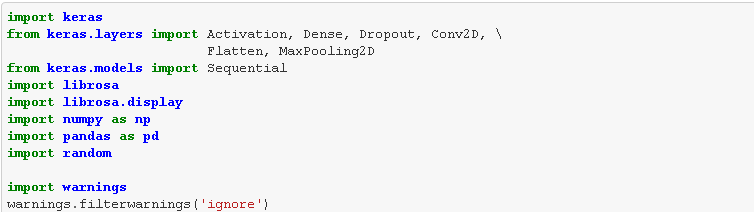


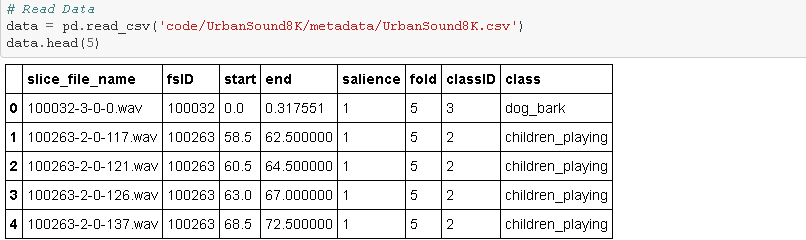
Figure 5 : Spectrogram

Code referenced from (ajhalthor/audio-classifier-convNet, 2020) Starting with importing libraries like keras , librosa to read write audio files , NumPy for mathematical implementation, panadas for reading meta data , random to shuffle data .

Keras is one of the open-source libraries for neural network which is written in Python which is pretty much capable of running on the top of TensorFlow for this system. TensorFlow in other hand is also one of the open-source math libraries for dataflow and several programming with a range of task and used for machine learning application which can be used here for neural network.



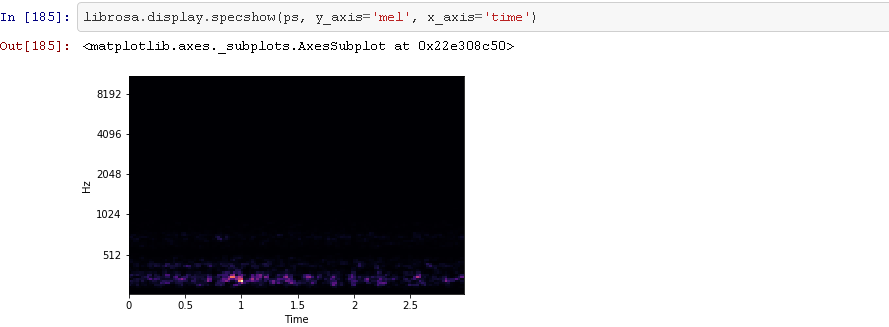
Reading .wav data from csv file in order to train the system and each tuple below have the label name, start and end time of audio where each audio file is almost 3 sec long.



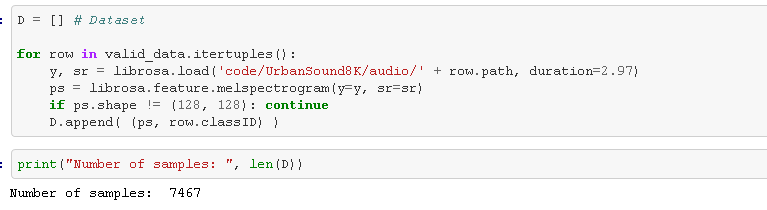
Now validating data length with given time and then using librosa to convert audio files into log scale spectrogram considering the first 3 seconds of the file uniform.



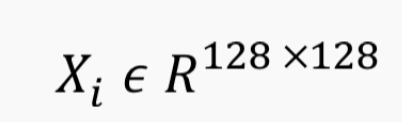
Below is the spectrogram for air conditioner where brighter the light louder the sound.



