A Comparative Study of Machine Learning Algorithms for Speaker Identification Using MFCC and Pitch

BY

MD. Asrafi Rahoman Tonmoy
Id: 232-25-001

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

Supervised By
Dr. Md Zahid Hasan
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised By
Dr. Sheak Rashed Haider Noori
Professor & Head
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

NOVEMBER 2024

APPROVAL

This Project/Thesis titled **"A Comparative Study of Machine Learning Algorithms for Speaker Identification Using MFCC and Pitch"**, submitted by **MD. Asrafi Rahoman Tonmoy Id: 232-25-001** 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 M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **11-01-2025**.

BOARD OF EXAMINERS

Chairman

Dr. S.M Aminul Haque
Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Fizar Ahmed
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Md Alamgir Kabir
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Mr. Sadat Hasan
Data Scientist
Risk Management Division,
BRAC Bank Limited

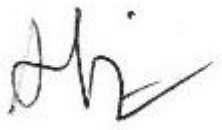
DECLARATION

I hereby declare that the project work entitled **"A Comparative Study of Machine Learning Algorithms for Speaker Identification Using MFCC and Pitch"** Submitted to the Daffodil International University, is a record of original work done by me. Except as acknowledged in the text and that the material has not been submitted, either in whole or in part for a degree at this or any other university.

Supervised By:

Dr. Md Zahid Hasan
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Submitted By:



Md. Asrafi Rahoman Tonmoy
ID: 232-25-001
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to almighty Allah for His divine blessing makes it possible to complete this project successfully.

I feel grateful to and wish my profound indebtedness to Supervisor **Dr. Md Zahid Hasan,** Associate Professor, Department of Computer Science and Engineering, Daffodil International University, Dhaka. My supervisor has deep knowledge and deep interest in computer science to accomplish 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 stages have made it possible to complete this project.

I would like to express my heartiest gratitude to **Dr. Sheak Rashed Haider Noori**, Professor and Head**,** Department of CSE, Daffodil International University for his kind help to finish my project and also to other faculty member and the staff of CSE department Daffodil International University.

I would like to thank my entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, I must acknowledge with due respect the constant support and patience of my parents.

ABSTRACT

The primary objective of this project " **A Comparative Study of Machine Learning Algorithms for Speaker Identification Using MFCC and Pitch**" is to create a reliable and effective system for accurately recognizing speakers through audio input. This project utilizes advanced machine learning methods and audio feature extraction techniques, with a primary focus on Mel-Frequency Cepstral Coefficients (MFCC) and pitch features, to effectively categorize speakers.

During the training process, the system obtains characteristics from clear audio sets and utilizes them to train various machine learning models like Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Gradient Boosting. These models are assessed using important performance criteria such as accuracy, precision, recall, and F1-score. During the testing stage, the trained models are assessed on audio data that has not been seen before in order to evaluate their strength and ability to apply to new situations.

This app showcases how feature extraction methods and machine learning algorithms can be combined effectively to improve the accuracy of speaker recognition. Python is used for the implementation of the system, making use of libraries like Librosa to extract features and Scikit-learn for training and evaluating models. The findings show that this method has promise for practical use in security systems, call centers, and smart assistants. This project offers a solid foundation for speaker recognition and also opens doors for more studies in audio-focused machine learning technologies.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURES **PAGE NO**

LIST OF TABLES

TABLES **PAGE NO**

CHAPTER 1

Introduction

1.1 Introduction

Voice identification, also referred to as speaker identification, involves determining the user's identity through their voice. While speaker verification aims to match a voice to a specific person, speaker recognition involves identifying a voice among a group of identified speakers. Identifying an individual's voice is achievable as every person possesses distinctive vocal characteristics influenced by biological and sociocultural factors. Some time it's easy to identify to us for known persons, because we hear their voice daily regularly. So, we can identify the known person voice even without seeing them. But now we are not limited in finding or identifying speakers by our own. In this modern era, we can use AI to do the same thing for us.

Speaker identification is becoming more difficult due to the growing popularity of voice-controlled technology in today's society. Speaker identification systems are present in many well-known systems today, including voice-user interfaces like Google Assistant and Amazon Alexa, as well as voice authentication in banking and forensic applications. It enhances security of voice authentication for baking, home security, smart device, access control and prevent fraud in law enforcement, it helps verify the suspect's identities in investigation and make it easier. It improves customer care or call center service. It also used in medical sector for monitoring patient. It can be used in talk-show or other situations like that like finding all speakers speech and summarize the context of their speech for newspaper or get information for other purposes.

