Text Independent Speaker Identification Model Using Deep Learning

Mahabuba Sultana and Jannatul Ferdous

A Thesis in the Partial Fulfillment of the Requirements

for the Award of Bachelor of Computer Science and Engineering (BCSE)

Department of Computer Science and Engineering

College of Engineering and Technology

IUBAT – International University of Business Agriculture and Technology

Summer 2022

Text Independent Speaker Identification Model Using Deep Learning

Mahabuba Sultana and Jannatul Ferdous

A Thesis in the Partial Fulfillment of the Requirements for the Award of Bachelor of Computer Science and Engineering (BCSE)
The thesis has been examined and approved,

Prof. Dr. Utpal Kanti Das
Chairman and Professor

Dr. Hasibur Rashid Chayon
Coordinator and Associate Professor

Md. Saidur Rahman
Supervisor and Assistant Professor

Department of Computer Science and Engineering

College of Engineering and Technology

IUBAT – International University of Business Agriculture and Technology

Summer 2022

Letter of Transmittal

3 September 2022
The Chairman
Thesis Defense Committee
Department of Computer Science and Engineering
IUBAT–International University of Business Agriculture and Technology
4 Embankment Drive Road, Sector 10, Uttara Model Town
Dhaka 1230, Bangladesh

Subject: Letter of Transmittal.

Dear Sir,

It gives us enormous please to submit the paper on "**Text Independent Speaker Identification Model Using Deep Learning**" as per instruction. We expect this paper to be informative as well as comprehensive.

While conducting the thesis paper, we have gathered lots of knowledge about machine learning. We have tried our level best to collect the relative information as comprehensive as possible in preparing the thesis. During preparation of the report we have experienced practically a lot that will help us a great in my career. We will be able to explain anything for more clarification if necessary.

We would like to thank you, for giving us the opportunity to do a report on the above mentioned topic

Yours sincerely,

Mahabuba Sultana	Jannatul Ferdous
18303045	18303002

Student's Declaration

We hereby declare that this thesis is based on results obtained from our own work. All the materials that were used for the purpose of completing this thesis are duly acknowledged and mentioned in reference. This thesis, neither in whole nor in part, has been previously submitted to any other University or Institute for the award of any degree or diploma. We carried our research under the supervision of Md. Saidur Rahman.

Mahabuba Sultana	Jannatul Ferdous
18303045	18303002

Supervisor's Certification

This is to certify that Thesis report on "**Text Independent Speaker Identification Model Using Deep Learning**" has been carried out by Mahabuba Sultana bearing ID# 18303045 and Jannatul Ferdous bearing ID# 18303002, of IUBAT – International University of Business Agriculture and Technology as a partial fulfillment of the requirement of thesis defense course. The report has been prepared under my guidance and is a record of the accomplished work carried out successfully. To the best of my knowledge and as per their declaration, no parts of this report has been submitted anywhere for any degree, diploma or certification.

Now they are permitted to submit the report. I wish them success in all their future endeavors.

Md. Saidur Rahman
Assistant Professor

Department of Computer Science and Engineering

IUBAT–International University of Business Agriculture and Technology

Abstract

Speaker identification methods are widely used in speech authentication, security and surveillance, electronic voice eavesdropping, and identity verification. This study examines the use of speech recognition to improve the security of people's lives and property while also making door systems more accessible to those with impairments. The objective of this research study is to create a door access control system that makes use of voice recognition algorithms to give consumers a quick way to unlock their doors while also ensuring their safety and security. Numerous research have demonstrated how well MFCC features work in accurately identifying speaker. Due to their capacity to effectively capture the repeating character of signals, short-time features like PLP coefficients and MFCC have been used in the majority of studies on speaker recognition. In order to identify the speaker, a unique architecture is presented in this study that uses a convolutional neural network (CNN) for classification and the mel frequency cepstral coefficient (MFCC) for feature extraction. This architecture is used in a text-independent setting. The system's ability to learn elements that are helpful for categorization is crucial to any text-independent speaker identification process. We created a dataset with 700 audio clips, 3 known classes of 3 people with 150–200 audio clips each, and 1 unknown class of 30 people with 5 voice samples for each speaker. Our proposed method achieved a 98% accuracy rate. This technique could be used for various studies that are related to the study of speaker identification with further optimization.

Acknowledgment

We would like to offer our sincere gratitude towards my thesis supervisor, Mr. Saidur Rahman for his patient guidance and constant support for our work. In addition to that, we would also like to thank sir for his critical analysis and regular feedback of our work's progress, which guided us to remain in the aligned path to our goal. Furthermore, we would like to extend our gratitude to all the other faculty members of IUBAT, Department of Computer Science and Engineering who played a vital part in the development process through the whole program. We are also grateful to all the resources found on the internet which helped us to grasp a better understanding of machine learning. "Super Data Science", a website which contains machine learning tutorials, proved out to be really helpful along with the sites related to group forum and discussions.

Table of Contents

List of Figures

List of Tables

Chapter I: Introduction

1.1 Background and context

Numerous high-tech devices are slowly making their way into our daily lives in this information age, changing the way we live and behave dramatically. Some traditional authentication procedures, which employ passwords or pin numbers to authenticate users, have been gradually overtaken by biometrics identification technology, which offers us simpler and more practical means to identify people. But because they might be mistyped or forgotten, passwords and pin numbers are no longer thought to provide a high level of protection. Two examples of biometric identification systems are the face recognition technologies employed in airport terminals1 and the iPhone's Siri speech assistant. On the one hand, speaking has always been the simplest and most natural way for us to express our thoughts, interact with others, and interact with one another. Therefore, a better way to actually simplify our daily life would be to identify people by their dialogue voice and content and then offer the appropriate services. Speech recognition technology has been well developed and used in our daily lives up until this point.

In general, two methods text-dependent and text-independent can be used to identify speakers. The text uttered during testing and training must exactly match for the text-dependent speaker identification system to function properly. In contrast, the text-independent speaker identification system does not require the speaker to speak any text in order to identify the speaker. In this study, the text-independent speaker identification task is taken into consideration due to its applicability in the speech technology that is currently being developed.

The task of supervised learning is speech recognition. The audio signal will be the input for the speech recognition issue, and we must predict the speaker from their audio signal. Since there would be a lot of noise in the audio signal, we cannot feed the raw audio signal into our model. It has been found that using features extracted from the audio signal as input to the basic model would result in significantly better performance than using the raw audio signal as input. The most popular method for removing characteristics from an audio signal is called Mel-Frequency Cepstrum Coefficients (MFCC). In this study, we have used MFCC method to get the feature extraction of speech signal. Finding a value or vector that can be used to identify an object or a person is the process of feature extraction. Because it is thought to be quite good at expressing signal, MFCC is the method that is most frequently utilized in the field of voice processing.

The Gaussian Mixture Model-Universal Background Model (GMM-UBM) and the i-vector are two examples of conventionally effective models for speaker verification. Since they were not objectively trained for speaker verification setup, these models' fundamental drawback is their unsupervised nature. In this study, we propose a CNN architecture for automatically identifying speakers without the need for text. Using a brief speech fragment, the main goal is to distinguish one speaker from a large group of others. Deep CNNs, which were initially created for computer vision problems, are the subject of most current research. Additionally, the majority of the present speaker identification techniques require query phase audio samples longer than 3 seconds to achieve high accuracy. For voice and speech-related classification problems, we developed a CNN architecture.