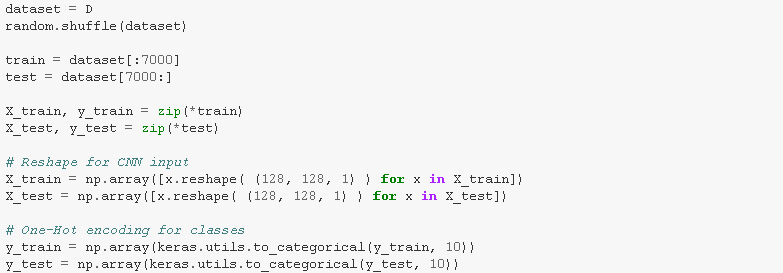
Collecting number of samples in a loop for metadata with the uniform duration of 3 seconds. Below are the testing samples 7467 in number which is huge. The Mel spectrogram is made by dividing frequency over 128 components and similarly for time dividing the sound of 3 sec into 128 frames of 23 Ms.



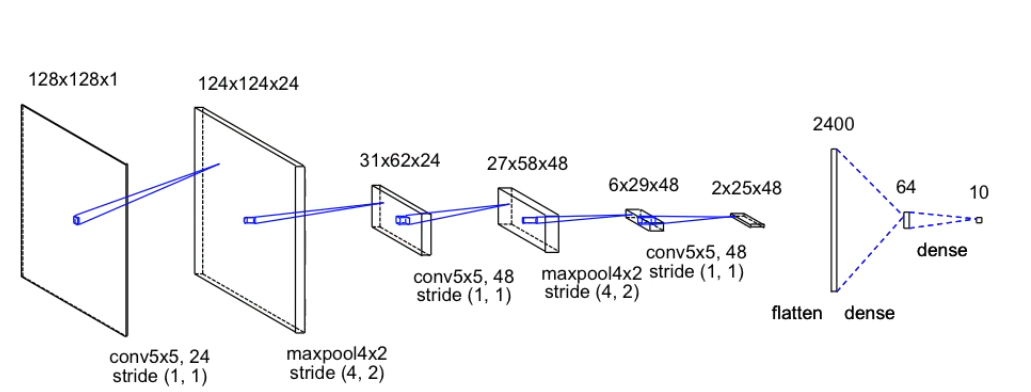
And the overall input is :



Now shuffling the dataset then split up test and train set. After then reshaping the inputs to 128 cross 128 cross 1 from CN n input and encode the labels of class with hot encoding for every ten classes.



And then using the reconstructed model from the research paper (Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification,2017) for conventional neural network which is show below:



From the above figure at first, we have 128 cross, 128 log scale Mel-spectra grounds as input where convolution layer is introduced at first to perform convolution with 24 5 cross 5 kernels with the stride of one. With these cross and kernels while convolution, time frequency signatures will be discovered in input which are distinct for sounds and help us distinguish different sound without the effect of noise.

In the above figure the last layer contains ten neurons due to our classes of output sounds and hereby convolutional neural network is defined. Every function discussed is coded below



## 3.2 Artificial neural network for voice and digits recognition

In this section with help of Keras model, speech classification will be done by working with .WAV files which are of three people and extract MFCC features to upload in dataset to train system. Many of the process followed during the testing and train is as same as CNN but in ANN each neuron is connected to every neuron which is no happening in CNN where only last layer is connected.

Code reference from (ravasconcelos/spoken-digits-recognition, 2020) Step 1 is after collecting the sample ,extracting the data files for cleaning and transformation task is done to extract Mel-frequency cep strum (MFCC) feature to transform into number array which is then exported to csv file.

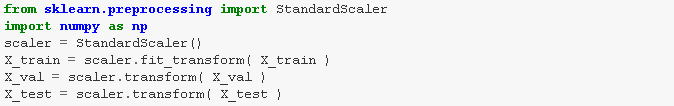
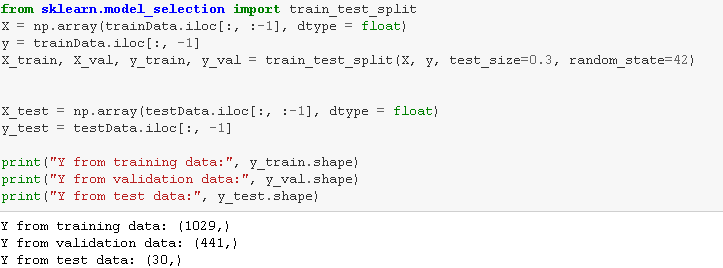
MFCC helps us to identify the audio on the basic of vocal track of the speaker where if the shape is determined accurate than any sound can be represented correctly.



