A NOBLE DEEP LEARNING APPROACH TO RECOGNIZE SPEAKER'S IDENTITY FROM BENGALI SPEECH

BY

Md. Fahad Hossain
ID: 172-15-9600

Hasmot Ali
ID: 172-15-9632
And
Md. Mehedi Hasan
ID: 172-15-9804

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Sheikh Abujar
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Co-Supervised By

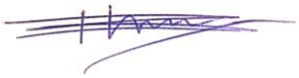
Dr. Sheak Rashed Haider Noori
Associate Professor and Associate Head
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
MAY 2021

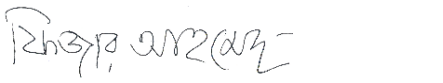
APPROVAL

This Research based Project titled **"A Noble Deep Learning Approach To Recognize Speaker's Identity From Bengali Speech**", submitted by Md. Fahad Hossain (172-15-9600), Hasmot Ali (172-15-9632) and Md. Mehedi Hasan (172-15-9804) and to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering (BSc) and approved as to its style and contents. The presentation has been held on 1th June 2021.

BOARD OF EXAMINERS



Dr. Touhid Bhuiyan — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Dr. Fizar Ahmed — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Azizul Hakim — **Internal Examiner**
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Mohammad Shorif Uddin — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

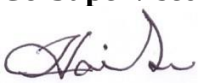
DECLARATION

We hereby declare that, this research project has been done by us under the supervision of **Sheikh Abujar, Lecturer (Senior Scale), Department of CSE,** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

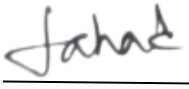
Sheikh Abujar
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Co-Supervised by:



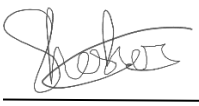
Dr. Sheak Rashed Haider Noori
Associate Professor and Associate Head
Department of CSE
Daffodil International University

Submitted by:



Md. Fahad Hossain
ID: 172-15-9600
Department of CSE
Daffodil International University

Hasmot Ali
ID: 172-15-9632
Department of CSE
Daffodil International University



Md. Mehedi Hasan
ID: 172-15-9804
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year research project successfully.

We really grateful and wish our profound our indebtedness to **Sheikh Abujar**, **Lecturer (Senior Scale)** and **Dr. Sheak Rashed Haider Noori**, **Associate Professor and Associate Head,** Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor and co-supervisor in the field of "*Natural Language Processing*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Professor Dr. Touhid Bhuiyan, Professor & Head, Department of CSE**, for his kind help to finish our research project and also to other faculty members and the staffs of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Speech is the most suitable form of communication. Speech-based applications are playing a vital role in modern technology for the last few decades. Because it has a lot of identical features for measuring performance and behavior of human voice. Speech-based application is not only the trend of modern and efficient technology but also a new shift of information and technology paradigm. Several research works have been completed on voice-based applications because it has more practical application than any other form of communication. In this work, we tried to recognize the feature of voice in term of identify the speakers from Bengali speech. We consider speakers Age, Division, Height, Weight, Gender, Occupation as the parameter to identify a speaker. But here we presenting the application of recognizing Bangladeshi speaker's age and division from Bengali Speech. We used our own dataset containing 16730 samples. Each sample is a wav format audio of 8-10 seconds duration. We consider MFCC, Delta, Delta-Delta, LSF, Spectral Bandwidth and mel spectrogram features to train our model. We tried some traditional Machine Learning algorithms early but we understand that the huge number of data does better with Deep Learning algorithms. We tried different Deep Learning algorithms such as Artificial Neural Network, Convolutional Neural Network, Region Based Convolutional Neural Network, Long Short-Term Memory with different types of features but ended with Artificial Neural Network with 85% accuracy for Division recognition and Convolutional Neural Network with 78% accuracy for Age recognition.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

ABBREVIATION	EXPLANATION
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
VoIP	Voice over Internet Protocol
NLP	Natural Language Processing
BNLP	Bengali Natural Language Processing
ASR	Automatic Speech Recognition
TTS	Text to Speech
STT	Speech to Text
MFCC	Mel-frequency Cepstrum
KNN	k-Nearest Neighbors
HCI	Human Computer Interaction
SVM	Support Vector Machine

CHAPTER 1
INTRODUCTION

1.1 Introduction

Speech is a powerful medium for efficient communication. Having a bunch of unique features like Qualities, Pitch, Tone, Rhythm, Resonance, Texture and others, speech is one of the most useful identifiers to recognize speakers. Each speech sound can be analyzed in terms of its phonetic features, the chunks of the sound that can each be autonomously controlled by the articulators. Having perfectly structured, multiple level regularized properties speech has only a few fragments of these properties become significant units, which is defined as phonemes [1-2]. So, by using speakers' speech and analyzing the features of the speech, voice-based applications can place a huge participation in speaker recognition term. In the era of Alexa, Siri, or Google Assistance, people are using voice command-based applications. So, voice also can be the detection mechanism of Age, Division, Gender, Occupation, Height, Weight or any other identical parameter of a speaker.

Age and Region detection is one of the important tasks of classifying people from a specific age range and geographical area. It has a lot of applications including detecting the speaker from an unknown age and region, verifying the unknown age and region of crime suspect and speaker recognition. As we know nowadays a lot of crime and harassment is being performed using telephone, cell phone, internet phone, voice fishing or Voice over Internet Protocol (VoIP). With the popularity of Internet phones, relevant criminal activities have emerged [3]. Some new kind of cybercrime using artificial intelligence and voice technology is one of the unfortunate developments of postmodernity. Nowadays people just cannot believe what they hear like they cannot just believe what they see after the introduction of deep fake video. A huge number of voice related harassment and crime cases are recorded every day in the perspective of Bangladesh.

Survey finds that about 165.57 million people are using mobile phones and about 100 million people in Bangladesh are using the internet. So, the possibility of such crime happening is high. Bangladesh has had such a negative impact of social network harassment and crime from the last few decades. So being prepared for avoiding such kinds of threats would lead to a civilized nation as well as a world. Control over voice-based communication can take part of this initiative as well.

Using their voice and digital voice-based communication devices, people are committing such crimes just because we don't have enough control over the voice-based communication system and have no effective system to look over them. Exactly here is our point of research to provide some practical contribution to determine the age and division. As we tried to recognize an individual speaker's age and division from Bangla speech. We know it is not completely possible to recognize a speaker's identity but we think it could be helpful for the overall process. It should mention that we consider three age labels like 8-18, 19-21, 22-26 for balancing the data and eight

divisions of Bangladesh named Barisal, Chittagong, Dhaka, Khulna, Mymensingh, Rajshahi, Rangpur and Sylhet as eight labels of Division recognition.

1.2 Objective

The research of Automatic Speech Recognition (ASR), Text to Speech (TTS) or Speech to Text (STT) are meant to recognize speakers and what they are trying to say. As well as our case we tried to recognize speakers' identical features such as the Age range or the Division they from, by processing the speech they deliver. The contribution can provide a fear medium of voice over internet communication systems. As we know ASR technology is not advanced enough in perspective of Bangladesh. The speech-based application and their implementation is quite not possible in Bangladesh. So, we hope we can contribute some knowledge and experience with the community of BNLP research. We also have the largest Bengali voice dataset available for every researcher over the country as well as Bengali researchers from every corner of the world.

Some of our key objective is to:

1. Contribute into Bengali Natural Language Processing (BNLP) research
2. Build one of the richest Bengali Speech Dataset
3. Recognize Speakers Age from Bengali Voice
4. Recognize Speakers Division from Bangla Voice
5. Identify speakers using Bangla voice

1.3 Motivation

There are a lot of applications of speech recognition technology like ASR, TTS, STT and other speech command-based processes in English as well as other languages. As far as we experience and know, there are no appropriate applications for working with Bangla Command and speech. So as children of a Bengali mother we obviously have some responsibility to give something to the Bengali community to present our beloved Bangla to the world. We know it is such a tiny contribution to the field of BNLP research but we hope it can be helpful in such a way.

Our motivation is to work with our roots, enrich the root and help the nation to go through the era of advanced science and technology. We just cannot forget the sacrifices of language martyrs from 21st February 1952. We believe improving technology is the only way of improving the generation and making a smooth way to drive them into creating a creative path to improve their next generation.

We see the researchers from another language making such a great contribution and invention for their own nation and community. So, it is also a great inspiration to learn from them and invest the learning for our own goods. We hope it would bring some good use of Bangla voice-based application over the Bangla language speaker.

1.4 Rationale of Study

It is already known to us that voice is one of the efficient forms of communication and it is also a hot topic in the research field. Many of the researchers are working with speech-based applications all over the world and we are experiencing the benefits. So, we think working with speech is obviously a worthy one. As this speech technology is not rich for Bangla language, so we think it is such a great initiative to work with Bangla voice. If we consider the fundamental principle of recognizing speakers' identity like age range and division it would be the first one in Bangladesh. As we discussed a lot of problems previously, we think this is the only consistent way of solving this. We also discuss about different technical possibilities like comparison between different machine learning and deep learning algorithms to examine this initiative. So, our overall observation about the topic and the process is quite worthy indeed.

1.5 Research Questions

If we want to know the contribution, we have to know the problem first. There is a saying like, before driving to the solution or approach, making the problem clear should be the ultimate goal of a researcher. As we intended to provide a method for recognizing speakers' identity from their speech, we considered some questions and over the process we tried to solve the question. Here are some major research questions we consider:

a. Is it possible to recognize speakers age limit from speech?
b. Is it possible to recognize speaker's division limit from speech?
c. How speech/voice feature work?
d. Which feature should we consider to increase the performance?
e. Which algorithm perform best for our data?
f. What about the practical implementation of this method?

1.6 Expected Output

As we have worked on Bangladeshi speaker's recognition using their speech data. So, we expect to recognize speakers Age range and the division they belong to so that we can recognize the person for any reason. Here are some core outputs we expect from the overall method:

a. Providing biggest Bangla speech dataset
b. Recognizing speakers Age range using their voice
c. Recognizing speakers Division from their voice
d. Helping other researchers to find best audio feature for voice recognition
e. Helping others researchers to choose best algorithm for speaker recognition

1.7 Report Layout

In **Chapter 1**, this report discusses about what we are going to do, why we are going to do and how we are going to do. In this section we point out about the impotency of speech recognition and their application, making clear about our motive and different question we consider about the process we are going to present. Our overall, the motivation behind this work with expected outcome is described briefly in this chapter.

In **Chapter 2**, related work of this sector has been described. And summarizing their work findings from their works also noted in this section. By finding their limitation we set our goals by explain the challenges.

In **Chapter 3**, this report discusses about methodology has been used in this work. Some theoretical topics are also discussed in this chapter which are related to this work. Process of data collection, data preprocessing, feature extraction, methodologies used in this work are briefly discussed in this chapter.

In **Chapter 4**, the result came from previous chapter have been presented and comparison and best process also showed in this chapter. We also present some comparison between different algorithm and showed their result for further clarification.

In **Chapter 5**, summery of the project is main focus. Future work, conclusion, limitation and recommendation also noted in this last chapter of the report. We also talked about Acknowledgement and Declaration of Competing Interest to clarify the gratitude some people and organization deserve.

CHAPTER 2
BACKGROUND

2.1 Introduction

Speech recognition is not a new topic of research and application as we know. So, there are a lot of research has been done in this field. Most of them are for English and other languages. As there are a lot of applications and possible applications of speech recognition researchers are driving through the topic day by day. If we consider some applications of speech recognition, we can see that some of the researchers are working on Automated speech recognition like automated voice command-based devices or robots, some of them are driving through the algorithm or performance improvement for different speech recognition applications. Some of them work on speech features and these feature-based classifications and recognition problems like gender, age, region, height weight estimation [4-7].