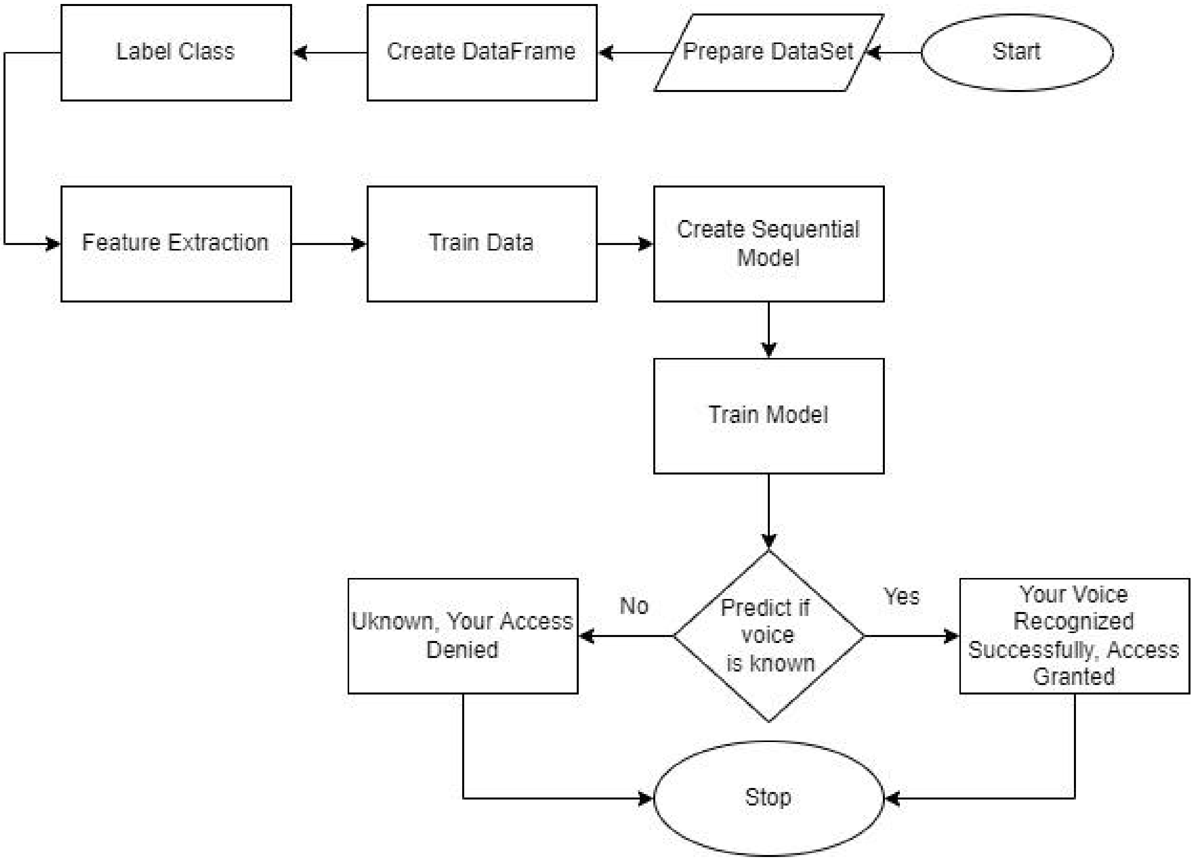


Figure 1: Workflow

1.2 Problem Statement

Security is one of the top concerns for any person or company, and as technology has developed through time, many methods have been employed to safeguard people and property through door access control systems. The conventional method of opening a door is by turning the door knob or using a physical key to open the door. Physical keys that are used to unlock doors are prone to duplication and might be lost by people. A person's finger can be cut off to do a fingerprint scan, a pin can be hacked using various methods or permutations, and a person's photo can be used for face recognition, among other failures, with common biometric technologies and other technologies. Furthermore, it is harder for those who have physical limitations. For instance, it is challenging for a person in a wheelchair to open a door system without additional assistance or help from someone else. Therefore, the requirement for a speech recognition access control system that can accommodate both those with disabilities and those who are able-bodied is unavoidable.

1.3 Research Aim and Objectives

The goal of this project is to build and implement a door access control system with voice recognition functionality, in order to make it easier to grant access to door systems and to provide protection for people's lives and property. The specific objectives of this research study are to:

- User verification: To create identities in the digital age, user verification goes beyond conventional and physical techniques of authentication. User verification can frequently be carried out using speech recognition. On users for whom the system has not yet been taught, our model's accuracy in user verification can be greatly improved. Model will be able to identify the speaker's audio sample with high accuracy.

- Biometric authentication: In comparison to traditional biometrics, speaker recognition has a variety of benefits, including low cost, high acceptance, and non-invasive voice acquisition. A speaker identification system can be built without the need of expensive equipment or direct speaker participation. The use of a debit card, credit card, keeping track of a bank account password or any other security measures, as well as many other online services, might be replaced by speaker recognition.

- Security: Speaker recognition could be used as an authentication mechanism in credit card transactions in conjunction with other verification techniques like facial recognition. Among other things, speaker recognition technology can be used for long-distance speech authentication, monitoring, and computer access control. Automatic speaker recognition, also referred to as speech biometrics, is a trustworthy method for confirming an individual's distinct identity for safe device access control. Automatic speaker recognition technology is used to identify who is speaking rather than what is being spoken. In order to offer frictionless security and personalization, speaker recognition is also being used in conversational interfaces, messaging apps, and IoT devices such as smart speakers and linked cars.

1.4 Significance of the study

- When compared to other biometric systems like iris and fingerprint scan, can speech recognition technology offer a better usability and performance?

- Would this study take into account a person's speech tone in relation to their moods? For instance, a person's vocal tone differs depending on whether they are just waking up or feeling angry.

1.5 Research Question

- When compared to other biometric systems like iris and fingerprint scan, can speech recognition technology offer a better usability and performance?
- Would this study take into account a person's speech tone in relation to their moods? For instance, a person's vocal tone differs depending on whether they are just waking up or feeling angry.

Chapter II: Literature Review

In the global tradition, people communicate with one another by speech. Speaker identification is the process of identifying speakers using the acoustic characteristics of the human voice. Due to the variety of uses for speaker recognition, including forensic voice verification for suspect identification by government law enforcement organizations, speaker recognition has become a focus of considerable research. Due to the enormous impact it has on a speaker identification classification model's performance, feature extraction is crucial to the speaker recognition process. Researchers in the field of speaker recognition have recently put forth innovative elements that have been effective in categorizing human voices.

Fong (2013) conducted a comparison study to categorize speakers using different time-

domain statistical variables and machine learning classifiers; they found that the multilayer perceptron classifier had the highest accuracy of about 94%. Because the experimenters used just 16 speaker voices from the PDA speech dataset, even if the experimental results of the study achieved good classification accuracy, the results cannot be applied to a larger scale. Additionally, in both the training and testing sets of the study, a scant number of speaker utterances were utilized [1].

Ali (2018) recently put up a speaker identification model that uses the dataset for the

Urdu language to recognize 10 different speakers. The support vector machine (SVM) approach was utilized in the study to combine MFCC and deep learning-based characteristics for speaker classification. The experimental findings have a classification accuracy of 92%. Therefore, the outcomes are encouraging. The dataset employed in the tests, however, has a number of flaws. First, the studies only included 10 speaker utterances. Second, there was only one word in each utterance. As a result, the authors' suggested fusion-based features may be inefficient and ineffectual for complex human voices [2].

Soleymanpour (2017) examined at the use of an ANN classifier and clustering-based

MFCC features to categorize 22 speakers from the ELDSR dataset. The study's experimental findings had a classification accuracy of 93% [3].

Nidhyananthan (2016) provided a set of discriminative features for the MEPCO voice

dataset's 50 speaker utterances. For the purpose of classifying speaker utterances, these writers retrieved RASTA-MFCC traits. To learn the classification rules, the collected characteristics were fed into a classifier built using a GMM-universal background model. 97% classification accuracy was attained by the findings. Although the results showed reasonable categorization accuracy, they cannot be generalized because the study only used six utterances, with one statement lasting barely three seconds. Because of this, RASTA-MFCC features may not be useful for speaker utterances that are longer than 3 s [4].

Laptik and Prasad (2017) examined MFFC features and a GMM classifier to categorize

50 and 138 speakers from the CMU and YOHO datasets. The experiment's outcomes showed 86% and 88% classification accuracy using the suggested feature extraction techniques [7].