Step 2 is choosing the module and train the data.

After confirming the dataset , it should be split out into training , testing and validation. Afterwards data should be normalized which gives the shape of data as below:

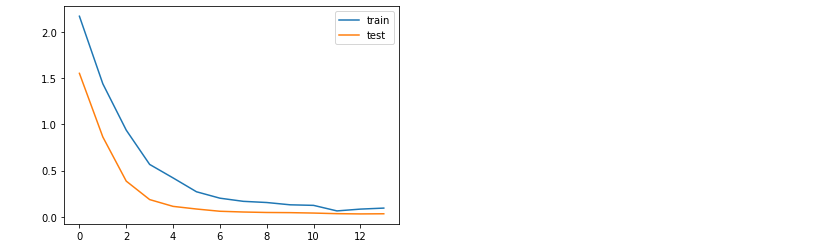
* Data for training(1029, 25)
* Data for validation(441, 25)
* Data for test(30, 25)



After finalizing the shape model should be created which is keras for both CNN and ANN. Nearly the implementation process is same as ANN beside the operations in hidden layer. The data can be used for train after creating model and then giving command with early stopping for avoiding overfitting during training.



Below is the plot for training history



# 4 Testing

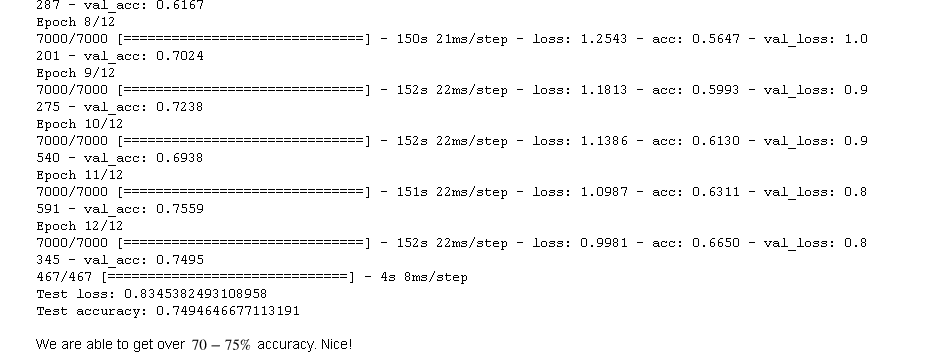
## 4.1 Accuracy Testing

### 4.1.1 Accuracy testing for CNN

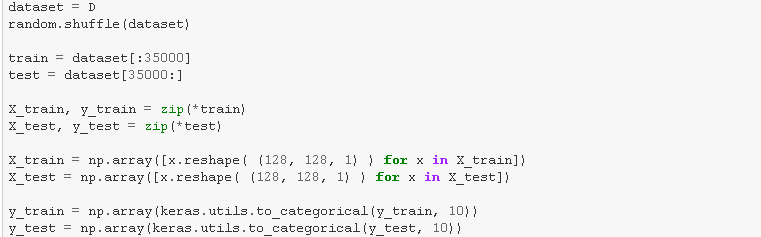
In this phase after evaluating the system with current dataset and compiling model with atom optimizer and measuring the loss with cross entropy loss which is coded out below where model learns every 128 samples:



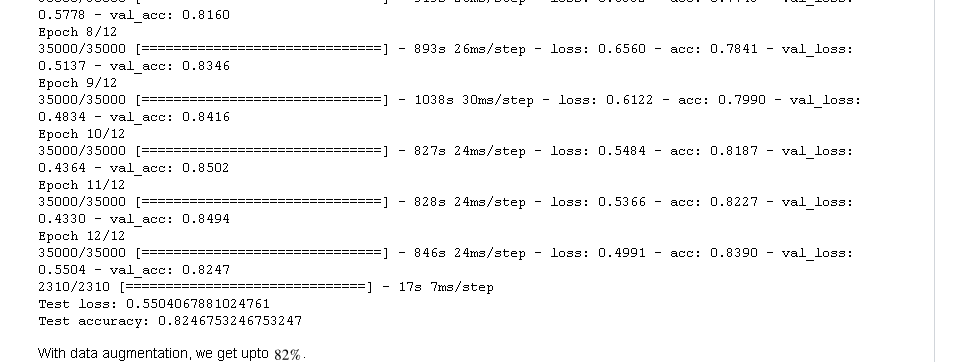
So, the current accuracy is between 70-75% as shown below:



This accuracy can be improved with data augmentation using two method which are varying time or pitch of sample for more sample. Below is the new dataset of 35000 from previous 7000.



And the new accuracy is 82% which is much better with increase in samples.

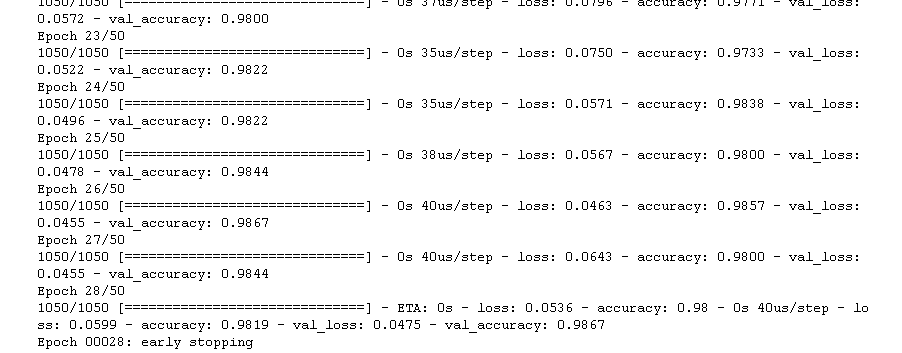


### 4.1.2 Accuracy testing for ANN

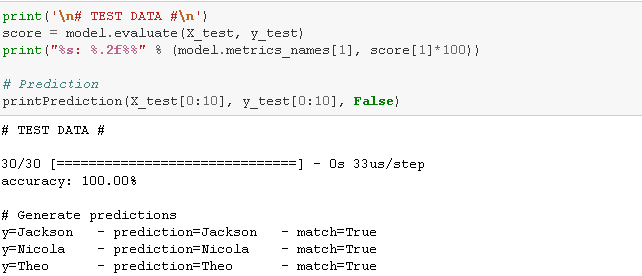
Testing the accuracy for the dataset accuracy



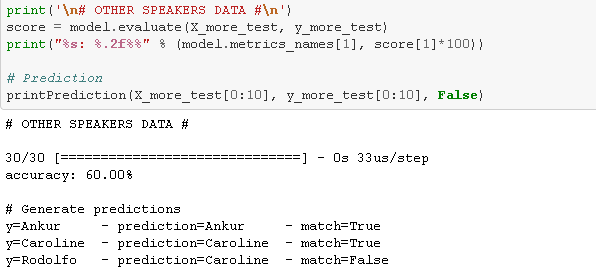
So, the accuracy for the test is 98% which is good



The accuracy for 3 people is 100% due too necessary amount of sample data.



Similarly, for other speakers, accuracy is 60% which can be uplifted with data augmentation.



Overall testing result for the implementation and testing data with neural network considering CNN and ANN for different purpose is justified in the below table with the overall report .