2.2 Related Work

As we intended to implement a speaker's age and division recognition system using Bangla Speech. We drive through a lot of related work available. Specially we are focusing on Speech Recognition application for Bangla Language and Age and Division recognition. As we tried this method with different Machine Learning and Deep Learning algorithms, we also reviewed some work to demonstrate how different Deep Learning algorithms perform for Speech Recognition. We also found some work related to speech preprocessing and feature extraction. Here are some most related works we study to demonstrate our work and compare with:

2.2.1 Speech Recognition: The research, development and the accuracy of automatic speech recognition (ASR) remains one of the most important research challenges over the years e.g. speaker and language variability, vocabulary size and domain, noise. There is a lot of research done for recognizing speakers from online meetings, television conversations, and live speech. Vinyals et al [8] performs a speaker recognition approach from an online meeting by using a single far-field microphone. A self-learning speech-controlled system for speaker identification and adaptation in terms of detecting unknown speakers is performed by Herbig et al [9-10] using Unsupervised Speech Controlled System. Amino et al [11] describe different factors that affect human speaker recognition application. They perform two different experiments to identify those effects. A speaker change detection in broadcast television is performed by Yin et al [12] where Bi-LSTM were used and said that Bi-LSTM method shows an improved result than conventional method. There is also a lot of speech-based speakers' gender, height, weight, age detection approach done by different researchers [13-16].

There are also a lot of preprocessing, feature extraction and selection tasks done for ASR. Some of the feature selection research is by Sharma et al [17], who have done a noble review of speech features for several machine learning applications. They have covered features of different domains such as temporal domain, frequency domain, cepstral domain, wavelet domain and time-frequency domain features. These domains include description of all the features within corresponding domains. Davis et al [18] compared several parametric representations of the acoustic signal for continuous speech recognition systems. They introduced a set of ten mel-frequency cepstrum coefficients known as MFCCs in their paper. MFCCs performed best with 96.5% and 95% accuracy for two speakers in the experiment. Furui et al [19] have experimented with speaker recognition tasks using statistical features and dynamic features. Their experiment shows that the accuracy difference is very low between the recognition for statistical features and dynamic features. They have also observed that the error rate can be reduced to half if we use a combination of dynamic features and statistical features.

We also study some review work in this field like Saksamudre et al [20] presents the basic idea of speech recognition, proposed types of speech recognition, issues in speech recognition, different useful approaches for feature extraction of the speech signal with its advantage and disadvantage and various pattern matching approaches for recognizing the speech of the different speaker. Haton et al [21] describes the recent progress and the author's perspective of ASR and gives an overview of major technological perspectives and appreciation of the fundamental progress of Automatic speech recognition. The last paper we read about GGGG by Benk et al [22], they give an overview of the main definitions of ASR which is an important domain of artificial intelligence and which should be considered during any related research (Type of speech, vocabulary size... etc.).

2.2.2 Speech Recognition in Bangla Language: Almost 200 million people worldwide, 160 million of whom are Bangladeshi speaking Bangla as their first language [23]. But research of ASR Bangla voice-based applications is rare. There is some research of Bangla Language based on voice and we review some of them for our contribution. We should mention some of the respective authors who have presented some great results of Bangla voice-based research to gear up the Bengali researchers. Rahman et al [24] discussed how Bangla speech recognition processes works and the steps of the ASR system. They mentioned 3 main steps of ASR and for each step, they mentioned many algorithms. LPC, PLP, MFCC, RASTA-PLP algorithm is used for extracting a feature from Bangla voice data. Ahammad et al [25] recognize Bangla digits using MFCC features of the segmented words (Bangla Digit) and these feature values are sent as the input to the Back-Propagation Neural Network (BPNN). BPNN model gives 98.46% accuracy. Paul et al [26] also work with speech of Bangla Digit and present a method which calculates the Linear Predictive Coding (LPC) and Cepstral Coefficients of speech and create an ANN model for recognized speech. It has a high accuracy of recognizing individual

digits. On the other hand, Hasnat et al [27] talked about isolated words and continuous speech recognition. They applied HMM for the pattern classification and incorporated a stochastic language model in this system and showed that the overall performance decreases by 20% when a different speaker performs the word. Finally, in a review paper presented by Badhon et al [28], they reviewed recent Bangla speech recognition papers and find some value of different attribute like Largest Bangla Datasets, Best accuracy performed in Bangla Speech research, best effective accuracy, most used algorithm, latest technique, tools and best feature extraction technique.

2.2.3 Deep Learning for Speech Recognition: As speech provides a continuous and sequential form of data it is quite easy to work with a sequential model. Many Machine Learning models and algorithms are used for speech recognition but somehow people ended up with the Deep Learning model when working with Speech Recognition and its huge amount of data. There is a lot of research that is already done for speech recognition using deep learning. Here we review some of them as deep learning is ruling this field. Gupta et al [29] propose two approaches for speech recognition via supervised and unsupervised learning. For supervised learning Bi-directional Recurrent Neural Network with Long Short-Term Memory model (LSTM), so that speech signal reconstruction can be done in a proper way without performance loss. For unsupervised learning, the model is designed on the basis of the Restricted Boltzmann Machine (RBM) which generates a reconstruction-based output and helps in conversion of voice into text, each letter by letter. A Raw Speech Recognition is done by Passricha et al [30] using CNN where they discuss Three major types of end-to-end architectures for ASR are attention- based methods, connectionist temporal classification, and convolutional neural network (CNN)-based direct raw speech model. They conclude that this system uses only few parameters and performs better than traditional cepstral feature-based systems.

Abdel-Hamid et al [31] showed that further error rate reduction can be obtained by using convolutional neural networks (CNNs). Experimental results show that CNNs reduce the error rate by 6%-10% compared with Deep Neural Networks. A review paper is done by Halageri et al [32] who shows the pattern matching abilities of neural networks on speech signals. They discuss different available speech recognition techniques, their new proposed technique and combine their thinking with the trend of working with automated speech. Hennebert et al [33] review some of the Artificial Neural Network (ANN) approaches used in speech recognition. Some basic principles of neural networks are briefly described as well as their current applications and performances in speech recognition. A systematic review is done by Nassif et al [34] where the author provides a thorough examination of the different studies that have been conducted since 2006, when deep learning was presented as a new area of machine learning, for speech applications. The results provided in this paper shed light on the trends of research in this area as well as bring focus to new research topics.

2.2.4 Region Detection: Region or Division detection is one of the important tasks of classifying people from a specific geographical area. We found some work related to speaker's region detection using their voice. The most similar work is done by M. F. Hossain et al [35] where they tried to recognize speakers from different regions of the United Kingdom using voice based on Accent classification. They used a traditional Machine Learning algorithm on speech features extracted by MFCC and showed maximum accuracy of 99% with k-NN on Crowdsourced high-quality UK and Ireland English Dialect speech dataset [36].

Another closely related work is done in different Indian languages to detect four different accents for Bengali, Gujarati, Malayalam and Marathi by Joseph et al [37]. They performed the region detection approach from different languages spoken in India based on their accent using Dynamic Time Warping (DTW) algorithm which shows an impressive result. Danao et al [38] perform a regional accent detection approach for Tagalog language in the Philippines using several classifiers but the Multi-Layer Perceptron (MLP) classifier did the best performance with 93.33% accuracy.

As most of the voice-based region recognition applications work on Accent. We also study some accent-based voice recognition applications. We found that Mannepalli et al [39] introduced a method to identify three different accents namely Coastal Andhra, Rayalaseema and Telangana of Telugu language using Nearest Neighbor Classifier and achieved 72% accuracy. For Chinese accent detection Long et al [40] perform a method based on RASTA – PLP algorithm extracting features known as short-time spectrum of each speech segment and record accuracy of 80.8% using Naïve Bayes classifier. Zheng et al [41] propose a new combination of accent discriminative acoustic features, accent detection, and an acoustic adaptation approach for accented Chinese speech recognition. We don't find any related papers for Bangla Language.

2.2.5 Age Range Recognition: One of the most likely contributions related to estimated age from voice is done by Buyuk et al [42]. They proposed a feed-forward DNN for age recognition from voice. They used long-term and short-term features to train two separate DNN models. They used MFCCs as short-term features to get long term features. Later long-term features were used to train DNN. They found that short-term features perform better than long term features in age identification. M. Mavaddati et al [43] also tried to recognize the age and gender of speakers. They have proposed a new system using generative incoherent models to train MFCC features by sparse non-negative matrix factorization. And as post-processing, they have used an atom correction step. They have used SpeechDat II corpus for this work.
We have also seen some reviewed papers for age estimation using voice data. Moyse et al [44] reviewed the literature on age estimation from faces and voices and highlights some similarities and differences. An Adolescent age estimation is presented by Marcin et al [45] where they introduce a method for evaluating the chronological age of adolescents on the basis of

their voice signal. Their approach suggests that the presented approach can be employed for accurate age estimation during rapid development in children. We cannot find any age or age range estimation contribution for Bangla Language.

2.3 Scope of the Problem

Literature presented a lot of statistics in the field of speech recognition and their application. So now we know the improvement done so far in this field as well as in Bangla Language. When we introduce our plan to present a speaker age and division recognition method from Bangla Speech using Deep Learning Algorithm, we talked about many sub processes like data preprocessing including Noise Reduction, Word Segmentation for some case, Pre-Emphasis, Post-Emphasis and feature selection, and performance comparison. So, the method demands some scope to contribute something new in the field of Speech Recognition as well as for the Bangla Language and its community.

2.4 Research Summary

Overall observation about the research field we are going to contribute is quite clear as we presented a lot of reviews related to Speech Recognition. As far we know that in the field of Speech Recognition application Bangla language is quite latecomer. And if we consider other terms of research with respect to our contribution, we have some good points to present. Most of the researchers work for providing automated speech identifiers or conversion of information form like speech to text. Many of them work with speech recognition algorithms and feature extraction algorithm performance. Speech preprocessing like Noise Reduction and Word Segmentation are also the concern of some researchers. But we tried many of them for our contribution as we intended to present something extraordinary.

2.5 Challenges

Working with speech is quite hard as speech data contains a lot of unwanted entities. Collection of speech data is also hard because of noise and other format issues. Also, there are some performance issues as we face while working with this method. Here are some challenges we face at the time of implementing this whole process:

2.5.1 Data Collection: Data collection was such a challenging issue for this project. As we said we have a collection of 16730 data. So, it was quite challenging to collect the huge amount of data. We have dealt with about 500 individual speakers from two campuses of our university, some children data from different schools and some data from Sylhet by creating Google Form to balance our data. It was becoming more tough after the lockdown by Covid-19 pandemic.

2.5.2 Feature Selection: Feature selection is one of the major technical challenges of this project when we focus on performance and accuracy for the time of implementing this model. We studied a lot of techniques about audio features but ended up with MFCC because of its best performance and the statistics of its use. When we implement the Deep Learning model, we consider four different types of feature set having 26, 58, 41, 128 MFCC features and raw audio. We tested a different Machine learning algorithm with 42 MFCC features of speech audio. We examine different algorithms with the combination of different feature sets for recording the best accuracy. The final result is Artificial Neural Network with 85% accuracy for Division recognition and Convolutional Neural Network with 78% accuracy for Age recognition.

2.5.3 Model Selection: Which model will perform best for our data is another big challenge to answer. We tried a different machine learning algorithm to test the method and compare the result with one another. Then we tried a different deep learning algorithm to test the performance. So before driving to the conclusion we have a lot of challenges to choose the perfect algorithm. We have to consider the amount of data, the number of features to choose the suitable model. We also thought about how an algorithm performs for which type and data and how much data is suitable for which algorithm. Then we build the model with ANN and CNN algorithms.