Chapter III: Research Methodology

3.1 Types of Machine Learning

Machine learning is a branch of artificial intelligence and computing that focuses on using data and algorithms to mimic the way human learns, gradually improving accuracy. Based on the collected data, the machine improved the computer programs according to the required performance. Because of this ability of a machine to learn by itself, there is no need for explicit programming of these computers. Virtually every machine we use, and the high - tech machines we've seen over the past decade, incorporate machine learning to improve product quality. Based on the methods of learning, machine learning is mainly divided into four types, namely:

- Supervised Machine Learning
- Unsupervised Machine Learning
- Semi-Supervised Machine Learning
- Reinforcement Learning

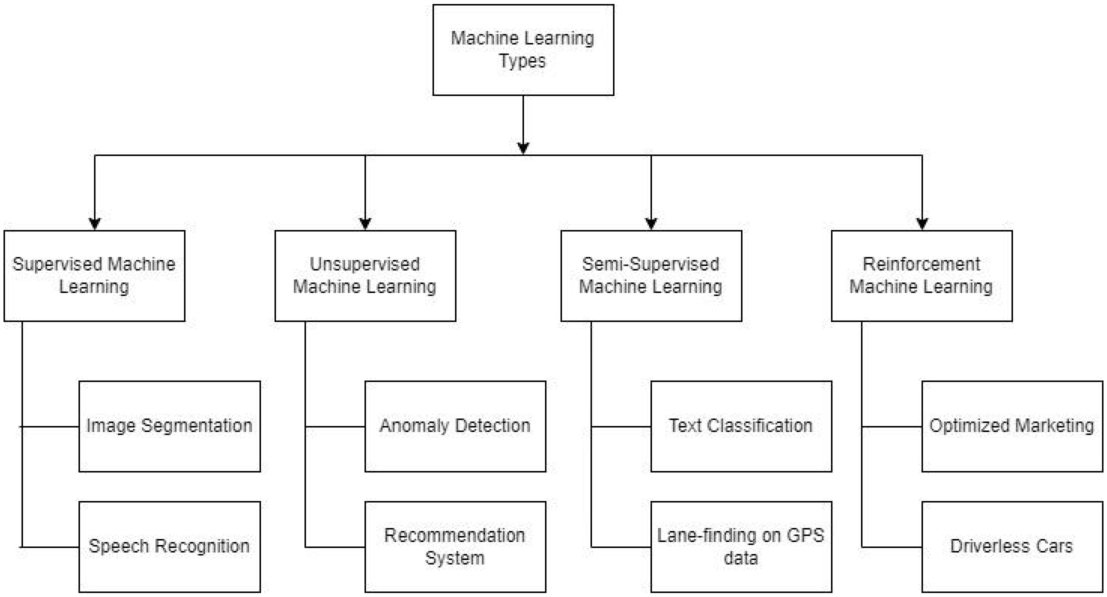


Figure 2: Types of Machine Learning

3.1.1 Supervised Learning

It is an algorithm used to form labeled facts and predict the consequences of unlabeled

facts. It uses well-labeled data to train the system. It indicates that some data's have already been tagged with the appropriate answers. It's like being with the help of teachers or supervisor while studying. Effectively creating, developing, and implementing appropriate supervised machine learning models requires time and hands-on expertise from highly skilled data analysts. Data analysts also need to update algorithm to ensure that the discoveries they make remain accurate even when the data is updated. Records training is used in supervised learning to achieve the desired results. When the inputs and outputs of these datasets are accurate, it helps this model learn faster. Example input-output pairs were used to monitor the machine learning process to learn an operation that translates an input into an output. The dataset on which we train our model is labeled in Supervised Learning. There is a clear and recognizable input-output mapping. The model can be trained in instances based on sample input. The spam filter is an example of supervised learning.

3.1.2 Unsupervised Learning

It is a type of machine learning where the algorithm is not monitored by humans.

Rather, it allows the model to work independently to uncover previously unnoticed data and patterns. This is usually data that is not labeled. Unlike the algorithms described earlier, these algorithms allow users to perform more sophisticated activities. These, on the other hand, provide more unexpected results than other natural learning methods. Neural networks, clustering, anomaly detection, and other unsupervised learning approaches are some examples. This methods does not used labeled data. The program recognizes and matches within the dataset. Based on intensity, the algorithm divided the data into several groups. It can be used to represent data that has many dimensions.

3.1.3 Semi-Supervised Learning

Semi-Supervised learning is a type of machine learning algorithm that falls between

supervised and unsupervised machine learning. It represents the intermediate step between supervised and unsupervised learning algorithms and uses the combination of labeled and unlabeled data sets during the learning. Although semi-supervised learning is halfway between supervised and unsupervised learning and works with data consisting of few labels, it mostly consists of unlabeled data. Because the tags are expensive but for business purposes, they may have few tags. It is completely different from supervise and unsupervised learning as they are based on the presence and absence of labels.

3.1.4 Reinforcement Machine Learning

This is a form of machine learning that depends on how systems should act in a given situation. It deep learning technique to maximize the result based on continuous feedback. This neural network learning method helps in learning method helps in learning achieve composite results and also takes advantages of the long time period to maximize the aspects provided. Reinforcement learning, semi-supervised learning, unsupervised learning, supervised learning are four main concepts or types of machine learning. It is the most common and widely used form of machine learning algorithm. It is used in a variety of autonomous systems, including automobiles and industrial robotics. The purpose of this algorithm is to achieve this objective in a changing environment. Anyone can achieve this goal through a variety of rewards provided by the system.

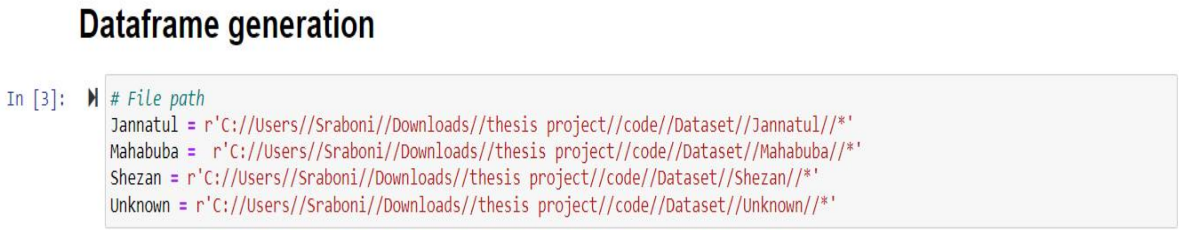
3.2 Dataset

We created a dataset of 700 audio clips, 3 known classes of 3 speakers, each with 150–

200 audio clips, and 1 unknown class of 30 speakers, each with five voice samples. The dataset contains audio files of both male and female speakers using a variety of accents and languages; however, the majority are speaking English. Each audio clips in the dataset are 10 seconds long. Moreover, 80% of the data was used for training, while 20% was used for testing. The classes were taken into the DataFrame and labelled to identify the speaker through MFCC and CNN architecture.

3.3 DataFrame

A 2-dimensional labeled data structure like a table with rows and columns is what the Pandas DataFrame is. The dataframe's size and values are mutable, or changeable. It is the panda thing that is used the most. In this work, a dataframe made from the dataset classes is shown as follows:

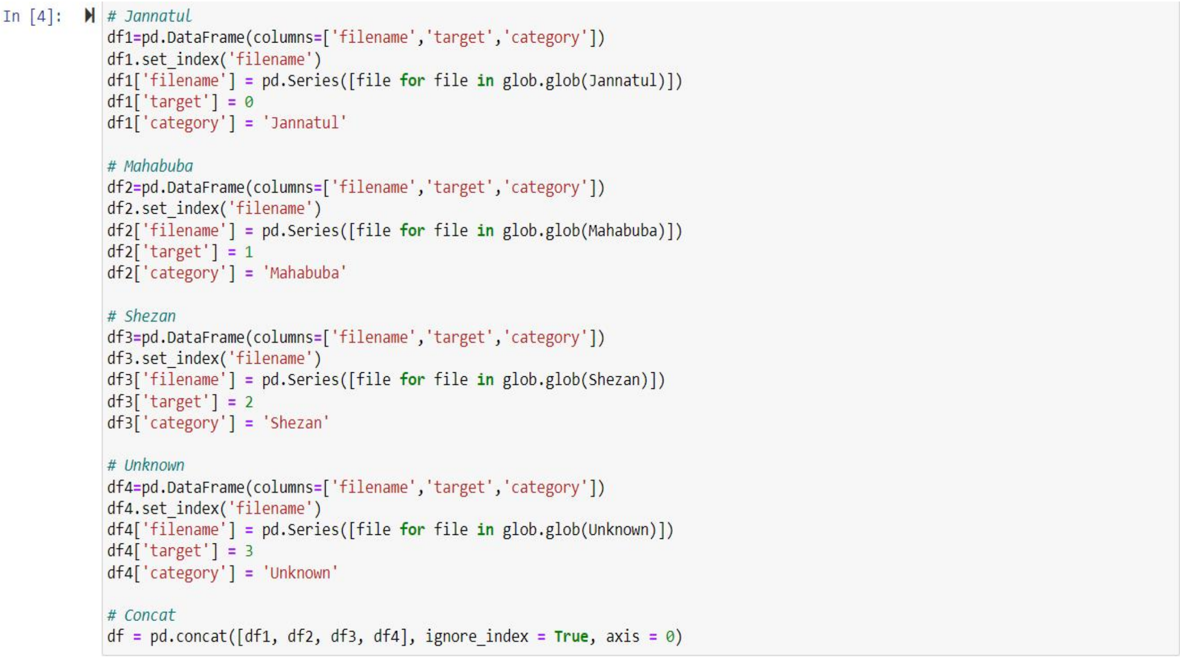


Dataframe generation

In [3]:

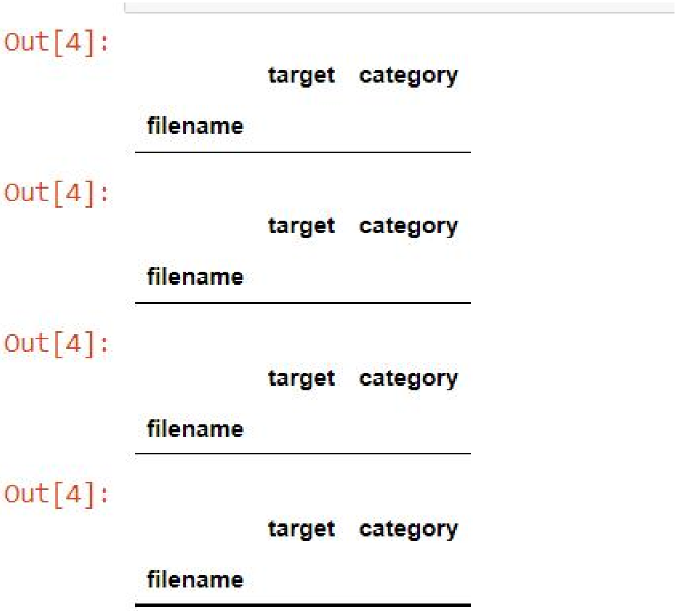
```
# File path
Jannatul = r'C://Users//Sraboni//Downloads//thesis project//code//Dataset//Jannatul//*'
Mahabuba =  r'C://Users//Sraboni//Downloads//thesis project//code//Dataset//Mahabuba//*'
Shezan = r'C://Users//Sraboni//Downloads//thesis project//code//Dataset//Shezan//*'
Unknown = r'C://Users//Sraboni//Downloads//thesis project//code//Dataset//Unknown//*'
```

Figure 3. Dataframe Generation (file path)



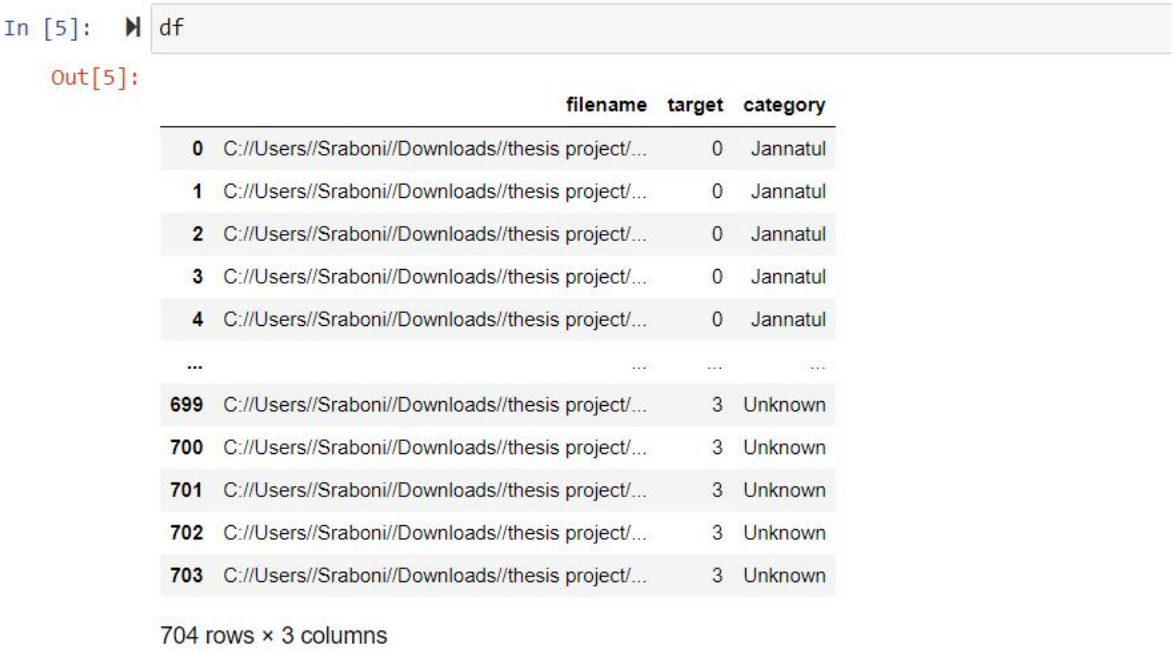
```
In [4]:  # Jannatul
         df1=pd.DataFrame(columns=['filename','target','category'])
         df1.set_index('filename')
         df1['filename'] = pd.Series([file for file in glob.glob(Jannatul)])
         df1['target'] = 0
         df1['category'] = 'Jannatul'

         # Mahabuba
         df2=pd.DataFrame(columns=['filename','target','category'])
         df2.set_index('filename')
         df2['filename'] = pd.Series([file for file in glob.glob(Mahabuba)])
         df2['target'] = 1
         df2['category'] = 'Mahabuba'

         # Shezan
         df3=pd.DataFrame(columns=['filename','target','category'])
         df3.set_index('filename')
         df3['filename'] = pd.Series([file for file in glob.glob(Shezan)])
         df3['target'] = 2
         df3['category'] = 'Shezan'

         # Unknown
         df4=pd.DataFrame(columns=['filename','target','category'])
         df4.set_index('filename')
         df4['filename'] = pd.Series([file for file in glob.glob(Unknown)])
         df4['target'] = 3
         df4['category'] = 'Unknown'

         # Concat
         df = pd.concat([df1, df2, df3, df4], ignore_index = True, axis = 0)
```

Figure 4. Dataframe Generation



```
Out[4]:
                 target   category
         filename

Out[4]:
                 target   category
         filename

Out[4]:
                 target   category
         filename

Out[4]:
                 target   category
         filename
```

Figure 5. Dataframe Output (1)



```
In [5]:  df
Out[5]:
                                             filename  target  category
0    C://Users//Sraboni//Downloads//thesis project/...       0  Jannatul
1    C://Users//Sraboni//Downloads//thesis project/...       0  Jannatul
2    C://Users//Sraboni//Downloads//thesis project/...       0  Jannatul
3    C://Users//Sraboni//Downloads//thesis project/...       0  Jannatul
4    C://Users//Sraboni//Downloads//thesis project/...       0  Jannatul
...                                                ...     ...       ...
699  C://Users//Sraboni//Downloads//thesis project/...       3   Unknown
700  C://Users//Sraboni//Downloads//thesis project/...       3   Unknown
701  C://Users//Sraboni//Downloads//thesis project/...       3   Unknown
702  C://Users//Sraboni//Downloads//thesis project/...       3   Unknown
703  C://Users//Sraboni//Downloads//thesis project/...       3   Unknown

704 rows × 3 columns
```

Figure 6. Dataframe Output (2)

3.4 Mel Frequency Cepstral Coefficients (MFCC)

The processes of MFCCs based feature extracted method are shown in Fig.1. The procedure begins by preprocessing and normalizing the voice or acoustic signal as necessary. Following that, the FFT process is followed by a temporal analysis that makes use of windowing and framing operations. The Mel-scale filter bank is then applied and wrapped around the logarithmic scale. Then, in order to calculate the Cepstral Mean Subtraction, the Discrete Cosine Transform (DCT) is applied (CMS). These CMS coefficients represent the characteristics obtained using the MFCCs approach. To progress the voice frame via the filter bank, the MFCCs-based feature extraction approach performed frequency analysis based on International Journal of Machine Learning and Computing. The method's output takes the form of MFCCs coefficients, which represent the retrieved features. As a result, these MFCCs coefficients can be used further for analysis or classification for any objective.