Table 1 : Overall Testing Report

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Neural Network Type** | **Case** | **Train Data** | **Epoch** | **Expected Accuracy** | **Accuracy Computed** |
| **Artificial Neural Network** | 1. Speaker voice recognition 2. Spoken digits recognition | |  | | --- | | 1029 | | 1050 |  |  | | --- | | 1029 | | 1050 | | |  | | --- | | 1/50  15/50  29/50 | | 1/50  15/50  29/50  44/50 |  |  | | --- | | 1/50  8/50  14/50 | | 1/50  22/50  28/50 | | |  | | --- | | Around 40%  Around 90%  Around 90% | | Around 40%  Around 90%  Above 90%  Around 100% |  |  | | --- | | Around 70%  Above 95%  Around 100% | | Around 80%  Above 95%  Around 100% | | |  | | --- | | 40.82%  88.66%  92.97% | | 38.67%  87.56%  94.44%  99.11% |  |  | | --- | | 70.29%  98.87%  99.09% | | 82.89%  98.00%  98.67% | |
| **Convolutional Neural Network** | 1. Environmental sound recognition | |  | | --- | | 7000 | | 35000 | | |  | | --- | | 1/12  8/12  12/12 | | 1/12  8/12  12/12 | | |  | | --- | | Around 30%  Around 70%  Around 80% | | Around 45%  Above 80%  Above 80% | | |  | | --- | | 32.33%  70.24%  74.95% | | 48.44%  83.46%  82.47% | |

# 5 Evaluation

In this section, all the research activity and the process used for this system will be reflected with every points of limitations and how to overcome such limitation is explained. The process first started with background research on sound processing and how events takes place I between sound processing so that it can help to justify how the process can be involved in neural network. After the proper research for types of neural network is done through which ANN and CNN came to be effective and then to implement the system knowing about programming languages was important where MATLAB and Python was severally explored through which phyton seems to be very effective and fat for this case so with the help of python in TensorFlow with use of keras the whole process is carried.

Data collection for each equivalent event was difficult due to limited amount of data where it was not easy to train with the available database so that further research on data augmentation was done due to which it was easy to train the system with the same dataset but with different frequency on which it was played although it took a lot of space from system so the training result was not completed as expected with 100% but if available more time then it can be used with more dataset which will result on accuracy of 100% although the present result of accuracy isn’t bad for this case as even above 60% accuracy found to be helpful accordingly.

# 6 Conclusions

With the implementation and testing of both ANN and CNN, neural network justified the problem domain . Though the accuracy rate of CNN for sound recognition is 82% and ANN wins by 98 % for all and 100% for 3 people, it’s not the case because CNN got less accuracy due to less amount of dataset which can be improved more with more dataset with data augmentation whereas considering the fact that only the 3 people got 100% accurate it is because of the suitable amount of required sample for training while other 3 got 60% accuracy.

Voice recognition and digit recognition was done for this project with ANN and environmental sound recognition with CNN but the fact that can be explored is that the process is same for all the three cases either to implement on ANN or CNN with the use of different label name where in voice recognition each speaker is labeled by alphabet where in digit recognition each number will be labelled from 0-9 number and similarly for environmental sound recognition.

Overall neural network is pretty much useful while it comes to sound processing either its CNN or ANN both the types justified beautifully with given dataset to train. Moreover, CNN is easier to build if looked at the implementation for this project but the sample data required for it is high due to which ANN is preferable and from the above section Artificial neural network is preferable when it comes to recognition of sound with few sample data. The only limitation found was data sample which is already explained how it can be lift and if given more time then the new strategy for all three event of sound recognition will be implemented with ANN as it found to be very effective and need less amount of dataset.