CHAPTER 3
Research Methodology

3.1 Introduction

A research method is an organized plan for regulating research. It composes the theoretical analysis of the figure of methods and principles related to a sector of knowledge. Our research is to provide a way of recognizing unknown speakers age and division from Bangladesh using their Bengali voice. We discuss a lot about the ultimate motive of choosing and working on this topic as well as the model and algorithm we present in this contribution. We used a deep learning algorithm to implement this model and talked about deep learning, which is actually the trending technology for supervised and unsupervised approach.

3.2 Research Subject and Instrumentation

Research subject is a research area which was reviewed and studied for clearing concepts and producing something innovative. For our work it was to provide a model which will help us to detect unknown speaker's division along with an age range. We learned not only for implementation but also to design models, collect data, process data and train the model which is a very important phase of performing professional projects. The other section is Instrumentation that is the technology and the methods we used. We used Windows platform, Python as a programming language with many packages like Numpy, Pandas, Scikit learn, Matplotlib, Keras, TensorFlow etc. Anaconda application was used for all the training and testing processes.

3.3 Process Workflow

The overall process workflow describes how we plan to perform this work and the steps we follow to reach our goal. We face a lot of problems and reconsider our workflow to generate the best result. The figure 3.3 shows the overall working process of the project.

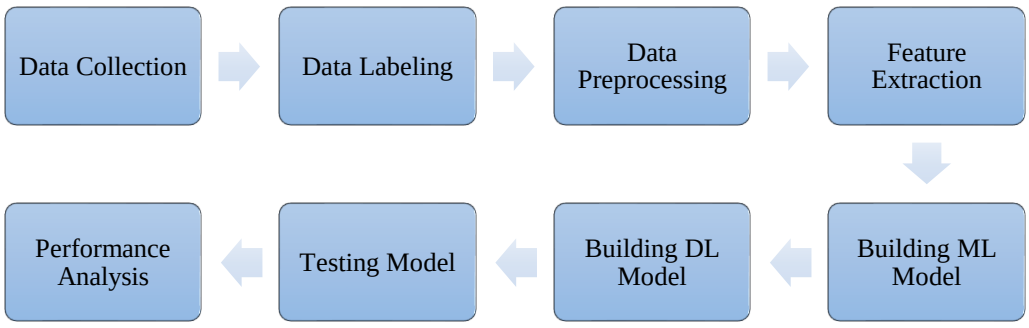


Figure **3.3**: Overall Working Process

3.4 Data Collection

Data collection for the method was one of the most challenging tasks over the entire process. As we said, we used our own dataset containing 613 individual speakers of Daffodil International University and some primary schools from Jamalpur Sadar Upazila. There are several processes we follow. We spend about 10 months collecting total data. First, we tried to collect data from six different booths of the Department of Computer Science and Engineering, Daffodil International University. We also held a day-long campaign to our Permanent Campus of Daffodil International University, Green City, Ashulia, Savar, Dhaka to collect data. Majority of our data collected from the main campus and permanent campus. But some of the data comes from our university male hall. Some of them come from the students of three different sections of OOP course. There are also two different ways of collecting our data we should mention. They are we collect some data from primary school to balance the data for age recognition. Because most of the data are between the age range of 18-26. And some of the data are collected through google form because we have a shortage of Sylhet division data. So, we requested some of our friends and seniors to provide data from the speaker of Sylhet division. Some of our faculty members also help us by providing data. We are highly indebted to all of the people who help us to complete the project by providing one of the biggest Bengali Voice Dataset.

Most of the data is collected from the student by providing project description, motive and outcome. We called the speakers one by one and recorded three different types of script like Story, Poem, News. We give them freedom to read as the speakers wish. We instruct them not to read English words or sentences if there are any mistakes. The data is recorded using Android Smart phone and for major data we use Easy Voice Record (Available on Play Store) software which gives us the flexibility to save the data into .wav format and easy renaming. We consider renaming the audio file as 1_p.wav which means the poem audio of script 1, 57_s.wav which means the story audio of script 57 and 300_n.wav which means the news audio of script 300.

We record Name, Occupation, Age, Gender, District, Height and Weight of every speaker along with the script id he/she read. So that the dataset can be used to recognize most of the characteristics of a speaker.

3.5 Data Description

As we describe the process of data collection, we already know that we have a collection of 613 individual speakers with corresponding speakers' information.
Table 3.5.1 shows some basic information of our dataset for better understanding the data.

TABLE 3.5.1: BASIC INFORMATION OF THE DATASET

Information	Values
Data Type	Audio Data
Format	.wav

Number of Speakers	613 individuals
Number of Raw Data	613 Poems, 613 Story, 613 News
Segmentation Unit	8-10 seconds
Number of data after Segmentation	16730
Total Data Duration	166961 seconds
Script Types	Story, Poem, News

When collecting data, we also collect some information about every single speaker. The attribute we choose is selected according to the usefulness of this dataset.

Table 3.5.2 shows the attribute we consider while collecting data and the varieties of the value of these attributes.

TABLE 3.5.2: DATA ATTRIBUTE AND VALUES

Attributes	Values
ID	Indicate the Script Numbers
Name	Speakers Name
Occupation	Students, Teachers
Age	8-26 Years
Gender	Male, Female
District	64 Districts of Bangladesh
Height	4’11’’ – 6’7’’
Weight	38 – 110 KG

Now as we clear about the nature of data we are driving through the number of data and some distribution of data. We are intended to apply two different applications from this data so label the data like this. We consider 8 different divisions of Bangladesh Barisal, Chittagong, Dhaka, Khulna, Mymensingh, Rajshahi, Rangpur and Sylhet for recognizing speakers from different divisions and we consider 6-18, 19-21, 22-26, three different age groups for speakers' age recognition.

Some graphical Representation of our data can help to understand and analyze our data more clearly. The following figure 3.5.1 shows the Male Female speakers ration of data we use to build our model.

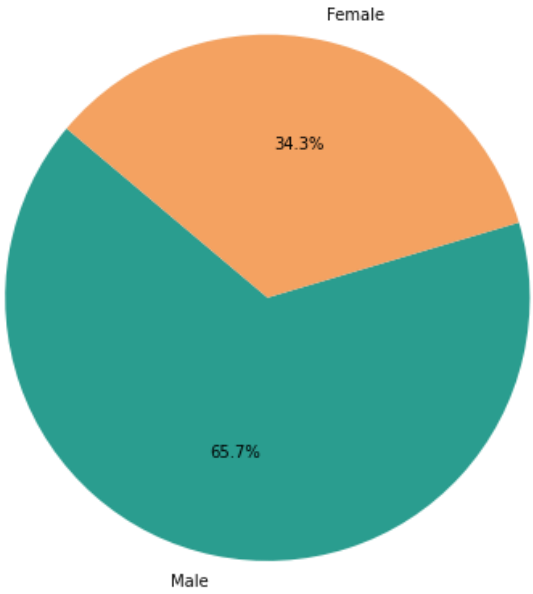


Figure **3.5.1**: Ratio of Male and Female of our Dataset

We already discussed that we label our whole data for Age and Division Recognition differently. The following figure 3.5.2 shows the number of data for different labels of Age recognition model.

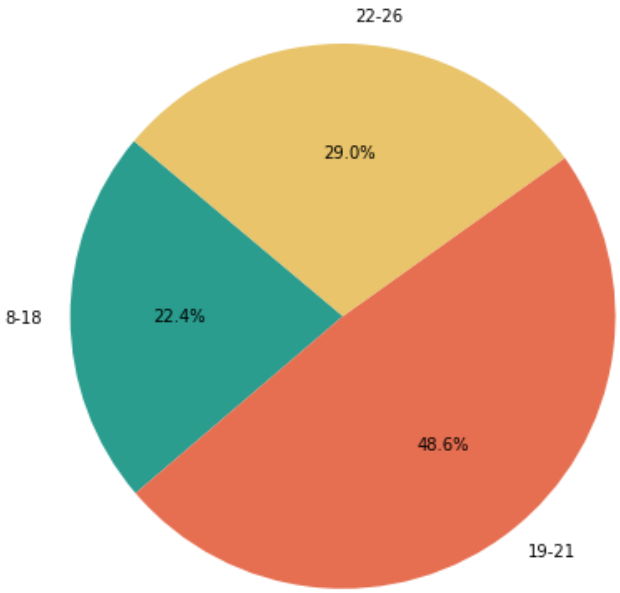


Figure **3.5.2**: Distribution of data for Age Recognition model

The following figure 3.5.3 shows the number of data for different label of Division recognition model.

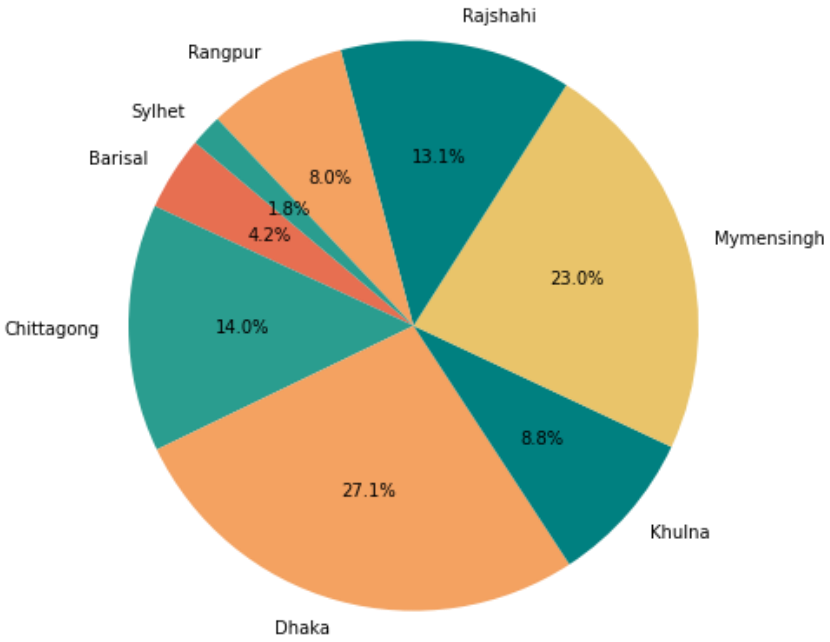


Figure **3.5.3**: Distribution of data for Division Recognition model

3.6 Data Preprocessing

For almost every application, voice signals need to be preprocessed before using it as the input signal. We also perform some preprocessing tasks to make our data usable by machine. As we talked earlier about the basic information about the data which is our data is audio data containing continuous words and collected from different types of recording environments. First of all, we remove the data which have no speaker's information because without the information of speakers the data has no value. Then we look for a duplicate script which means we take note about the script which is recorded more than one time. We also look for the speakers who read more than one script then we perform crossmatching between them and take the unique one.

We are intended to collect all data to a format of .wav so that the audio processing technique would be much easier. Then we receive some data of different audio format types like .mp3, .mpeg and we convert the data into .wav format. As we said our data is collected from different recording environments so some of the speakers record their data in a noisy environment. So, we delete the data containing much noise. And we also perform a noise reduction process over our full dataset to provide clean and accurate data. The noise reduction method is described in our other paper by Fahad et al [46].

Then we focus on labelling them according to the requirements. For region detection we make 8 data labels named Barisal, Chittagong, Dhaka, Khulna, Mymensingh, Rajshahi, Rangpur and Sylhet. For age recognition we make three different age ranges according to the data we collected. For balancing the data, we make the three labels like 6-18, 19-21, 22-26.

Then we perform data segmentation. As our goal is to provide maximum performance with minimum system requirements, time and memory complexity. We intended to segment our data into 10 second. But some of the data segmented into 8 second as the shortage of 10 second. So, on average the data is segmented into 8-10 second. The data segmentation is done by some python script. And finally, the segmented data is ready to be processed by machine.

3.7 Feature Extraction

In the term of Machine Learning or Artificial Intelligence, data features have always been playing one of the most important roles. As features are the representation of data, based on which the model is trained, we need to make sure that we select the features that represent the raw data in the most accurate way.