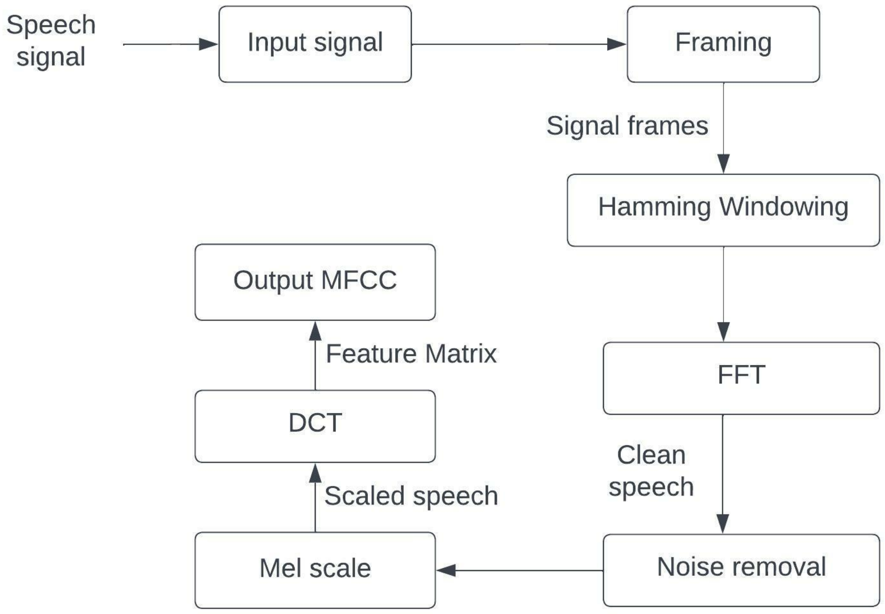


Figure 7. Mel Frequency Cepstral Coefficients(workflow)

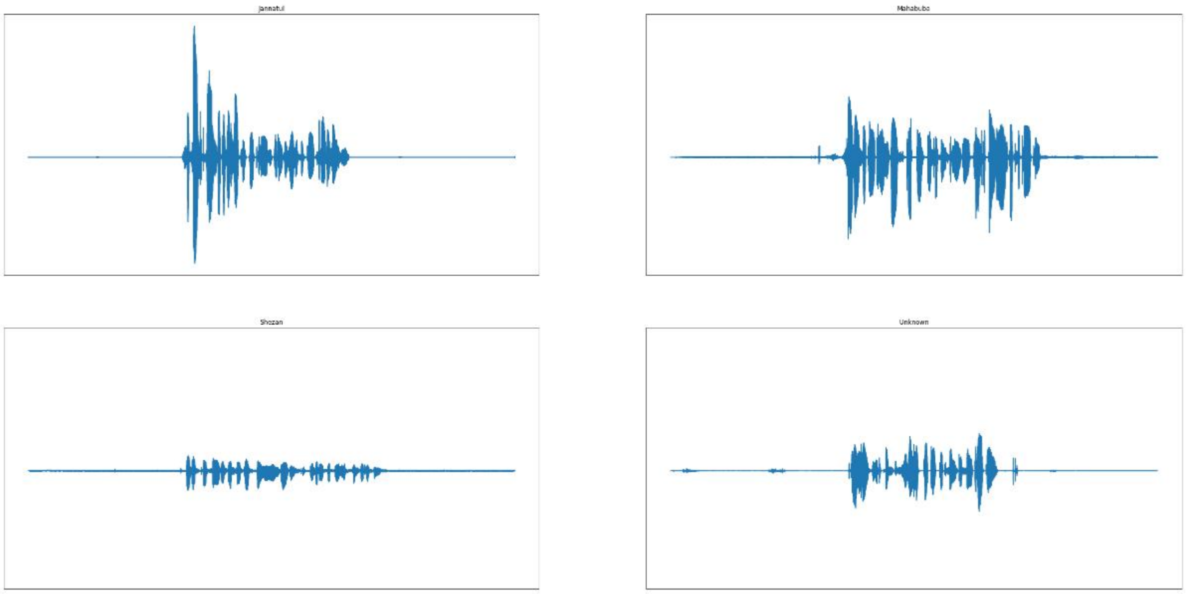


Figure 8. Mel Frequency Cepstral Coefficients

3.4.1 Noise Reduction

With the help of the Python noise reduction method noise reduce, time-domain signals

like voice, bioacoustics, and physiological signals can now be made quieter. It uses a technique known as "spectral gating," a type of Noise Gate. It computes a signal's spectrogram (and, if desired, a noise signal as well), and then estimates a noise threshold (or gate) for each frequency band of the signal and noise. In order to gate noise below the frequency-varying threshold, a mask is computed using that threshold.

3.5 Convolutional Neural Network

Convolutional Neural Network (ConvNet/CNN) is a Deep Learning technique that takes a picture as input and assigns significance (Weights and biases that can be taught) to a number of aspects/objects in the image while also differentiating them. ConvNet needs a lot of money. This classification technique requires fewer pre-processing steps than earlier classification techniques. ConvNets can train themselves to learn these filters and properties, whereas manual engineering is required for primitive approaches. The structure of a ConvNet, a connection network analog, influenced by the arrangement of the visual cortex, was described in the journal Neurons in the Human Brain. Only the Receptive Field, a small portion of the visual field, is responsive to inputs by individual neurons.

3.5.1 Convolutional Neural Network (CNN) architecture

The two components of a CNN architecture are:

- Feature extraction is a procedure that uses a convolution tool to separate and identify the distinct characteristics of a data for analysis.

- There are numerous pairs of convolutional or pooling layers in the feature extraction network.
- A fully connected layer that makes use of the convolutional process's output and determines the class of the data using the features that were previously extracted.
- This CNN feature extraction model seeks to minimize the quantity of features in a dataset. It develops new features that compile an initial collection of features' existing features into a single new feature.

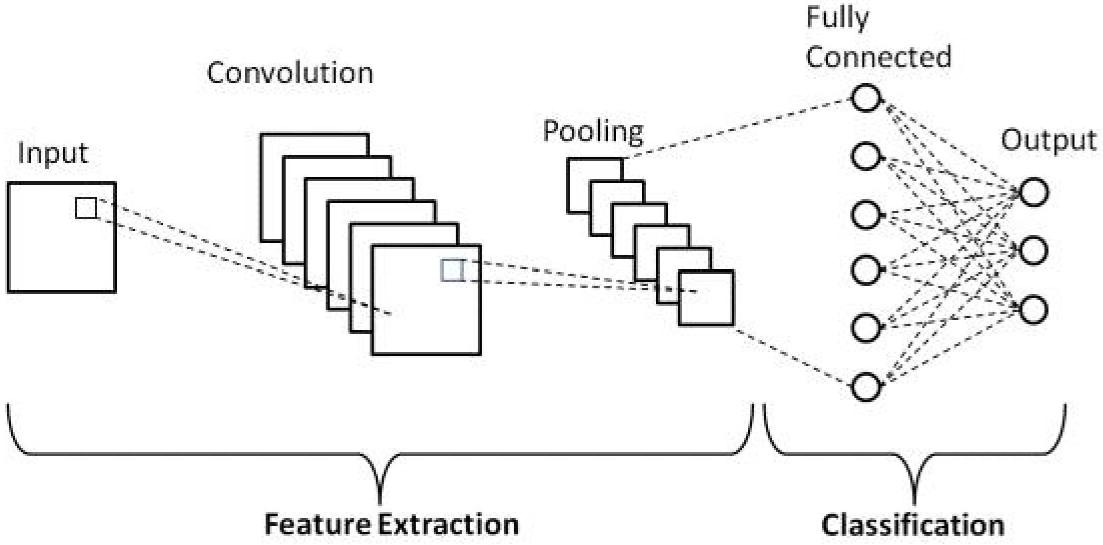


Figure 9. CNN Architecture

The convolutional layer determines the yield of neurons related to nearby input regions

by analyzing the weights scalar product position of each input volume. Convolutional neural networks work like a sigmoid to apply the activation from the preceding layer to the straight unit points that have been adjusted. Convolution is a specific kind of easy operation applied throughout the extraction process. Tensors, which are collections of numbers, are the input.

A portion is a small group of connected numbers that are small in size and have low spatial dimensions. The kernels consequently distributed equally across the input density. Once the data enters a convolutional layer, each channel is convolved over the spatial dimensions of the input.

Currently, a wide variety of computer vision techniques, other just CNN, use kernel

convolution. Using this technique, we can alter our image based on the values of a small network of numbers that we project onto it. The following equation, where f stands for the input image and h for our bit, is used to identify the pattern's elements. m and n, respectively, are used to validate the consequence network's line and column lists. The following formula demonstrates this:

$G[m, n] = (f * h)[m, n] = \Sigma \Sigma h[j, k]f [m - j, n - k]$

We locate our network over a particular pixel and then repeat each evaluation using bit

pairs of values taken from the images. In the end, we totaled everything up and placed the result in the appropriate place within the result, having the total.