These systems utilize ML methods to improve the analysis of voice characteristics such as volume, tone, and spectral features. However, environmental noise and overlapping speech in typical situations have prevented the attainment of high levels of accuracy.

This study aims to address these issues by validating multiple automated speaker recognition models. The study's focus includes feature extraction methods, tolerance to

noise, and comparison to find the most appropriate model for real-world use. Furthermore, this study contributes to the development of robust speaker recognition systems.

1.2 Motivation

The motivation of the study on this for this research arises from the growing reliance on voice control system and pressing need of the accurate speaker identification in diverse applications. Voice identification systems have broad research topics to ubiquitous technologies in homes, industries and workplace. Normal authenticating system that we use now have several limitations and voice-based system provide and additional layer of security and make them essential for any sensitive applications or systems like financial transaction and personal device access. Voice interfaces allow us for hands free interaction to making technology more accessible to people with disabilities and improve our user experience in many aspects.

Because of noise, accent variations and overlapping of speech the performance of the existing systems often get degraded. This thing motivates the need of robust models capable of handling these challenges and this study aligns with these motivations by focusing on building a model that will perform well in any type of noisy and real-world conditions.

1.3 Rationale of the study

Speaker identification or recognition system are the core of many cutting-edge tech, including secure access control, virtual assistants and forensic analysis. Most of the existing systems are work good at specific scenarios. This study seeks to bridge this gap by making or testing multiple systemic step and find out best-performing way on under various conditions. All the systems and process work very well in normal or specific scenarios but most of the difficult time this system didn't work properly. To improvement in technology of speaker identification system like voice base security system and speaker's speech summarization there are many reasons behind this study.

This study's finding will provide valuable insights for researchers, developers and industries looking for implement robust speaker identification solutions.

1.4 Research Questions

To achieves the objectives this research, address the following key questions:

i. Which machine learning models will achieve higher accuracy for this speaker identification on my dataset?
ii. Which feature extraction method are effective to capture the unique characteristics of speakers?

1.5 Expected Output

This study aims to provide us a comprehensive evaluation of the speaker's identification systems by comparing the accuracy and the preferences of some model like Random Forest, Gradient Boosting, SVM, Adaboost, LightGBM and CatBoost models by using various metrics. It delves into effects of noise in model accuracy, presenting quantitative insights and strategies to mitigate the impact of the model. This research also focuses on identify the best effective feature extraction process like MFCCs, delta features and pitch-based attributes for improvement model performance.

This study will offer real-world deployment for speaker identification systems and it also discuss the ethical implementation of such technology like privacy issues and potential misuse, emphasize importance of responsible and secure application.

1.6 Report Layout

This thesis report has been organized in six chapters and ensure logical flow of information, offering comprehensive understanding on this research process and findings:

Chapter 1 sets the stage of the study and explain the motivations behind this research and its significance in the field of speaker identification. It clarifies the outlines of the objectives on this research study and specific research questions it seeks to answer. It provides an overview on this research challenges and the impact of the outcomes.

Chapter 2 delves into the theoretical framework underpinning speaker identification, provide a detailed insight of related works and advancement on this field. it also explores existing literature, highlight key process and find out when identifying gaps

that in this research aims to address. This chapter also outlines the challenges inherent to speaker identification like deal with noisy audio data, variability in any voice patterns and providing context to the chosen methodology and focus areas.

Chapter 3 is the research methodology chapter and it describes in detail approach for this study. It includes a comprehensive overview of full dataset and specifying its structure, source and preprocessing methods. The feature extraction techniques like MFCCs, Delta features and pitch-based attributes will be explained. The experimental setup is described including which machine learning models are employed. This chapter will ensure a roadmap for implementing similar systems and ensure similar system.

Chapter 4 this chapter in about experimental result and discussion. This chapter presents the finding of the research study include the detailed results of those model which are used here. The performance of this models is compared by using various metrics and impact of the noise on the accuracy. Visualizations part of this chapter by using some chart or other image will enhance the understanding.