Continuous speech is one of the toughest data to represent. Starting from the 50s, a huge number of experiments have been done on speech feature extraction and a good number of feature extraction algorithms have been developed. These algorithms can extract different characteristics of speech as features. Among all these features, Cepstral domain features such as MFCCs, LPCCs etc have been the most successful feature to represent continuous speech. And Mel Frequency Cepstral Coefficients (MFCCs) has overperformed Linear Prediction Cepstral Coefficients (LPCCs) in many experiments becoming one of the most popular features in current trends of speech recognition,

In our work, we have experimented with different features such as MFCCs, deltas, delta-deltas, zero crossing rate, spectral flux, pitch, chroma features, rms, spectral centroid, spectral bandwidth, linear spectral features and mel-scaled features. Testing with different algorithms with different subsets of these features, we have found that a combination of MFCCs and delta features with Artificial Neural Network have performed best for division classification and mel-scaled features with CNN have performed best for age classification.

3.7.1 Mel Frequency Cepstral Coefficients (MFCCs): MFCCs were first introduced in the late 70s [47]. After that lots of developments have been made over years. Different implementations have been proposed for different scope of uses. Todor Ganchev, with the team, have compared the most popular four implementations for speaker recognition including the first implementation known as MFCC FB-20 implementation [48].
In our work, we have used MFCC FB-40 implementation to extract the 13 MFCC features. This implementation consists of several steps-

- **Framing:** 13 MFCC features represents 13 portions of the audio signal. To make this more convenient, these 13 portions needed to be at the same stage. But an audio signal is not static, it changes over time. So, the whole signal is divided into some frames assuming that each segment acts as a stationary signal. Default length of the frame is 25 ms [49]. In our dataset, the duration of most audio files is 10 sec and the sample

rate is 16kHz. So, after the framing, there are 400 segments or frames of each audio and each frame has (0.025*16000) = 400 samples.

- **Calculating Power Spectrum:** Power spectrum gives an estimation of frequencies present in each frame. To calculate that, Discrete Fourier Transform has been applied to each frame [49]. There is an equation (1) for this step-

$$P_i(k) = 1/N \left(|\Sigma^{N}_{n=1} (S_i(n)h(n)e^{-i2\pi kn/N}|^2\right) \qquad 1 \leq k \leq k \ldots\ldots..(1)$$

  Here, Pi(k) is the power spectrum of the ith frame, Si(n) is the signal representation of each sample of the ith frame, h(n) is a hamming window and K is the length of the Discrete Fourier Transform.

- **Applying Mel Filtebank:** It is the most important part of MFCC implementation. In this implementation, a set of 40 equal area filters have been used [48]. Mathematical representation (2) of each filter is given below-

$$H_i(k) = \begin{cases} 0 & \text{for} \quad k < f_{b_{i-1}} \\ \dfrac{2\left(k - f_{b_{i-1}}\right)}{\left(f_{b_i} - f_{b_{i-1}}\right)\left(f_{b_{i+1}} - f_{b_{i-1}}\right)} & \text{for} \quad f_{b_{i-1}} \leq k \leq f_{b_i} \\ \dfrac{2\left(f_{b_{i+1}} - k\right)}{\left(f_{b_{i+1}} - f_{b_i}\right)\left(f_{b_{i+1}} - f_{b_{i-1}}\right)} & \text{for} \quad f_{b_i} \leq k \leq f_{b_{i+1}} \\ 0 & \text{for} \quad k > f_{b_{i+1}} \end{cases}, \qquad (2)$$

  Here $\int$ bi represents the boundary of the i^{th} filter and k represent k-th coefficient of the N - point DFT.
  These 40 filters are basically a set of 40 vectors with some non-zero values in different positions of different vectors. Rest of the values of the vectors are zero. When each of the vectors are multiplied with the power spectrum, the sum of each vector represents the energy of the corresponding segment of the power spectrum. After this, there will be 40 energy representations for 40 filters.

- **Taking the log of energies:** Now the logarithm of the energy estimation is taken for each filterbank section.
- **Taking the DCT of log energies:** In the final step, a Discrete Cosine Transform (DCT) is applied to all the log energies. This represents the 40 features of the audio as energy estimation.

After calculating the 40 energy features, we have only taken the first 13 features. Because these first 13 features are linear features and mostly represent the signal. Figure 3.7.1.1 shows a comparison of MFCC feature values between each label.

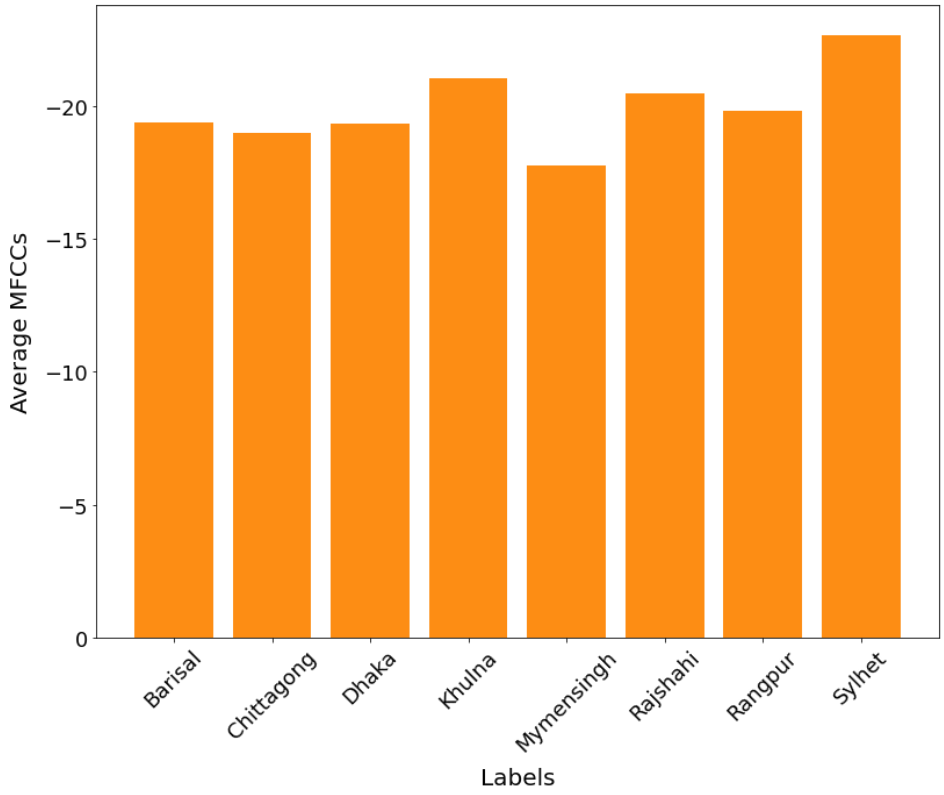


Figure **3.7.1.1**: Average MFCCs for each label

3.7.2 Delta Features: Though MFCC has been the most successful way to represent speech, its major drawback is it contains only static information. Later it has been proven that dynamic features also contain some useful information and a combination of static features and dynamic features can reduce the error rate to half [50] making the model better. So, we used one set of dynamic features known as delta features. Delta features are calculated from MFCC features by the first order derivation. We used 13 MFCC features to calculate 13 delta features. We used this set of dynamic features as an additional set of features to make our model better.

3.7.3 Mel-Scaled Features: This also known as mel spectrogram. In this spectrogram, audio is measured in mel-scale instead of frequency. Humans do not perceive all the frequency ranges in the same scale. They respond better to lower frequencies than higher. For this specific reason, the concept of mel-scale has been proposed. Mel spectrogram is related to pitch which is the closest representation of what humans hear.

In our work, we splitted the whole audio into some chunks using an overlapping window. Then we converted the time series of each chunk into its mel-scaled representation or mel spectrogram [51]. For this, we have used a formula (3) which is given below.

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \qquad 3$$

Here, f denotes the time series values in frequency. The values of spectrogram have been used as mel-scaled features. The following figure 3.7.3.1 shows the Average mel-scaled feature for corresponding age label

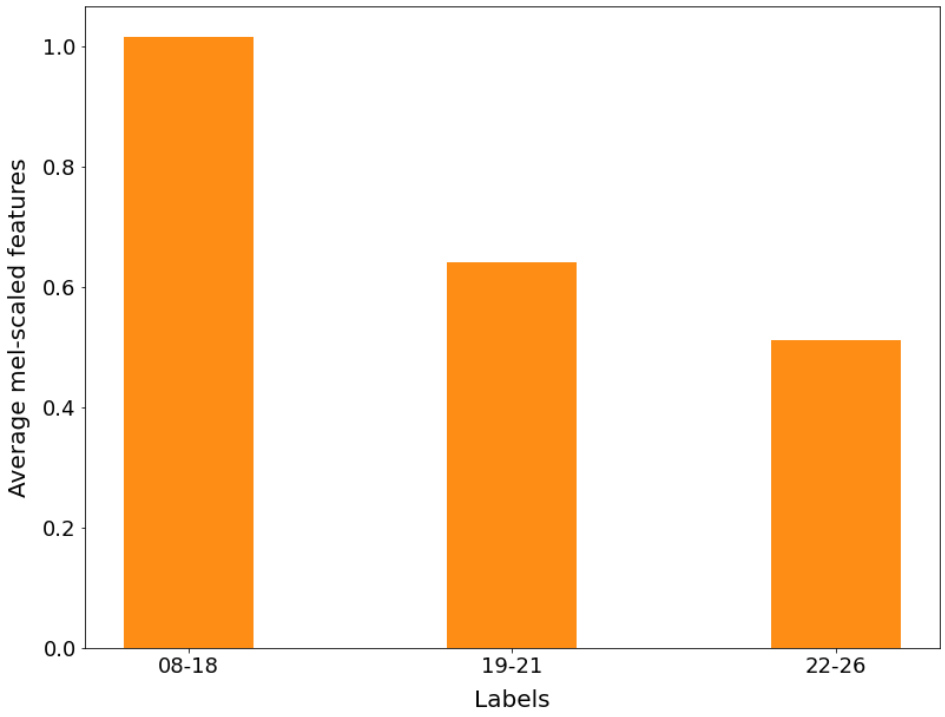


Figure **3.7.3.1**: Average mel-scaled features for each label.

3.8 Machine Learning Model

After the feature extraction, initially we perform some Machine Learning Algorithm to recognize speakers' age and division. We have a lot of data and for supervised learning technique division recognition provides 8 labels to classify. We are having complexity with traditional machine learning approaches. This is why we move into Deep Learning Algorithms.

Before that we tried Decision Tree classifier with random_state=2, k-Nearest Neighbour classifier with n_neighbors = 5, Logistic Regression classifier, Random Forest classifier with n_estimators = 30, Support Vector Machine classifier with kernel= ‘rbf’ for both Age and division recognition with number of features = 43. We will discuss the performance in result section.

3.9 Division Recognition Model

As we already know for division recognition, we use Artificial Neural Network as our final algorithm according to the number of data, label, feature type, system requirement and performance. The raw model with required number of layers and optimization is discussed here. We consider best performance for our data along the model complexity to reduce time and resources.

3.9.1 Bengali Division Recognition Model Architecture

Division Recognition Model

1 : ADAM (learning rate)
2 : For 106 iterations in all batch do:
3 : Dense (Node, Activation)
4 : Dense (Node, Activation)
5 : Dropout (Rate)
6 : Dense (Node, Activation)
7 : Dropout (Rate)
8 : Dense (Node, Activation)
9 : Dense (Node, Activation)
10 : end for

Proposed Bengali Division Recognition Model used fully connected Dense layer for classifying Bengali Division. This model used a fully connected dense layer and some regularization methods like batch normalization [52] and dropout [53].

layer 1 and layer 2 is dense with 256 nodes with ReLU (4) activation. The output of these layers later connected with 20% dropout layer 3.

$$\text{ReLU}(Y) = \text{MAX}(0, Y) \quad (4)$$

The output of layer 3 then goes into layer 4. Layer 4 is a dense layer with 64 nodes. The output of this layer later connected with 20% dropout layer 5. The output of layer 5 then goes into layer 6. Layer 6 is a dense layer with 32 nodes.