3.5.2 Pooling Layer

Pooling, which makes promoting layers easier, is best described as down-sampling. It

is analogous to lowering the determination in the area of image processing. It's not very typical to see a Pooling layer combined with the intermediate progressive Conv layers in a ConvNet design. By minimizing the entity's borders and computations, it should be possible to manage over-fitting. Each spatially sliced depth cut of the data is resized by the Pooling

Layer utilizing the maximal operation, acting separately on each profundity cut. The biggest part of Max Pooling is made using the feature map. Using average pooling, the components in a section of an image with a predetermined size are averaged. The total sum of all the elements in a given segment is determined via sum pooling. Linking the Convolutional and Fully Connected layers is frequently done using the Pooling Layer.

3.5.3 Dropout

Normally, overfitting in the training dataset might result from all features being

connected to the FC layer. When a given model performs so well on training data that it has a negative effect on the model's performance when applied to new data, this is known as overfitting. To solve this issue, a dropout layer is used, in which a small number of neurons are removed from the neural network during training, reducing the size of the model. A machine learning model performs better thanks to dropout since it reduces overfitting by simplifying the network. During training, neurons are removed from the neural networks.

3.5.4 Activation Function

The activation function is one of the most crucial elements of the CNN model. They

are used to discover and approximation any type of continuous and complex link between network variables. In layman's terms, it determines which model information should shoot ahead and which should not at the network's end. The network gains nonlinearity as a result. The ReLU, Softmax, tanH, and Sigmoid functions are a few examples of commonly used activation functions.

- Rectified Linear Unit (ReLU): A ReLU is a piecewise linear function that yields a straight output from an input sequence. It is available in several sizes (0, 0). The most popular ReLU is an activation function in the deep learning era, which is a brief overview shown in the formula below:

y = max (0, x) (2.4)

The main advantage of using ReLU is that it solves the fading gradient issue; TanH, on

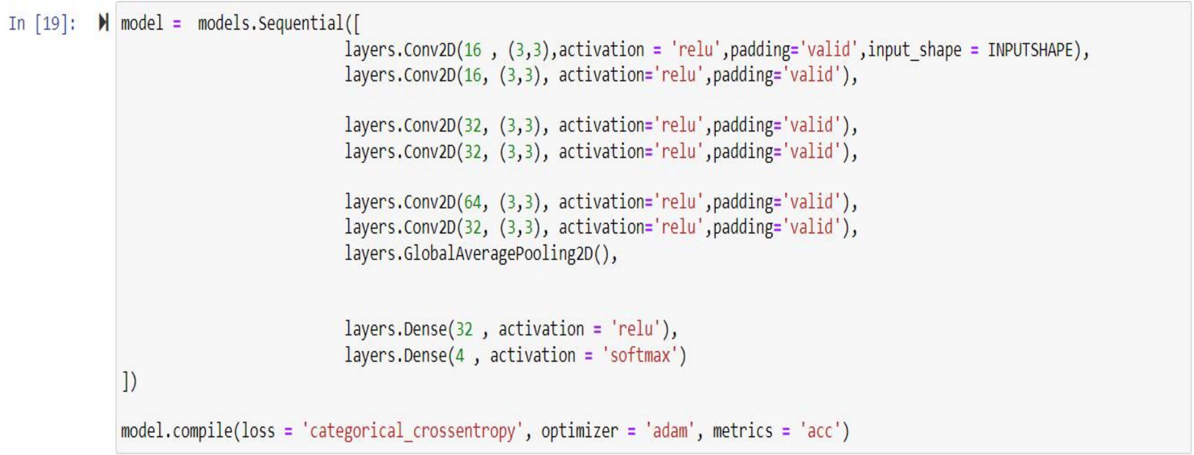
the other hand, has one-sided activation (50%) but sparse activation (0%) propagation error. Both this function and its derivatives are monotonic. The fact that this function is not zero-centered and cannot be differentiated by zero is a drawback. The dying ReLU problem, which happens when half of the results for non-zero-cantered action are inactive, is another drawback (returned as 0).

- Softmax: Using the Softmax function, you can transform a vector of K real values into a vector of 1 from a vector of K real values. The Softmax transforms the input data into probabilistic-sounding values between 0 and 1. It will always fall between 0 and 1, regardless of whether one of the inputs is tiny or negative, or huge. The Softmax translates small or negative inputs into small probabilities, and large or positive inputs into large probabilities. Multi-class logistic regression and Softmax are other names for the Softmax function. This is because the Softmax, a logistic regression extension for multi-class classification, has an equation that is extremely similar to the sigmoid used in logistic regression. The Softmax function can only be used in classifiers when the classes are mutually exclusive.

3.5.5 Final Sequential Model

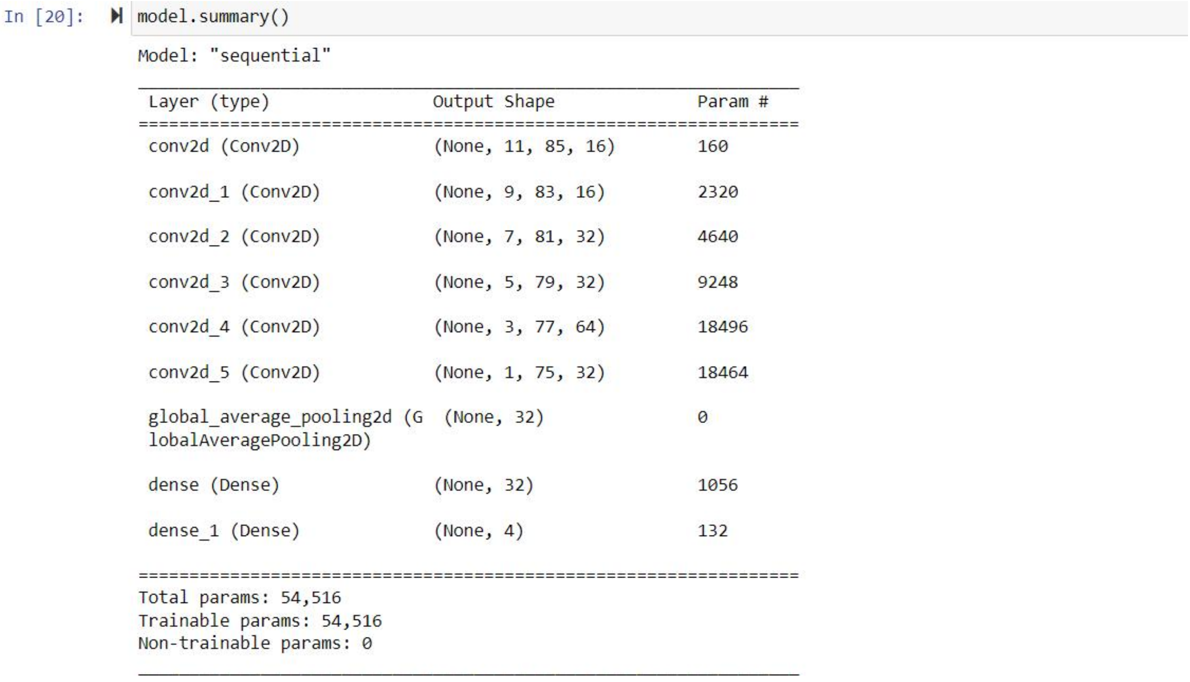
Finally, our model satisfies all of the conditions of the CNN (Convolutional Neural

Network). As can be seen below:



```
In [19]:  model =  models.Sequential([
                            layers.Conv2D(16 , (3,3),activation = 'relu',padding='valid',input_shape = INPUTSHAPE),
                            layers.Conv2D(16, (3,3), activation='relu',padding='valid'),

                            layers.Conv2D(32, (3,3), activation='relu',padding='valid'),
                            layers.Conv2D(32, (3,3), activation='relu',padding='valid'),

                            layers.Conv2D(64, (3,3), activation='relu',padding='valid'),
                            layers.Conv2D(32, (3,3), activation='relu',padding='valid'),
                            layers.GlobalAveragePooling2D(),


                            layers.Dense(32 , activation = 'relu'),
                            layers.Dense(4 , activation = 'softmax')
          ])

          model.compile(loss = 'categorical_crossentropy', optimizer = 'adam', metrics = 'acc')
```