Chapter 5 is about impact analysis. The environmental, socials and ethical implication systems are critically examined in here. It will discuss privacy concerns that connected with the data collection and the storage and potential misuse of the data. It also discus about minimize of the model deployment risk while maximizing benefits.

Chapter 6 will discuss about the conclusion of the thesis and it's works. This Final chapter will summarize the key findings of this thesis and provide practical recommendations for the real-world applications.

CHAPTER 2

Background

2.1 Terminologies

This section defines key terminologies and concepts critical to the thesis, ensuring a clear understanding of the problem. Noise in audio refers to any unwanted sound that disrupts the clarity of the intended signal, such as white noise, electrical noise, or natural sounds. Noise reduction techniques aim to enhance audio clarity by minimizing such disruptions, using tools like Librosa, NoiseReducer, PyDub, and others, each with distinct benefits.

MFCCs (Mel-Frequency Cepstral Coefficients) are essential features that describe the short-term power spectrum of audio signals. They focus on sound frequencies relevant to human perception, making them crucial for tasks like speaker identification. Similarly, pitch represents the perceived frequency of sound and is a key feature for analyzing tonal characteristics in audio.

Classification metrics like accuracy, precision, recall, and F1-score are used to evaluate model performance. Accuracy measures the proportion of correct predictions, while precision and recall focus on the balance between false positives and true positives. The F1-score combines these metrics for a balanced assessment.

Machine learning models like Random Forest, SVM, and Gradient Boosting play key roles in audio signal classification. Random Forest improves accuracy and reduces overfitting through decision tree aggregation, while SVM identifies optimal hyperplanes for data classification. Gradient Boosting captures non-linear relationships through sequential learning. KNN classifies data based on proximity in a non-parametric manner. AdaBoost combines weak classifiers to improve performance by focusing on misclassified data points. LightGBM offers efficient, scalable gradient boosting for large datasets, while CatBoost specializes in categorical data and reduces overfitting. These models provide robust solutions for complex audio classification tasks.

2.2 Related Works

This Part of the report will review prior research and advancements related to noise reduction and audio classification. The research in audio noise reductions and speaker identification is vast, spanning decades and integrated traditional signal processing with modern machine learning. Extensive research has conducted on noise reduction and speaker identification, focusing on various technologies and models:

2.2.1 Noise reduction

Noise reduction methods have undergone significant evolution over decades, driven by advancements in digital signal processing and machine learning.

Foundational works: On of the earliest contribution on this noise reduction was Boll's 1979 paper, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction" [1]. This work introduced spectral subtraction, a technical method that estimate noise by analysis silent segments in signals and subtract it from the noisy segment in spectral domain. It laid the foundation for numerous subsequent algorithms in noise suppression.

Contemporary tools and libraries: Now modern noise reduction methods have some incorporated machine learning techniques and adoptive filtering. Libraries such as Noisereducer, PyDub, Audacity and adobe Audition etc has aim to address a variety of noise types, such as background chatter, hums and white noise provides flexibility across diverse applications like podcast editing and speech enhancement.

2.2.2 Speaker Identification

Speaker identification is a subset od audio classification and it has focused on the extracting and analysis of voice features to distinguish between speakers. Davis and Mermelstein's research of 1980 [2] had established MDCCs as a standard feature for speaker identification by their ability to accurately model the humane vocal tract. Their work remains critical for applications when clean voice is available.

S. M. Kamruzzaman et al. (2010) [3] extended the use of MFCCs by integrate SVMs for identifying speakers, archive high precision and recall in in experimental datasets.

2.2.3 Cancelable Biometrics and Robust Features

Recent advancements like cancelable biometrics have explored distorting MFCCs to enhance data security in speaker recognition. Techniques such as comb filtering have been proposed to address acoustic challenges, like reverberation, in speaker identification systems. [4]

2.2.4 Deep Learning and Advanced Models

Deep learning methods involve embeddings, for example, in the paper titled "Deep Speaker Embeddings for Speaker Recognition" from the year 2018 which are superior to classical features such as MFCCs through deep neural networks for feature extraction. The models are more robust against noisy and degraded speech. [5]

Applications of XGBoost and Gradient Boosting to audio classification tasks have achieved remarkable improvements, especially in handling imbalanced datasets. [4]

2.2.5 Impact of Reverberation

Research by PLOS ONE (2021) has outlined the challenges reverberation places on MFCCs and other cepstral features. Using modeling techniques such as comb filtering, it shows ways in which to improve the reliability of speaker identification systems when in enclosed spaces with high reverberation [5].