The output of layer 6 is connected with a fully connected dense layer 7 with 8 nodes with SoftMax (5) activation which is also the output layer for the model.

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \; for\; j = 1, \ldots k \quad (5)$$

Code and details about Division Classification can be found on GitHub [54].

Figure 3.9.1.1 showing the proposed Division Recognition architecture.

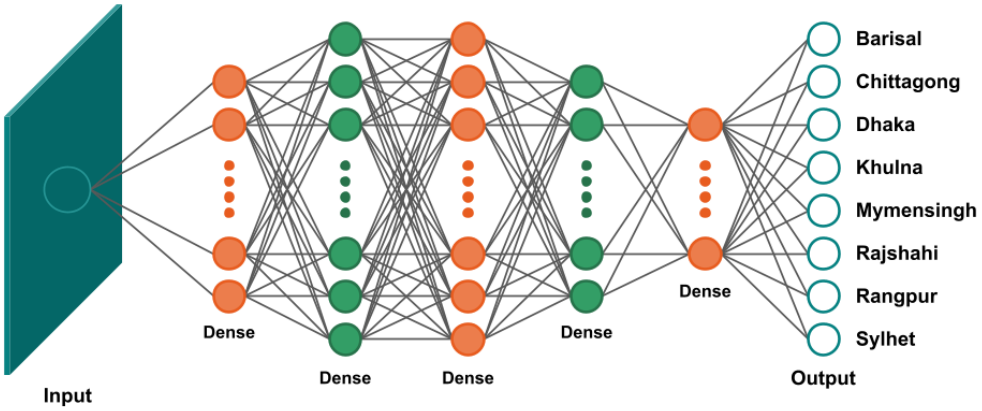


Figure **3.9.1.1**: Architecture of Division Recognition Model

The model architecture summary will help us to determine the model more clearly. Table 3.9.1.1 shows the summary for clear visualization the whole model.

TABLE 3.9.1.1: DIVISION RECOGNITION MODEL ARCHITECTURE SUMMARY

Layer No(type)	Output Shape	Parameter	Connected to
(Input Layer)	128	3456	-
1(Dense)	256	33024	Input Layer
2(Dense)	256	65792	1
3(Dropout)	256	0	2
4(Dense)	64	16448	3
5(Dropout)	64	0	4
6(Dense)	32	2080	5
7(Dense)	8	264	6

Total params: 121,064
Trainableparams:121,064
Non-trainable params: 0

3.9.2 Optimizer and Learning Rate

Researchers will minimize neural network algorithm error with the help of Optimizer algorithm Proposed Bengali Division Classification model used Adam Optimizer [55]. Most of the researchers use it to get better performance. This Optimizer is a reform of stochastic gradient descent algorithm. Updating network weight is a mandatory part of hyper-parameter

tuning which can be done by this optimizer. Proposed Bengali Division Classification used Adam optimizer (6) with a learning rate of 0.001.

$$\theta_{t+1} = \theta_t - \frac{\eta}{\frac{1}{\sqrt{\hat{v}_t}+\varepsilon}} \widehat{m}_t \quad (6)$$

We calculated the error of our model using a function named categorical cross entropy (7) which is needed to optimize the algorithm. A study says that cross entropy has outperformed other loss calculating functions such as classification error and mean squared error etc [56].

$$L_i = -\sum_j t_{i,\ j} \left| \log(p_{i,\ j}) \right| \quad (7)$$

Learning rate has a great impact on hyper-parameter tuning for convolutional neural networks. Low learning rate is good for accuracy as it takes smaller steps towards global optima minimizing the chances of overshooting. But it takes lots of time to reach the global optima. High learning rate can solve this time issue taking bigger steps but there is a chance of overshooting the global minima. And also, accuracy may be compromised. Automatic learning rate reduction method [57] was used to overcome this problem. Initially we set a higher learning rate of 0.001 and it was changed based on the validation accuracy.

3.9.3 Training the Model

The Bengali Division Classification model was trained on our dataset with a batch size of 128. After 35 epochs the model got almost 90% accuracy. The optimizer converged faster by reducing the learning rate with the help of the automatic learning rate method.

3.10 Age Recognition Model

As we already know for age recognition, we use Convolutional Neural Network as our final algorithm according to the number of data, feature type, system requirement and performance. The raw model with required number of layers and optimization is discussed here. We consider best performance for our data along the model complexity to reduce time and resources.

3.10.1 Age Recognition Model Architect

Age Recognition Model

1 : ADAM (learning rate)
2 : For 945 iterations in all batch do:
3 : Convolution 1 (Filter, Kernel Size, Stride, Padding, Activation)
4 : MaxPooling 1 (Pool Size, Stride)
5 : Dropout (Rate)
6 : Convolution 2 (Filter, Kernel Size, Stride, Padding, Activation)
7 : MaxPooling 2 (Pool Size, Stride)
8 : Dropout (Rate)
9 : Convolution 3 (Filter, Kernel Size, Stride, Padding, Activation)
10 : MaxPooling 3 (Pool Size, Stride)

11 : Dropout (Rate)
12 : Convolution 4 (Filter, Kernel Size, Stride, Padding, Activation)
13 : MaxPooling 4 (Pool Size, Stride)
14 : Dropout (Rate)
15 : Convolution 5 (Filter, Kernel Size, Stride, Padding, Activation)
16 : MaxPooling 5 (Pool Size, Stride)
17 : Dropout (Rate)
18 : Dense (Units, Activation)
19 : Dropout (Rate)
20 : Dense (Units, Activation)
21 : Dropout (Rate)
22 : Dense (Units, Activation)
23 : Dropout (Rate)
24 : Dense (Units, Activation)
25 : Dropout (Rate)
26 : Dense (Units, Activation)
27 : end for

Proposed Age Recognition Model used a multilayer CNN for classifying Bangla Handwritten Characters. This model used convolution, Max pooling layer, fully connected dense layer and dropout [53].

Convolution 1 layer consists of 32 filters, kernel of size (3 x 3), stride of size (1 x 1), padding "same" and elu (8) activation function. Next layer is a MaxPooling layer with a pool of size (2 x 2), stride of size (2 x 2) and padding "same". Next layer is a Dropout layer when the dropout rate is 25%.

$$\begin{cases} \alpha\left(e^{x}-1\right) & \text{if } x \leq 0 \\ x & \text{if } x > 0 \end{cases} \quad \text{with parameter } \alpha \tag{8}$$

Convolution 2 layer consists of 32 filters, kernel of size (3 x 3), stride of size (1 x 1), padding "same" and elu activation function. Next layer is a MaxPooling layer with a pool of size (2 x 2), stride of size (2 x 2) and padding "same". Next layer is a Dropout layer when the dropout rate is 25%.

Convolution 3 layer consists of 64 filters, kernel of size (3 x 3), stride of size (1 x 1), padding "same" and elu activation function. Next layer is a MaxPooling layer with a pool of size (2 x 2), stride of size (2 x 2) and padding "same". Next layer is a Dropout layer when the dropout rate is 25%.

Convolution 4 layer consists of 128 filters, kernel of size (3 x 3), stride of size (1 x 1), padding "same" and elu activation function. Next layer is a MaxPooling layer with a pool of size (2 x 2), stride of size (2 x 2) and padding "same". Next layer is a Dropout [8] layer when the dropout rate is 25%.

Convolution 5 layer consists of 128 filters, kernel of size (3 x 3), stride of size (1 x 1), padding "same" and elu activation function. Next layer is a MaxPooling layer with a pool of size (3 x 3), stride of size (3 x 3) and padding "same". Next layer is a Dropout layer when the dropout rate is 25%.

Then flatten the layer and use a Dense layer with 150 units with elu activation and 25% dropout. The output of this layer connected with the Dense layer with 100 units with elu activation and 25% dropout. Then output of this layer connected with the Dense layer with 50 units with elu activation and 25% dropout. After that output of this layer connected with the Dense layer with 20 units with elu activation and 25% dropout. At the final output layer, use 3 units with SoftMax (9) activation.

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \; for\; j = 1, \ldots k \tag{9}$$

Code and details about Age Classification can be found on GitHub [54].

The model architecture summary will help us to determine the model more clearly.

Figure 3.10.1.1 showing the proposed Age Recognition Model Architecture.

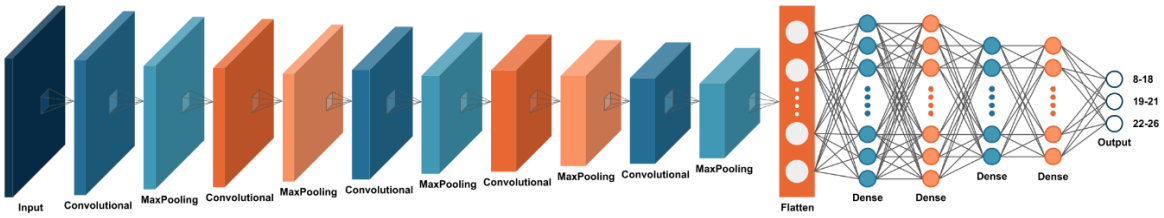


Figure **3.10.1.1**: Architecture of Age Recognition Model.

Table 3.10.1.1 shows the Age Recognition model architecture summary for better visualization of the model.

TABLE 3.10.1.1: AGE RECOGNITION MODEL ARCHITECTURE SUMMARY

Layer No(type)	Output Shape	Parameter	Connected to
(Input Layer)	128, 26, 32	320	-
1(MaxPooling2D)	64, 13, 32	0	Input Layer

2(Dropout)	64, 13, 32	0	1
3(Conv2D)	64, 13, 32	9248	2
4(MaxPooling2D)	32, 7, 32	0	3
5(Dropout)	32, 7, 32	0	4
6(Conv2D)	32, 7, 64	18496	5
7(MaxPooling2D)	16, 4, 64	0	6
8(Dropout)	16, 4, 64	0	7
9(Conv2D)	16, 4, 128	73856	8
10(MaxPooling2D)	8, 2, 128	0	9
11(Dropout)	8, 2, 128	0	10
12(Conv2D)	8, 2, 128	147584	11
13(MaxPooling2D)	3, 1, 128	0	12
14(Dropout)	3, 1, 128	0	13
15(Flatten)	384	0	14
16(Dense)	150	57750	15
17(Dropout)	150	0	16
18(Dense)	100	15100	17
19(Dropout)	100	0	18
20(Dense)	50	5050	19
21(Dropout)	50	0	20
22(Dense)	20	1020	21
23(Dropout)	20	0	22
24(Dense)	3	63	23

Total params: 328,487
Trainable params: 328,487
Non-trainable params: 0

3.10.2 Optimizer and Learning Rate

Optimization algorithms help NN algorithms to minimize the error. Proposed Age Classification used Adam optimizer [55]. Adam optimization algorithm that can be used to update network weights iteratively in training data. Adam is an update of extension to stochastic gradient descent algorithm. For its better performance, it is widely used by numerous researches. Proposed Age Classification used Adam optimizer (10) with a learning rate of 0.001.

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt[1]{\hat{v}_t} + \varepsilon} \widehat{m}_t \quad \textbf{(10)}$$

We calculated the error of our model using a function named categorical cross entropy which is needed to optimize the algorithm. A study says that cross entropy (4) has outperformed other loss calculating functions such as classification error and mean squared error etc.