# 7 References

## 7.1 Articles, Journals, Blogs , Research references

* En.wikipedia.org. 2020. *Artificial Neural Network*. [online] Available at: <https://en.wikipedia.org/wiki/Artificial\_neural\_network> [Accessed 27 March 2020].
* En.wikipedia.org. 2020. *Convolutional Neural Network*. [online] Available at: <https://en.wikipedia.org/wiki/Convolutional\_neural\_network> [Accessed 27 March 2020].
* Li, S., Yao, Y., Hu, J., Liu, G., Yao, X. and Hu, J. (2018). An Ensemble Stacked Convolutional Neural Network Model for Environmental Event Sound Recognition. *Applied Sciences*, [online] 8(7), p.1152. [online] Available at: <https://www.mdpi.com/2076-3417/8/7/1152> [Accessed 27 March 2020]
* Bankinfosecurity.com. 2020. *Voice Authentication In The Future For Online Banking*. [online] Available at: <https://www.bankinfosecurity.com/voice-authentication-in-future-for-online-banking-a-254> [Accessed 28 March 2020].
* Bull, R., 1981. Voice Identification by Man and Machine: a Review of Research. *Psychology in Legal Contexts*, pp.28-42.
* Salamon, Justin, and Juan Pablo Bello. “Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification.” *IEEE Signal Processing Letters*, vol. 24, no. 3, Mar. 2017, pp. 279–283, arxiv.org/abs/1608.04363, [online] Available at :<10.1109/lsp.2017.2657381> [Accessed 28 March. 2019].
* Parascandolo, Giambattista, et al. “Recurrent Neural Networks for Polyphonic Sound Event Detection in Real Life Recordings.” *IEEE Xplore*, 1 Mar. 2016, [online] Available at :<ieeexplore.ieee.org/abstract/document/7472917>. [Accessed 28 March 2020].
* Inoue, Tadanobu, et al. “DOMESTICACTIVITIESCLASSIFICATIONBASEDONCNNUSINGSHUFFLINGAND MIXINGDATAAUGMENTATION.” *Detection and Classiﬁcation of Acoustic Scenes and Events 2018*, 2018.
* Xiaodong Cui, Goel, V. and Kingsbury, B. (2015). Data Augmentation for Deep Neural Network Acoustic Modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(9), pp.1469–1477.
* Zaatri, A., Azzizi, N. and Rahmani, F., 2015. Voice Recognition Technology Using Neural Networks. *Journal of New Technology and Materials*, 5(1), pp.27-31.
* En.wikipedia.org. 2020. *Audio Signal Processing*. [online] Available at: <https://en.wikipedia.org/wiki/Audio\_signal\_processing> [Accessed 28 March 2020].
* Xing, Y. F., et al. “Sound Quality Recognition Using Optimal Wavelet-Packet Transform and Artificial Neural Network Methods.” *Mechanical Systems and Signal Processing*, vol. 66–67, 1 Jan. 2016, pp. 875–892, www.sciencedirect.com/science/article/abs/pii/S0888327015002241, 10.1016/j.ymssp.2015.05.003. [Accessed 28 March 2020].
* *Sound Processing Unit, Sound Processing System, Audio Output Unit and Display Device*.[online] Available at:< patents.google.com/patent/US20030185400A1/en>. [Accessed 29 March 2020].
* *Sound Processor, Sound Processing Method, and Computer Program Product*. [online] Available at:<patents.google.com/patent/US9014383B2/en>. [Accessed 29 March 2020].
* “INTERSPEECH 2015 Abstract: Ko et Al.” *Isca-Speech.Org*, 2015, [online] Available at :<www.isca-speech.org/archive/interspeech\_2015/i15\_3586.html> [Accessed 30 March 2020].
* Su, Yu, et al. “Environment Sound Classification Using a Two-Stream CNN Based on Decision-Level Fusion.” *Sensors*, vol. 19, no. 7, 11 Apr. 2019, p. 1733, www.ncbi.nlm.nih.gov/pmc/articles/PMC6479959/pdf/sensors-19-01733.pdf, 10.3390/s19071733. [Accessed 31 March 2020].
* Cowling, Michael, and Renate Sitte. “Comparison of Techniques for Environmental Sound Recognition.” *Pattern Recognition Letters*, vol. 24, no. 15, Nov. 2003, pp. 2895–2907, 10.1016/s0167-8655(03)00147-8
* 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.

## 7.2 References for codes

* GitHub. 2020. *Ravasconcelos/Spoken-Digits-Recognition*. [online] Available at: <https://github.com/ravasconcelos/spoken-digits-recognition> [Accessed 28 March 2020].
* GitHub. 2020. *Ajhalthor/Audio-Classifier-Convnet*. [online] Available at: <https://github.com/ajhalthor/audio-classifier-convNet> [Accessed 28 March 2020].