Figure 10. Sequential Model

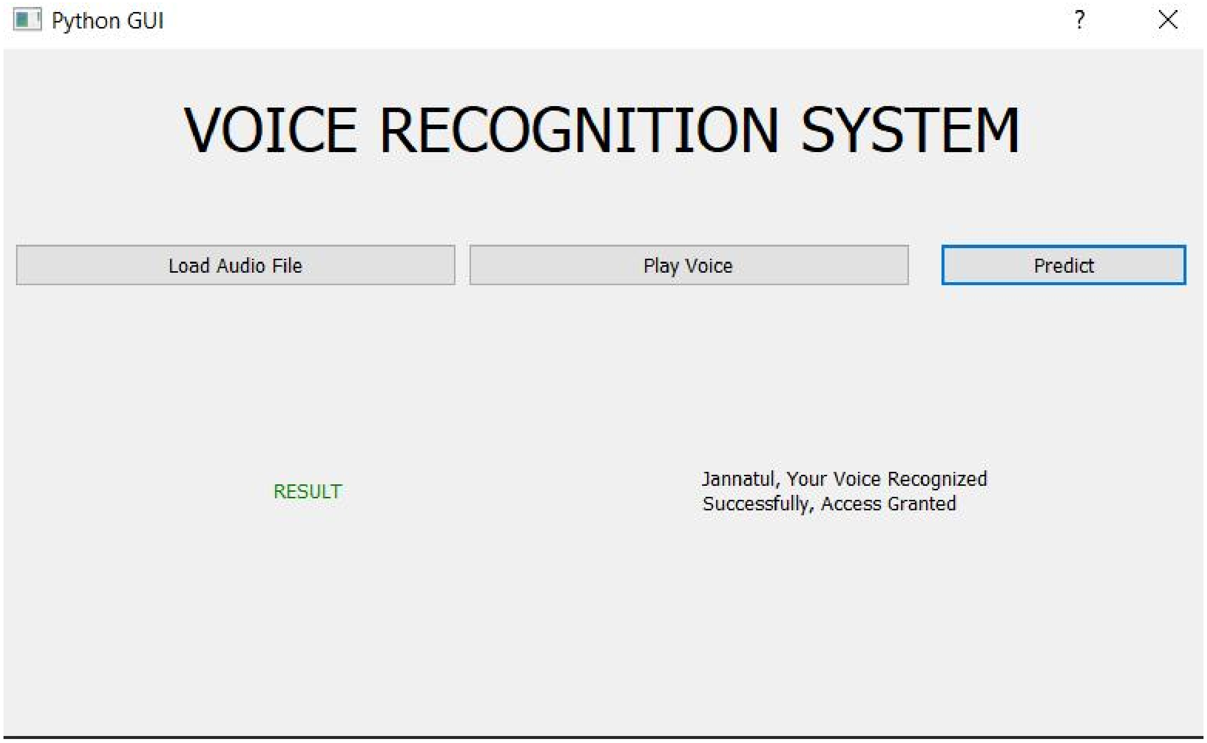


```
In [20]:  model.summary()

          Model: "sequential"
          _________________________________________________________________
           Layer (type)                Output Shape              Param #
          =================================================================
           conv2d (Conv2D)             (None, 11, 85, 16)        160

           conv2d_1 (Conv2D)           (None, 9, 83, 16)         2320

           conv2d_2 (Conv2D)           (None, 7, 81, 32)         4640

           conv2d_3 (Conv2D)           (None, 5, 79, 32)         9248

           conv2d_4 (Conv2D)           (None, 3, 77, 64)         18496

           conv2d_5 (Conv2D)           (None, 1, 75, 32)         18464

           global_average_pooling2d (G  (None, 32)               0
           lobalAveragePooling2D)

           dense (Dense)               (None, 32)                1056

           dense_1 (Dense)             (None, 4)                 132

          =================================================================
          Total params: 54,516
          Trainable params: 54,516
          Non-trainable params: 0
          _________________________________________________________________
```

Figure 11. Sequential Model Summary

3.6 Python Graphical User Interface

The GUI for this system was made using PyQt4, a complete set of Python bindings for Digia's Qt cross-platform GUI toolkit. The GUI contains a predict button and shows the speaker's name if the system successfully predicts the name. The GUI contains a predict button and if the system successfully predicts the speaker, the speaker's name will display on the screen along with the message "accept granted." otherwise, it will display "accept denied."



VOICE RECOGNITION SYSTEM

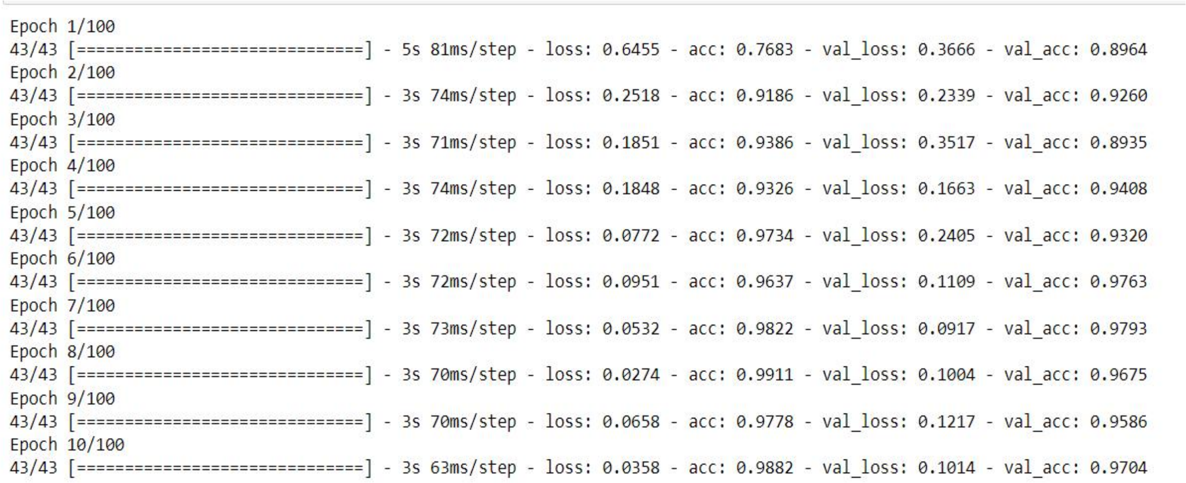
Jannatul, Your Voice Recognized
Successfully, Access Granted

Figure 12. Python Graphical User Interface

Chapter V: Result and Discussion

When evaluating the performance of a novel, we need to consider three qualities: test, accuracy and loss function, and epoch. It shows the accuracy rate of a given model when evaluating accuracy per epoch and shows the difference between our generated and accuracy results when the test loss is determined. We chose the highest possible epoch while using a large enough dataset to ensure that the model was properly trained on the given data to obtain the minimum achievable loss function in our experiment.

We achieved the highest accuracy through our model. Epochs and accuracy shown below:



```
Epoch 1/100
43/43 [==============================] - 5s 81ms/step - loss: 0.6455 - acc: 0.7683 - val_loss: 0.3666 - val_acc: 0.8964
Epoch 2/100
43/43 [==============================] - 3s 74ms/step - loss: 0.2518 - acc: 0.9186 - val_loss: 0.2339 - val_acc: 0.9260
Epoch 3/100
43/43 [==============================] - 3s 71ms/step - loss: 0.1851 - acc: 0.9386 - val_loss: 0.3517 - val_acc: 0.8935
Epoch 4/100
43/43 [==============================] - 3s 74ms/step - loss: 0.1848 - acc: 0.9326 - val_loss: 0.1663 - val_acc: 0.9408
Epoch 5/100
43/43 [==============================] - 3s 72ms/step - loss: 0.0772 - acc: 0.9734 - val_loss: 0.2405 - val_acc: 0.9320
Epoch 6/100
43/43 [==============================] - 3s 72ms/step - loss: 0.0951 - acc: 0.9637 - val_loss: 0.1109 - val_acc: 0.9763
Epoch 7/100
43/43 [==============================] - 3s 73ms/step - loss: 0.0532 - acc: 0.9822 - val_loss: 0.0917 - val_acc: 0.9793
Epoch 8/100
43/43 [==============================] - 3s 70ms/step - loss: 0.0274 - acc: 0.9911 - val_loss: 0.1004 - val_acc: 0.9675
Epoch 9/100
43/43 [==============================] - 3s 70ms/step - loss: 0.0658 - acc: 0.9778 - val_loss: 0.1217 - val_acc: 0.9586
Epoch 10/100
43/43 [==============================] - 3s 63ms/step - loss: 0.0358 - acc: 0.9882 - val_loss: 0.1014 - val_acc: 0.9704
```