Learning rate has a great impact on hyper-parameter tuning for convolutional neural networks. Low learning rate is good for accuracy as it takes smaller steps towards global optima minimizing the chances of overshooting. But it takes lots of time to reach the global optima. High learning rate can solve this time issue taking bigger steps but there is a chance of overshooting the global minima. And also, accuracy may be compromised. Automatic learning rate reduction method [57] was used to overcome this problem. Initially we set a higher learning rate of 0.001 and it was changed based on the validation accuracy.

3.10.3 Training the Model

The Age Classification model was trained on our dataset with a batch size of 128. After 35 epochs the model got almost 78% accuracy. The optimizer converged faster by reducing the learning rate with the help of the automatic learning rate method.

3.11 Implementation Requirements

To run this model in the whole algorithm we had to use a good system and environment. The system requirements are given below:

3.11.1 Python 3.8: Python is a high-level programming language. It can be used for desktop GUI and web applications but most importantly it has a rich resource in data science and machine learning. Which actually helps us to complete our work easily. The less complexity and easiness of python programming language increase its acceptance to all the tech enthusiasts. And we used the 3.8+ version in our research work which is the updated version of the time.

3.11.2 Anaconda 4.5.12: This the free and open source distribution of python. This is also available for the R programming language. This is actually a bundle installer. By installing a single thing, it installs lots of necessary tools for data science. Even it comes with a concept of a virtual environment. We can isolate different projects from each other so that we can use different requirements for each of them. We used the 4.5.12 version of anaconda, the updated version of the time.

3.11.3 Jupyter Notebook: We used a Jupyter notebook for writing the code. This is actually a web based open source which allows you to write codes, visualize the data, using equations and much more. We used the 6.0.3 version of jupyter notebook.

3.11.4 Keras: Keras is a deep learning library which is written in python language. This library actually makes the work easy. They already developed the calculation parts. We just use them in the proper place according to our needs. In the backend, keras is using tensorflow.

There are some other libraries which are using keras as backend but keras is the most developed and rich. We used keras 2.3 with tensorflow 2.0 version. This work used keras and tensorflow for experimenting the deep neural network on its dataset.

3.11.5 Scikit learn: This is a python free library which features various classification, clustering and regression algorithms. It makes it easy to use those algorithms. We used multiple algorithms on our work from the library. The used algorithms are listed below:

- Support vector machine
- Logistic regression
- K-nearest neighbor
- Random forest
- Decision Tree

3.10.1 Librosa: Librosa is a feature extraction library from signals. Computers can't work with analog data, so we need to convert them into numeric values. Librosa did this work for us. We extracted 26 features from a single audio file. We used the 0.8 version of librosa.

CHAPTER 4
Experimental Results and Discussion

4.1 Introduction

For recognizing speakers Age and Division we presented two Deep Learning models in this report. Before that we applied some traditional Machine Learning algorithms. The whole performance with necessary parameters is discussed in this chapter. We will also evaluate the performance of different models and their effectiveness. Comparison between different algorithms is also given.

As we already said, we perform some machine learning and deep learning algorithms for our model but we conclude with ANN for division recognition and CNN for age recognition.

4.2 Performance Evaluation

The performance of this whole model can be determined by the ultimate result of the final model. These three models were given after the final simulation of different machine learning, Artificial Neural Network and Convolutional Neural Network. All these models are implemented using our own datasets. For Age and Division Recognition the dataset is labeled as intended and divided into two categories: the training and testing category. The accuracy of the test dataset is known as validation accuracy. Loss is the measurement of how the prediction is missing the actual data. The training loss determines the loss of the training dataset. The validation set determines the loss on the test dataset.

4.3 Performance of Machine Learning Algorithm

When planning to implement the idea of recognizing speakers' age and division we tried some traditional machine learning algorithms. As we have a huge amount of data and we tried to train our model with 75% data where we separate the rest 25% for the test. The problem with machine learning algorithms is it takes a huge time to train and the performance is quite unsatisfactory. For training all machine learning algorithm we take 42 features and the selected features are 13 MFCCs, 13 Deltas, 13 Delta-Deltas, 12 Chroma Features, Zero Crossing Rate, Spectral Flux, Pitch, RMS, Spectral Centroid, Linear Spectral Frequency, Spectral Bandwidth.

Table 4.3.1 shows the performance of different Machine Learning algorithms for Speakers Age and Division Recognition.

TABLE 4.3.1: PERFORMANCE OF DIFFERENT MACHINE LEARNING ALGORITHMS

Algorithm	Accuracy (Age Recognition)	Accuracy (Division Recognition)
KNN	83.19%	73.99%
Logistic Regression	61.18%	43.17%
Random Forest	87.00%	83.10%

Support Vector Machine	79.96%	65.07%
Decision Tree	72.17%	59.83%

As we can see from table 4.3.1, For Age recognition Random Forest classifiers perform the best with the accuracy of 87.00% with n_estimators = 30, random_state=5, max_features=11, StandardScaler parameters. And also, for the Division recognition model Random Forest performs the best accuracy of 83.10% with the same parameters.

4.4 Performance of Division Recognition with Artificial Neural Network

For Division Recognition we finalize the Artificial Neural Network model as our final model. We already discuss the model structure and description of necessary parameters we used to implement this model in the Methodology section. The model was trained and validated on our dataset and gave a promising result on a train set, test set, and validation set.

We have a total data of 16730 segmented audio as we used our 80% data for training the model, we used 10% for test and 10% for validation.

The following figure 4.4.1 shows the accuracy and loss of the division recognition model for training and validation set.

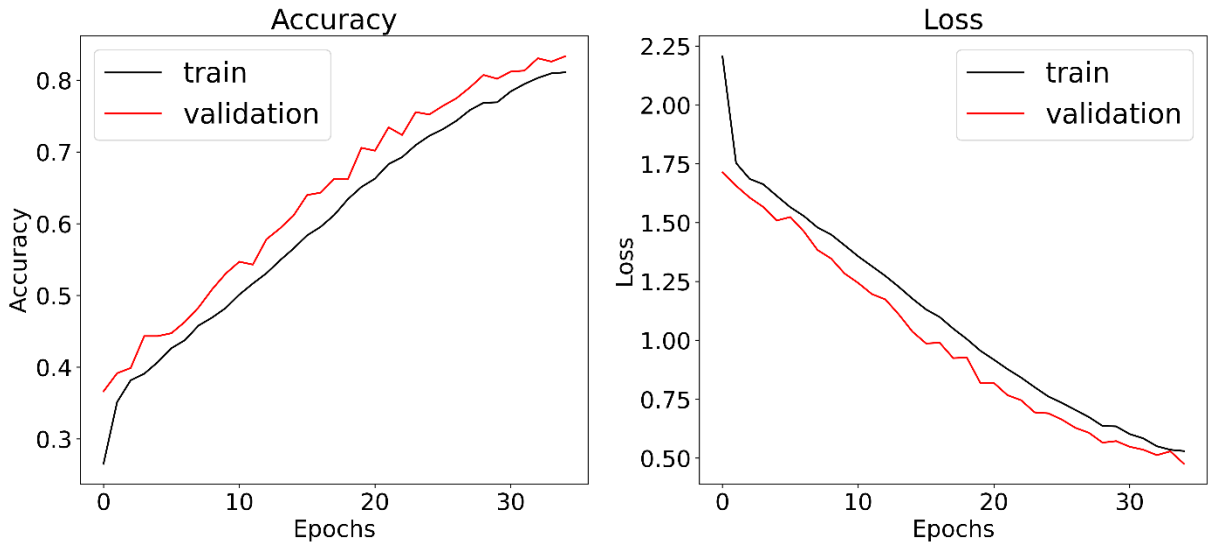


Figure **4.4.1**: Training and Validation Accuracy and Loss of Age Recognition Model

After 35 epochs, the proposed model got 83.99% accuracy on the training set and 85.44% accuracy on a validation set of our datasets.

Figure 4.4.2 shows the confusion matrix of this model so that the model performance will be clearer to all.

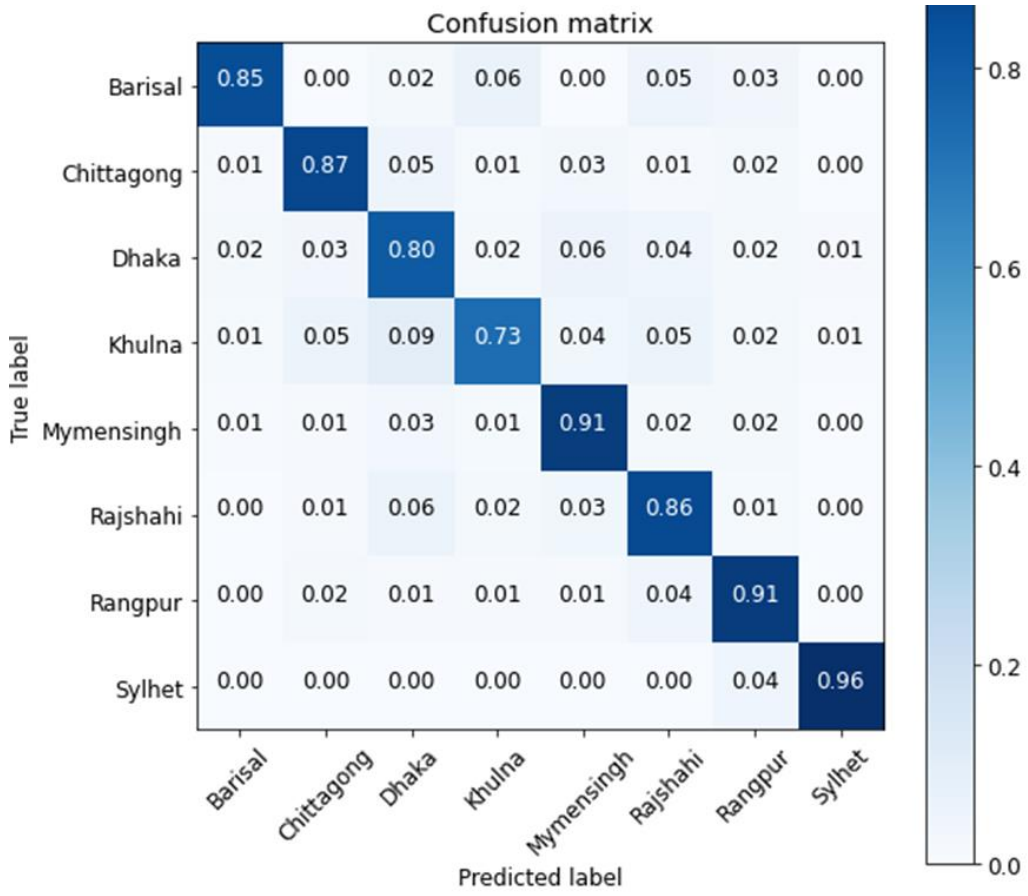


Figure **4.4.2**: Confusion matrix of Division Recognition model

4.5 Performance of Age Recognition with Convolutional Neural Network

For Age Recognition we finalize the Convolutional Neural Network model as our final model. We already discuss the model structure and description of necessary parameters we used to implement this model in the Methodology section. The model was trained and validated on our dataset and gave a promising result on a train set, test set, and validation set.

When we worked with Age recognition, we had a collection of 149283 segmented audio as we used our 80% data for training the model, we used 10% for test and 10% for validation.

The following figure 4.5.1 shows the accuracy and loss of the division recognition model for training and validation set.