2.2.6 Advancements in Neural Network

Artificial neural networks have been implemented to enhance robustness for noisy and reverberant environments in speaker identification systems. It has been shown that they give very good classification accuracy for distorted signals hence furthering the scope for secure speaker recognition.

Noise reduction techniques have greatly improved the clarity of audio and reduced distortion, hence providing a reliable foundation for downstream applications like speaker identification. Features such as MFCC and pitch have remained effective for clean audio, while achieving robustness under degraded conditions often involves advanced techniques like reverberation modeling or the use of embeddings. Among the classification models Random Forest presents the highest accuracy with low interpretability, whereas SVM provides a good trade-off between accuracy and noise robustness. Random Forest and Gradient Boosting are relatively good at handling imbalanced data and show high performance in various conditions. Deep learning approaches, specifically those that use embedding-based methods, have been able to

outperform traditional approaches in noisy and challenging environments by far pushing speaker recognition systems further.

2.3 Comparative Analysis and Summary

The speaker identification techniques have evolved from classical MFCCs established by Davis and Mermelstein to the state-of-the-art techniques with deep learning. Traditional models, such as SVMs and Random Forest work effectively for clean audio, with the best balance between accuracy and robustness given by SVM and with the capability of handling imbalanced datasets given by Random Forest. More advanced techniques like Gradient Boosting and Random forest showed a better performance in different scenarios. However, the classical methods have difficulties in noisy or reverberant environments, while deep learning approaches, especially embedding-based models, have shown much better resistance and accuracy. Recent works like cancelable biometrics enhance data security, while techniques such as comb filtering mitigate issues like reverberation. Neural networks and noise reduction techniques have further bolstered performance, enabling reliable speaker identification even in degraded conditions. While classical methods remain foundational, modern deep learning approaches have redefined robustness and adaptability in complex acoustic environments.

2.4 Scope of the Problem

The problem of speaker identification has a wide scope distinguishing between different speakers based on voice features under various, sometimes adverse conditions. While time has brought developments, noisy, reverberated or otherwise degraded conditions still pose significant challenges in achieving high reliability. Traditional features like MFCCs perform well for clean audio and degrade with environmental distortions, while modern models using deep neural networks are much more robust but come with increased computational overhead. Data security by techniques such as cancelable biometrics, class imbalance problems, and adaptation of the systems to diverse acoustic environments are other major concerns. The problem scales to optimize model performance in various real-world applications, including authentication, forensic data analysis, and personalized human-machine interaction, which demand reliability, adaptability and security.

2.5 Challenges

Several challenges are involved in implementing an effective noise reduction and speaker identification system. These are:

Diverse Noise Environments: Background noise in marketplaces or train stations varies in intensity and frequency, thus demanding the development of a one-size-fits-all solution in practice.

Feature Robustness: Higher-order features such as MFCCs and pitch need sophisticated pre-processing and feature enhancement techniques to remain informative and discriminative in noisy conditions.

Imbalanced Data: Speaker datasets often have varying amounts of data for different individuals, which can bias machine learning models.

Model Complexity vs. Interpretability: While deep learning models offer high accuracy, their complexity makes them less interpretable and harder to deploy in resource-constrained environments.

Computational Requirements: Real-time speaker identification systems demand high computational efficiency, especially in resource-limited devices.

Scalability: The practical scenario implies modification of the models to support new speakers without retraining of the overall system, which turns out to be pretty challenging.

Ethical and Privacy Concernations: Speaker identification systems raise many ethical and data privacy questions regarding misuse in various surveillance contexts.

These challenges necessitate the design of innovative algorithms, robust preprocessing, and careful consideration of deployment scenarios. It is expected that the results and observations obtained from this work will add to the general knowledge about speaker identification systems and their practical applications in real-world noisy environments.

CHAPTER 3

Research Methodology

This chapter describes in detail the approach, tools, and techniques adopted to conduct the research from the collection of data to the final implementation of speaker identification and noise reduction methodologies.

This research projects take up the following subjects for investigation with respect to speaker identification in noisy environments:

Noise Reduction: Using techniques and algorithms that are new to purify the quality of the audio.