Figure 13. Test Accuracy

So, we’ve got almost 98% of accuracy which is much better than other researches previously done on this topic. And about the detection result our model successfully recognized the voice that is being tested by the machine. As our research topic is based on text independent voice recognition system, so when the user speaks, it doesn’t need to speak only some specific phrase rather he/she can speak anything. As we have taken a wide range of dataset containing almost 700 audio clips. The audio clips contain independent audio content. Here the majority of the audio clip is English and both male and female voices are included in the dataset to train the model to predict. And our model successfully recognized the audio clips by recognizing their voice from the audio clip and gives a confirmation message that the voice has been successfully recognized. With that it also shows the permission to access the door if the voice is known by the machine it shows that access granted by the machine but if the voice is unknown, it denied the access of that user. Here we show the accuracy of our model by using a bar diagram:

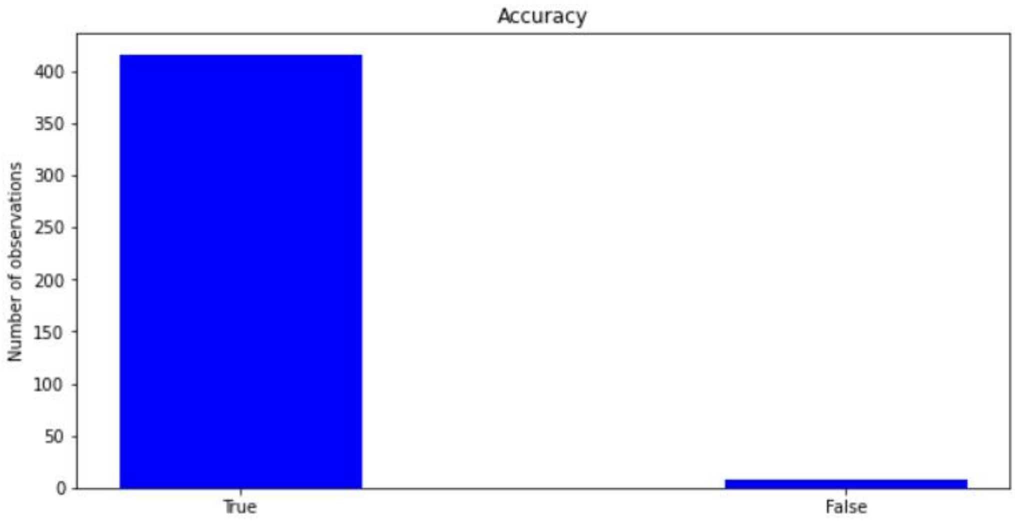
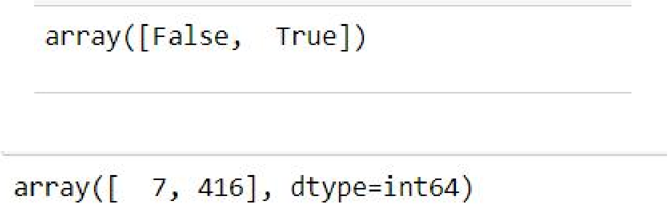


Figure 14. Accuracy in Bar Diagram

Here from the bar diagram, we can clearly see our accuracy. From the dataset it correctly (True) predicts the voice up to 416 records and mispredict (False) only 7 records.



```
array([False,  True])
```

```
array([  7, 416], dtype=int64)
```

Figure 15. Prediction Count

Here is also a comparison table for comparing our work at the end of the table and some of the works that we have studied we conducted from different research papers before it:

Table 1: Accuracy Comparison of different studies

Sl	Research Name	Accuracy
1	Text-Independent Speaker Identification Through Feature Fusion and Deep Neural Network	83.5%–93%
2	An MFCC-based Text-Independent Speaker Identification System for Access Control	94%-95%
3	Speaker Verification Using Machine Learning for Door Access Control Systems	97%
4	Voice Recognition Door Access Control System	96%
5	Speaker recognition with hybrid features from a deep belief network	92%
6	Classifying human voices by using hybrid SFX time-series preprocessing and ensemble feature selection	94%
7	Text-independent speaker identification based on selection of	93%

	the most similar feature vectors	
8	Noise robust speaker identification using RASTA–MFCC feature with quadrilateral filter bank structure	97%
9	Fast binary features for speaker recognition in embedded systems	86%
10	Frame selection for robust speaker identification: A hybrid approach	88%

Chapter VI: Conclusion

The proposed methodology provides a significant study of text independent voice recognition system for opening a door system with voice recognition. We have developed a system that can recognize a user voice. We have prepared a dataset including four classes having almost 700 audio clips. When a user wants to access the door, he/she gives the input to the local machine as soon as the machine gets the voice by using the CNN architecture SoftMax will give a probability of four class. From that four class which one have the most probability among other class the architecture gives that class as the output or predict that the given voice is from that class.

The experimental results showed that the performance of the proposed MFCC functions

in terms of overall accuracy were approximately 97%-98%. Furthermore, DNN was found to be suitable for identifying speakers through the proposed MFCC functions. Experimental results demonstrate that the proposed voice recognition system is efficient, accurate and robust compared to the other identification model. Promising results show that the proposed voice recognition system can be used in many application areas, including access control and security. The problem we have face in figuring out how to get the unknown audio clips to the system to improve the result.

We have tried to make a model on the basis of independent voice input, so that the

model can be recognized the user independent voice audio clips, and the model successfully recognized the voice clips and gives a good prediction. Though we have tried our best to get the best result from the proposed system but for time limitation and limited number of data we didn’t make it to 99%-100% and for this limited data our model is unable to predict the known voice in real time. In the future, our desire is to improve categorical accuracy by reducing categorical errors/loss by using deep learning with deeper architectures. We believe, with a large dataset we can do the speaker recognition in real-time having a much better accuracy and much better interface than now.

References

[1] Fong S., Lan, K. and Wong, R. (2013). *Classifying human voices by using hybrid SFX time- series preprocessing and ensemble feature selection*, BioMed Res. Int

[2] Ali, H., Tran, S. N., Benetos, E. and Garcez, A. S. D. A. (2018). *Speaker recognition with hybrid features from a deep belief network*, vol. 29, pp. 13–19, Neural Comput. Appl

[3] Soleymanpour, M., and Marvi, H. (2017). *Text-independent speaker identification based on selection of the most similar feature vectors*, vol. 20, no. 1, pp. 99–108, Int. J. Speech Technol.

[4] Nidhyananthan, S. S., Kumari, R. S. S. and Selvi, T. S. (2016). *Noise robust speaker identification using RASTA–MFCC feature with quadrilateral filter bank structure*, vol. 91, no. 3, pp. 1321–1333, Wireless Pers. Commun.

[5] Vacher, M., Lecouteux, B., Serrano-Romero, J., Ajili, M., Portet, F., Rossato, S. (2015). *Speech and speaker recognition for home automation: preliminary results*, 8th International Conference Speech Technology and Human-Computer Dialogue.

[6] Prasad, S. Z., Tan, H. and Prasad, R. (2017). *Frame selection for robust speaker identification: A hybrid approach*, vol. 97, no. 1, pp. 933–950, Wireless Pers. Commun.

[7] Laptik, R. and Sledevič, T. (2017*). Fast binary features for speaker recognition in embedded systems*, pp. 1–4, in Proc. Open Conf. Elect., Electron. Inf. Sci., New York, NY, USA.

[8] Jahangir, R., Teh, Y. W., and Memon, N. A. (2020). *Text-Independent Speaker Identification Through Feature Fusion and Deep Neural Network,* IEEE Access, Malaysia.

[9] Olubukola, A., Adeoluwa, A., and Abraham, O. (2019). *Voice Recognition Door Access Control System,* vol. 21, pp. 01-12, IOSR Journal of Computer Engineering (IOSR-JCE).

[10] Alaliyat, S., Dyvik, K., and Oucheikh, R. (2021). *Speaker Verification Using Machine Learning for Door Access Control Systems,* ResearchGate.

[11] Liu, J. C., Leu, F. Y., Lin, G. L., and Susanto, H. (2017). *An MFCC-based text-independent speaker identification system for access control,* Wiley.

[12] Bunrit, S., Inkian, T., Kerdprasop, N., and Kerdprasop, K. (2019). *Text-Independent Speaker Identification Using Deep Learning Model of Convolution Neural Network,* Vol. 9, No. 2, International Journal of Machine Learning and Computing.