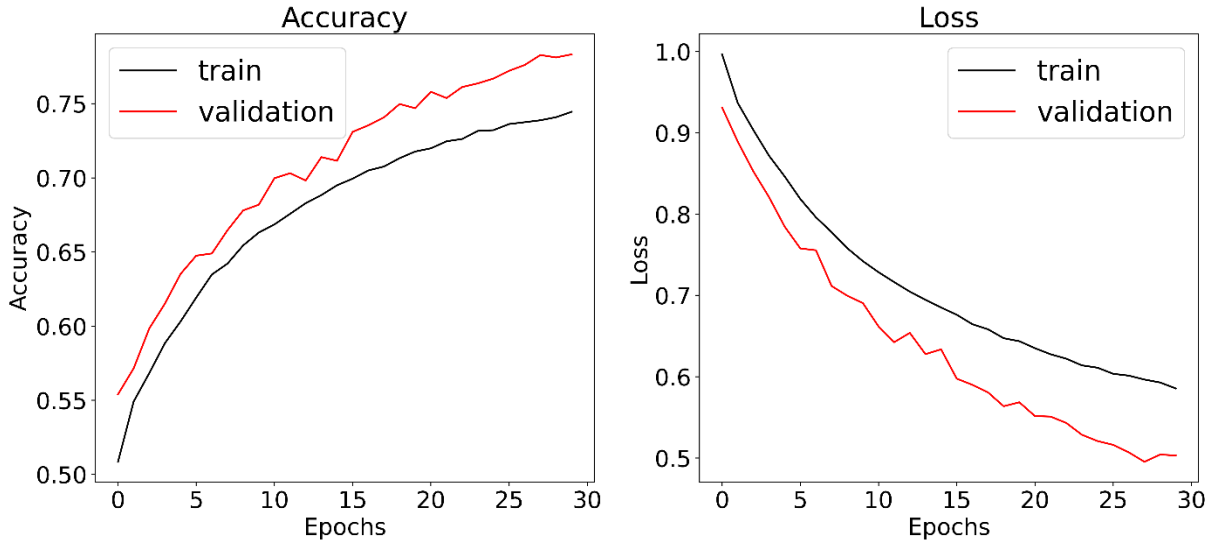


Figure **4.5.1**: Training and Validation Accuracy and Loss of Division Recognition Model

After 30 epochs, the proposed model got 75.30% accuracy on the training set and 78.30% accuracy on a validation set of our datasets.

Figure 4.4.2 shows the confusion matrix of this model so that the model performance will be clearer to all.

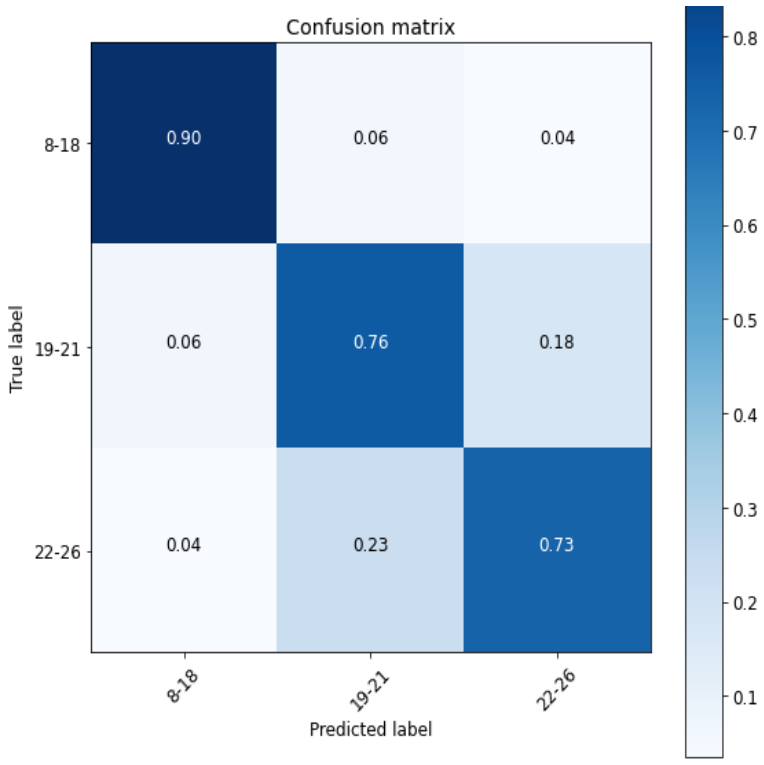


Figure **4.5.2**: Confusion matrix of Age Recognition model

4.6 Summary

In a nutshell, we can say that we have implemented our idea that we had dreamt of. We are intended to recognize speakers' age and division from Bangladeshi local speakers using Bengali voice. And we implemented two final models with 85% and 78% accuracy. We also tried some traditional machine learning algorithms and evaluated their results. If we perform some comparison between all the results, we discussed in this report we can conclude that our final model of ANN for division recognition and CNN for age recognition was the best. We also can improve the result by improving different layers of our model as well as examine the feature type of the audio data. So, we can finally say that we are pretty much successful with our attempt.

CHAPTER 5
Summary, Conclusion, Recommendation and Implication for Future Research

1.1 Summary of the Study

Voice is the most suitable and easy form of human to human communication. But the world is changing from human to computer and vice versa from human to human. So, our research project, which we presented in this report, is intended to provide some application for recognizing a speaker's identity from the speech he/she delivered specially for Bengali Speakers. For this research we implemented only the division recognition and age recognition of a Bengali speakers. We collected a lot of our own data, we studied a lot of related works and all the techniques for voice and audio feature extraction. We applied different traditional machine learning algorithms like SVM, Decision Tree, Random Forest, Logistic Regression, KNN. Then we finalize our two models with Artificial Neural Network and Convolutional Neural Network. The whole research is done successfully as we planned to implement.

1.2 Conclusions

Bangla is the most precious and most beloved language for the Bengali people. Many memories are attached to this language. Making a little contribution in this area satisfies us a lot. Native speakers have their own way of speaking. This way or style is different according to the areas. Huge number of speakers produce a huge amount of audio data and data is valuable when it's processed. From the audio data of Bengali speakers' speech, we tried to recognize the speaker's identity.

To understand these two models, we notice that the Deep Learning model can achieve better performance in classifying and recognizing audio or speech data as well as the field of ASR. For our Division recognition model ANN does a very good job of recognizing Bengali Speakers' Division for its distinctive features as well as CNN for recognizing Bengali Speakers Age recognition. A large number of datasets helps us to train and test our model perfectly and to provide the accuracy of 85% and 78% for Division and Age recognition respectively.

1.3 Recommendations

As now-a-days the area of speech recognition and human-computer interaction is getting vast and larger every single passing day so concentrating on a particular area is a must. Any specific working idea in any specific language is a must to implement something better for the sake of inventing something unique. So, such recommendations regarding these would help.

Firstly, we would recommend the dataset we create to use. The researchers from Bengali community and who are working with Bangla Language base Automated Speech Recognition application.

We also recommend our models in terms of recognizing Speakers age and division as well as recognizing unknown speakers and their identities.

1.4 Implication for Further Study

- We would ty to recognize speakers' gender, occupation, height and weight from Bengali speech using our data.
- We would try to provide some product-based device according to our presented method.
- We would try to improve our proposed algorithm for better performance.
- We would increase our dataset with more speakers from different age and occupation.

REFERENCES

[1] Kuhl, P.K.(2004). Early language acquisition: cracking the speech code. Nat. Rev. Neurosci. 5, 831–843

[2] Yurovsky, Daniel & Yu, Chen & Smith, Linda. (2012). Statistical Speech Segmentation and Word Learning in Parallel: Scaffolding from Child-Directed Speech. Frontiers in psychology. 3. 374. 10.3389/fpsyg.2012.00374.

[3] G. Hongtao, "Forensic Method Analysis Involving VoIP Crime," 2011 Fourth International Symposium on Knowledge Acquisition and Modeling, 2011, pp. 241-243, doi: 10.1109/KAM.2011.71.

[4] Badhon, S M & Rahaman, Md & Rupon, Farea. (2019). A Machine Learning Approach to Automating Bengali Voice Based Gender Classification. 55-61. 10.1109/SMART46866.2019.9117385.

[5] Bahari, Mohamad & Van hamme, Hugo. (2011). Speaker age estimation and gender detection based on supervised Non-Negative Matrix Factorization. 1 - 6. 10.1109/BIOMS.2011.6052385.

[6] Hansen, John & Williams, Keri & Boril, Hynek. (2015). Speaker height estimation from speech: Fusing spectral regression and statistical acoustic models. The Journal of the Acoustical Society of America. 138. 1052. 10.1121/1.4927554.

[7] Mporas, I., Ganchev, T. Estimation of unknown speaker's height from speech. Int J Speech Technol 12, 149–160 (2009). https://doi.org/10.1007/s10772-010-9064-2

[8] Vinyals, Oriol & Friedland, Gerald. (2008). LIVE SPEAKER IDENTIFICATION IN MEETINGS - "WHO IS SPEAKING NOW?

[9] Herbig, Tobias & Gerl, Franz & Minker, Wolfgang. (2010). Detection of Unknown Speakers in an Unsupervised Speech Controlled System. 6392. 25-35. 10.1007/978-3-642-16202-2_3.

[10] Herbig, Tobias & Gerl, Franz & Minker, Wolfgang. (2012). Selflearning speaker identification for enhanced speech recognition. Computer Speech & Language. 26. 210-227. 10.1016/j.csl.2011.11.002.

[11] Amino, Kanae & Arai, Takayuki. (2009). Effects of linguistic contents on perceptual speaker identification: Comparison of familiar and unknown speaker identifications. Acoustical Science and Technology. 30. 89-99. 10.1250/ast.30.89.

[12] Ruiqing Yin, Hervé Bredin, Claude Barras. Speaker Change Detection in Broadcast TV Using Bidi-rectional Long Short-Term Memory Networks. Interspeech 2017, Aug 2017, Stockholm, Sweden. 10.21437/ Interspeech.2017-65. hal-01690244

[13] Badhon, S M & Rahaman, Md & Rupon, Farea. (2019). A Machine Learning Approach to Automating Bengali Voice Based Gender Classification. 55-61. 10.1109/SMART46866.2019.9117385.

[14] Bahari, Mohamad & Van hamme, Hugo. (2011). Speaker age estimation and gender detection based on supervised Non-Negative Matrix Factorization. 1 - 6. 10.1109/BIOMS.2011.6052385.

[15] Hansen, John & Williams, Keri & Boril, Hynek. (2015). Speaker height estimation from speech: Fusing spectral regression and statistical acoustic models. The Journal of the Acoustical Society of America. 138. 1052. 10.1121/1.4927554.

[16] Mporas, I., Ganchev, T. Estimation of unknown speaker's height from speech. Int J Speech Technol 12, 149–160 (2009). https://doi.org/10.1007/s10772-010-9064-2

[17] Sharma, Garima & Umapathy, Kartikeyan & Krishnan, Sridhar. (2020). Trends in audio signal feature extraction methods. Applied Acoustics. 158. 107020. 10.1016/j.apacoust.2019.107020.

[18] Davis, S.V. & MERMELSTEIN, PAUL. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. IEEE Transactions on Acoustics, Speech and Signal Processing. 28. 57-366. 10.1016/B978-0-08-051584-7.50010-3.

[19] Furui, Sadaoki. (1981). Comparison of speaker recognition methods using statistical features and dynamic features. Acoustics, Speech and Signal Processing, IEEE Transactions on. 29. 342 - 350. 10.1109/TASSP.1981.1163605.

[20] Saksamudre, Suman & Shrishrimal, P.P. & Deshmukh, Ratnadeep. (2015). A Review on Different Approaches for Speech Recognition System. International Journal of Computer Applications. 115. 23-28. 10.5120/20284-2839.