Feature Extraction: MFCCs and Pitch are employed, which are proven features for classifying speakers.

Machine Learning Models: Training advanced models constituting Random Forest, Support Vector Machine (SVM), and KNN on classification tasks.

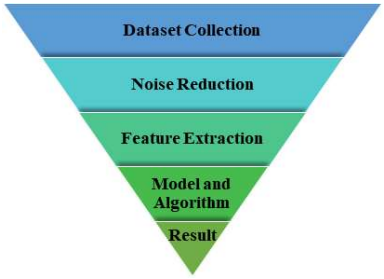


Figure 3.1: The working process to train Speaker Identification Model.

3.1 Data Collection Strategy/Dataset Adopted

A careful data collection and processing approach was adopted to have a strong dataset on which to base the speaker identification process. The following are the steps involved in collecting, processing, and organizing the dataset.

3.1.1 Sources of Data

Social Networking Sites: Audio data were taken from sites like Facebook and YouTube. These provide broad scenarios such as casual conversations, debates, and interviews.

Broadcast Media: Collected recordings from TV programs, which included talk shows and discussions in panels.

Public Addresses: Speeches from open-mic events, public congregations, and lecture sessions provided a greater variance in the acoustic environmental conditions like background noise.

3.1.2 Formats and Tools

The following steps were done so that uniformity and compatibility for the audio analysis are maintained:

Audio Conversion

The multimedia files were first converted into the .wav format, best suited for audio processing tasks, using the **Format Factory** software.

Conversion standardized all the audio files with the same parameters: sampling rate at 16 kHz and outputting into mono-channel sound.

Audio Segmentation

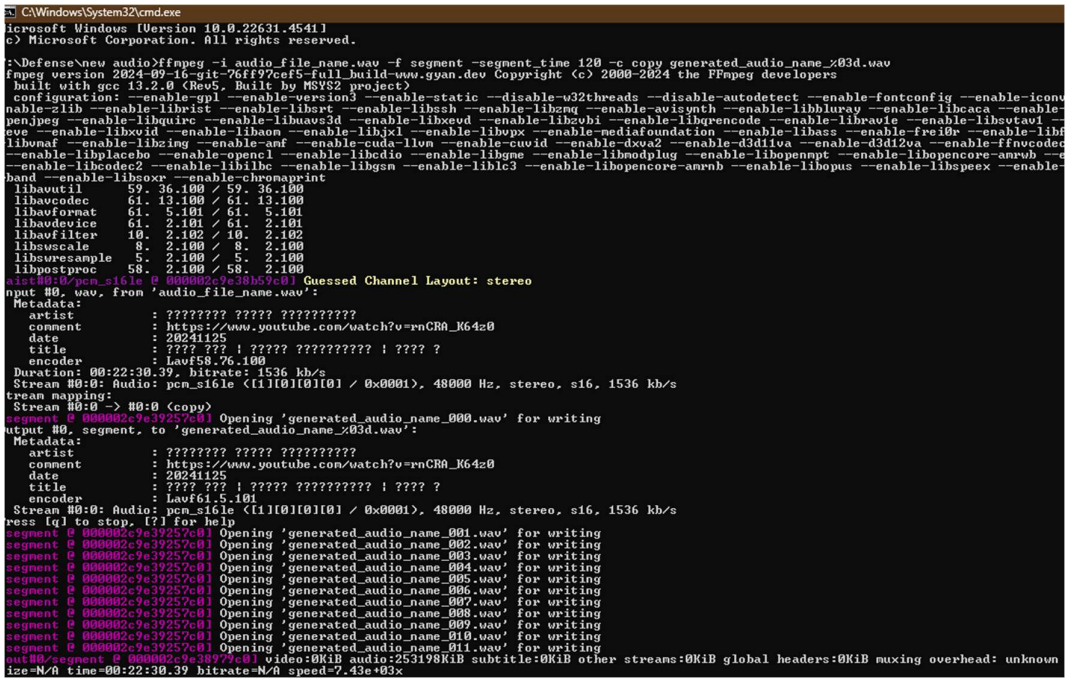


Figure 3.2: Audio segmentation by FFmpeg.