[21] Haton JP. (2004) Automatic Speech Recognition: A Review. In: Camp O., Filipe J.B.L., Hammoudi S., Piattini M. (eds) Enterprise Information Systems V. Springer, Dordrecht. https://doi.org/10.1007/1-4020-2673-0_3

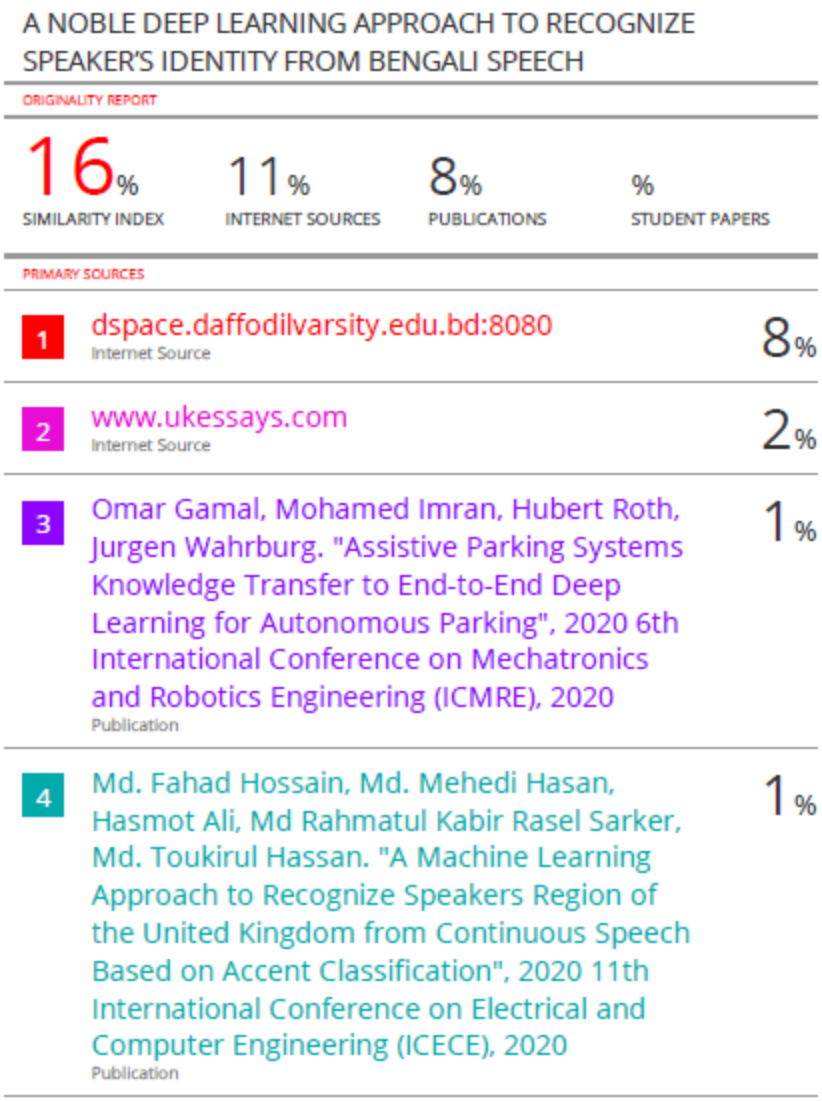
[22] Benk, Sal & Elmir, Youssef & Dennai, Abdeslem. (2019). A Study on Automatic Speech Recognition. 10. 77-85. 10.6025/jitr/2019/10/3/77-85.
[23] Banglapedia, Bangla Language.: (2016 August 30). Available http://en.banglapedia.org/index.php?title=Bangla_Language
[24] Rahman, Md & Roy, Debopriya & Hasan, Md. (2018). Dynamic Time Warping Assisted SVM Classifier for Bangla Speech Recognition. 1-6. 10.1109/IC4ME2.2018.8465640.
[25] Ahammad, Khalil & Rahman, Md. Mahfuzur. (2016). Connected Bangla Speech Recognition using Artificial Neural Network. International Journal of Computer Applications. 149. 38-41. 10.5120/ijca2016911568.
[26] A. K. Paul, D. Das and M. M. Kamal, "Bangla Speech Recognition System Using LPC and ANN," 2009 Seventh International Conference on Advances in Pattern Recognition, 2009, pp. 171-174, doi: 10.1109/ICAPR.2009.80.
[27] Hasnat, M.A., Mowla, J., & Khan, Md. (2007). Isolated and continuous bangla speech recognition: implementation, performance and application perspective
[28] S. M. S. I. Badhon, M. H. Rahaman, F. R. Rupon and S. Abujar, "State of art Research in Bengali Speech Recognition," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-6, doi: 10.1109/ICCCNT49239.2020.9225650.
[29] Bahari, Mohamad & Van hamme, Hugo. (2011). Speaker age estimation and gender detection based on supervised Non-Negative Matrix Factorization. 1 - 6. 10.1109/BIOMS.2011.6052385.
[30] Passricha, Vishal & Aggarwal, Rajesh. (2018). Convolutional Neural Networks for Raw Speech Recognition. 10.5772/intechopen.80026.
[31] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn and D. Yu, "Convolutional Neural Networks for Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 10, pp. 1533-1545, Oct. 2014, doi: 10.1109/TASLP.2014.2339736.
[32] Halageri, Akhilesh, et al. "Speech recognition using deep learning." Int. J. Comput. Sci. Inf. Technol 6.3 (2015): 3206-3209.
[33] Hennebert, Jean & Hasler, Martin & Dedieu, Hervé. (1994). Neural Networks In Speech Recognition.
[34] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh and K. Shaalan, "Speech Recognition Using Deep Neural Networks: A Systematic Review," in IEEE Access, vol. 7, pp. 19143-19165, 2019, doi: 10.1109/ACCESS.2019.2896880.
[35] M. F. Hossain, M. M. Hasan, H. Ali, M. R. K. R. Sarker and M. T. Hassan, "A Machine Learning Approach to Recognize Speakers Region of the United Kingdom from Continuous Speech Based on Accent Classification," 2020 11th International Conference on Electrical and Computer Engineering (ICECE), 2020, pp. 210-213, doi: 10.1109/ICECE51571.2020.9393038.
[36] Demirsahin, Isin & Kjartansson, Oddur & Gutkin, Alexander & Rivera, Clara. (2020). Opensource Multispeaker Corpora of the English Accents in the British Isles
[37] J. Joseph and S. S. Upadhya, "Indian accent detection using dynamic time warping," 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Chennai, 2017, pp. 2814-2817, doi: 10.1109/ICPCSI.2017.8392233.
[38] G. Danao, J. Torres, J. V. Tubio and L. Vea, "Tagalog regional accent classification in the Philippines," 2017IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), Manila, 2017, pp. 1-6, doi: 10.1109/HNICEM.2017.8269545.
[39] K. Mannepalli, P. N. Sastry and V. Rajesh, "Accent detection of Telugu speech using prosodic and formant features," 2015 International Conference on Signal Processing and Communication Engineering Systems, Guntur, 2015, pp. 318-322, doi: 10.1109/SPACES.2015.7058274.
[40] Long, Zhang & Yunxue, Zhao & Peng, Zhang & Ke, Yan & Wei, Zhang. (2015). Chinese accent detection research based on RASTA - PLP algorithm. 31-34. 10.1109/ICAIOT.2015.7111531.
[41] Zheng, Yanli & Sproat, Richard & Gu, Liang & Shafran, Izhak & Zhou, Haolang & Su, Yi & Jurafsky, Daniel & Starr, Rebecca & Yoon, Su- Youn. (2005). Accent detection and speech recognition for Shanghaiaccented Mandarin. 217-220.
[42] Buyuk, Osman & Arslan, Levent. (2018). Age identification from voice using feed-forward deep neural networks. 1-4. 10.1109/SIU.2018.8404322.

[43] Mavaddati, S.. “Voice-based Age and Gender Recognition Based on Learning Generative Sparse Models.” International Journal of Engineering - Transactions C: Aspects 31 (2018): 1529-1535.
[44] Moyse, E., 2014. Age Estimation from Faces and Voices: A Review. Psychologica Belgica, 54(3), pp.255–265. DOI: http://doi.org/10.5334/pb.aq
[45] Bugdol, Marcin D., Bugdol, Monika N., Bieńkowska, Maria J., Lipowicz, Anna, Wijata, Agata M. and Mitas, Andrzej W.. "Adolescent age estimation using voice features" Biomedical Engineering / Biomedizinische Technik, vol. 65, no. 4, 2020, pp. 429-434. https://doi.org/10.1515/bmt-2018-0082
[46] M. M. Hasan, H. Ali, M. F. Hossain and S. Abujar, "Preprocessing of Continuous Bengali Speech for Feature Extraction," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-4, doi: 10.1109/ICCCNT49239.2020.9225469.
[47] Davis, S.V. & MERMELSTEIN, PAUL. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. IEEE Transactions on Acoustics, Speech and Signal Processing. 28. 57-366. 10.1016/B978-0-08-051584-7.50010-3.
[48] Ganchev, Todor & Fakotakis, Nikos & George, Kokkinakis. (2005). Comparative evaluation of various MFCC implementations on the speaker verification task. Proceedings of the SPECOM. 1.
[49] Haton JP. (2004) Automatic Speech Recognition: A Review. In: Camp O., Filipe J.B.L., Hammoudi S., Piattini M. (eds) Enterprise Information Systems V. Springer, Dordrecht. https://doi.org/10.1007/1-4020-2673-0_3
[50] Furui, Sadaoki. (1981). Comparison of speaker recognition methods using statistical features and dynamic features. Acoustics, Speech and Signal Processing, IEEE Transactions on. 29. 342 - 350. 10.1109/TASSP.1981.1163605.
[51] Douglas O'Shaughnessy (1987). Speech communication: human and machine. Addison-Wesley. p. 150. ISBN 978-0-201-16520-3.
[52] Sergey Ioffe, Christian Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”, arXiv:1502.03167 [cs.LG]
[53] Srivastava, Nitish & Hinton, Geoffrey & Krizhevsky, Alex & Sutskever, Ilya & Salakhutdinov, Ruslan. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research.15.1929-1958.
[54] Github[online], https://github.com/fahad35/division-and-age-classfication-model
[55] Kingma, Diederik P. and Ba, Jimmy. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG], December 2014.
[56] Katarzyna Janocha and Wojciech Marian Czarnecki. On loss functions for deep neural networks in classification. Arxiv, abs/1702.05659, 2017
[57] Schaul, Tom, Zhang, Sixin, and LeCun, Yann. No more pesky learning rates. arXiv preprint arXiv:1206.1106,2012.

APPENDIX

Appendix A: Research Reflection

The purpose of this appendix is about research reflection. From Fall 2019 semester we started our journey to research in this field. We collected a lot of data in duration of about 1 years. We have faced a lot of challenges and difficulties. Covid-19 pandemic hamper our research and forces us to do the research from home so that we miss the spirit of working together. We studied a lot about speech feature and audio preprocessing technique. We published three different research paper from our learning we achieve for doing this research. We are planned to implement more application to recognize speakers but conclude with Age and Division recognition.



A NOBLE DEEP LEARNING APPROACH TO RECOGNIZE SPEAKER'S IDENTITY FROM BENGALI SPEECH

ORIGINALITY REPORT

16%	11%	8%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	8%
2	www.ukessays.com Internet Source	2%
3	Omar Gamal, Mohamed Imran, Hubert Roth, Jurgen Wahrburg. "Assistive Parking Systems Knowledge Transfer to End-to-End Deep Learning for Autonomous Parking", 2020 6th International Conference on Mechatronics and Robotics Engineering (ICMRE), 2020 Publication	1%
4	Md. Fahad Hossain, Md. Mehedi Hasan, Hasmot Ali, Md Rahmatul Kabir Rasel Sarker, Md. Toukirul Hassan. "A Machine Learning Approach to Recognize Speakers Region of the United Kingdom from Continuous Speech Based on Accent Classification", 2020 11th International Conference on Electrical and Computer Engineering (ICECE), 2020 Publication	1%

Figure: Plagiarism Report