Large audio files were segmented into 10-second clips using the FFmpeg open-source tool with the help of command that given below:

ffmpeg -i audio_file_name.wav -f segment -segment_time 10 -c copy generated_audio_name_%03d.wav

This segmentation facilitated the creation of manageable file sizes for both manual cleaning and feature extraction.

3.2 Data Cleaning and Preprocessing

The collected data underwent meticulous processing to enhance its quality and suitability for training machine learning models.

3.2.1 Manual Cleaning

Each 10-second audio segment was manually reviewed to ensure:

- Only the target speaker's voice was present.
- Audio segments containing other speakers, overlapping conversations, or significant noise were excluded.

This step was quite crucial because contamination with other voices or excessive noise would degrade the model's accuracy both during the training and testing phases.

3.2.2 Noise Removal

Noise reduce and other tools were applied for automatic noise suppression, but that needed to be validated manually to filter out all residual distortions or irrelevant sounds

3.2.3 Characteristics of Dataset

The final dataset was structured such that all the requirements related to effective model training and validation were satisfied. The key characteristics are listed below.

Clean Audio Files

Organized into directories based on the identity of the speaker. Every speaker folder contained cleaned-up and noisy-free audio segments, which were feature extraction ready and train-ready. Then extra noise will be reduced on time of feature extraction process with Noise reducers library.

Train/Test Splits

The dataset had been split into training and testing subsets for model evaluation purposes. The training data was approximately 80% of the dataset, while 20% was used for testing. Care was taken to ensure that the test set samples varied in different scenarios and acoustic environments, and as such, reflected realistic situations.

3.3 Feature Extraction

Here with the help of statistical analysis, the basis for understanding the dataset and evaluating the performance of feature extraction techniques to train any model will be provided.

3.3.1 Exploratory Data Analysis of extricated feature

Speaker's Audio Analysis: The dataset was explored to check if the samples were balanced between speakers-that is, every speaker should have approximate numbers of samples. This balance is very crucial to avoid biases during model training and evaluation.

Table 3.1: Speaker's audio data Amount

Speaker	Train Audio Files	Test Audio Files	Total Audio Files
andalib_rahman	379	160	539
dr_younus	347	144	491
mizanur_rahman	354	145	499
khaled_mohiuddin	290	109	399

3.3.2 Feature Extraction and Data Preparation

The major task in this work is the extraction of suitable features from audio data. This work is done to capture spectral and tonal features of the audio signal, and these features can be used further for tasks such as classification and regression in speech and audio processing applications.

Audio Feature Extraction

The audio features were extracted using a combination of Mel-Frequency Cepstral Coefficients (MFCCs) and pitch.

Mel-Frequency Cepstral Coefficients (MFCCs): MFCCs are one of the most common features in speech and audio processing, capturing the spectral characteristics

of an audio signal in a way that approximates the human auditory system's response. These coefficients are calculated by taking the Fourier transform of an audio signal, then mapping the frequencies onto the Mel scale, which is a scale designed to better align with the perceived pitch of sounds. In this dataset, MFCCs were extracted and stored as the first 13 features (indices 0 to 12), each representing a different frequency component in the audio signal. In this respect, these coefficients represent the timbral features of the sound: its texture or quality.

Pitch: The pitch, which is the fundamental frequency of the sound, was also extracted to capture features of the audio in terms of tone. This was stored as a single feature at index 13. Pitch plays an important role in applications like speech recognition, music analysis, and emotion detection since it may carry information on melody or intonation in speech.

	mfcc_0	mfcc_1	mfcc_2	mfcc_3	mfcc_4	mfcc_5	mfcc_6	mfcc_7	mfcc_8	mfcc_9	mfcc_10	mfcc_11	mfcc_12	pitch	label
0	0.513075	0.435218	0.715368	-0.227198	0.606182	0.262674	-0.479027	-0.061695	0.571843	-0.157947	0.186368	0.243848	-0.246118	0.125145	andalib_rahman
1	0.657652	0.564459	0.730477	0.231358	0.153333	-0.082310	0.337220	-0.214847	-0.110076	0.080439	0.113870	-0.039944	0.065810	0.074256	andalib_rahman
2	0.591905	0.670058	0.811439	0.331165	0.268236	0.224585	0.495061	-0.368993	0.353025	0.181311	0.005897	0.069853	0.088988	-0.008541	andalib_rahman
3	0.594906	0.461390	0.886954	0.120373	0.216855	0.138663	0.519867	-0.269087	0.255363	0.251444	0.201315	0.199732	0.069126	-0.043342	andalib_rahman
4	0.636276	0.669495	0.846709	0.187068	0.041018	0.028654	0.927780	-0.394484	-0.140111	0.321272	-0.016103	0.047631	0.118647	-0.005701	andalib_rahman

	mfcc_0	mfcc_1	mfcc_2	mfcc_3	mfcc_4	mfcc_5	mfcc_6	mfcc_7	mfcc_8	mfcc_9	mfcc_10	mfcc_11	mfcc_12	pitch
count	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000
mean	0.542207	0.454229	0.483435	-0.025910	0.496065	0.301499	0.184030	0.131879	0.069143	-0.115448	0.279903	0.158773	-0.125941	0.278877
std	0.130056	0.234508	0.309621	0.366411	0.232020	0.373744	0.322803	0.342534	0.349822	0.381714	0.364568	0.350228	0.371374	0.234665
min	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000
25%	0.465386	0.311192	0.213348	-0.326805	0.311590	-0.011447	-0.043497	-0.014706	-0.191778	-0.366097	0.123784	-0.004367	-0.343315	0.096516
50%	0.539553	0.486971	0.518948	-0.009169	0.510631	0.266108	0.151112	0.114417	0.047123	-0.148604	0.295024	0.181978	-0.162896	0.251340
75%	0.598794	0.623142	0.775653	0.265650	0.684901	0.644397	0.444153	0.286649	0.303557	0.156497	0.544594	0.388291	0.036857	0.453111
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Figure 3.3: Exploratory Data analysis.

Data Labeling

Each audio sample in this dataset had a class label; that was the target variable to be used for supervised learning. It was placed at index 14 of the feature set and designates the category or class of every audio sample. The class labels are crucial in providing the actual truth for a classification task while training machine learning models.

3.3.3 Descriptive Statistics of Feature's

Descriptive statistics provide an overview of the dataset, summarizing its mean, median, standard deviation, and range for the extracted features. These metrics help

understand the variability and central tendency of the dataset's key variables, including MFCC coefficients and pitch.

Justification of Variables

The primary variables analyzed include MFCC_0 to MFCC_12 and pitch, which are crucial for distinguishing speakers in the audio data.

Mean: Represents the average value of each variable and gives an idea of the central tendency.

Median: The middle value in the sorted data, reflecting the dataset's central point.

Standard Deviation: Quantifies the dispersion or variability around the mean, calculated as:

$$\text{Standard Deviation} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - mean)^2} \quad \text{(ii)}$$

Minimum: The smallest value in the dataset.

Maximum: The largest value in the dataset.

Table 3.2: Descriptive statistics of key variables

Feature	Mean	Median	Std. Dev.	Min	Max
mfcc_0	0.563	0.550	0.125	-1.000	1.000
mfcc_1	0.349	0.365	0.237	-1.000	1.000
mfcc_2	0.306	0.373	0.248	-1.000	1.000
mfcc_3	0.004	-0.029	0.327	-1.000	1.000
mfcc_4	0.495	0.509	0.206	-1.000	1.000
mfcc_5	0.203	0.100	0.331	-1.000	1.000
mfcc_6	0.160	0.182	0.333	-1.000	1.000
mfcc_7	0.042	0.033	0.301	-1.000	1.000
mfcc_8	0.009	0.003	0.306	-1.000	1.000
mfcc_9	-0.015	-0.048	0.334	-1.000	1.000
mfcc_10	0.051	0.056	0.297	-1.000	1.000
mfcc_11	0.118	0.118	0.314	-1.000	1.000
mfcc_12	0.188	0.183	0.238	-1.000	1.000
pitch	0.163	0.100	0.232	-1.000	1.000

Correlation Analysis

The correlation matrix provides a graphic overview of the linear relationships that exist between features in a dataset, with values varying from -1 (perfect negative correlation) to 1 (perfect positive correlation). Features with high positive values are indicative of similar trends, while negative values imply inverse relationships. This gives an idea of identifying redundant variables that may need to be eliminated or reduced in dimensionality to avoid overfitting. Moreover, features that are not highly correlated with others could carry unique information that is useful for speaker identification. This matrix is important in feature interaction and optimization of the dataset for machine learning models.

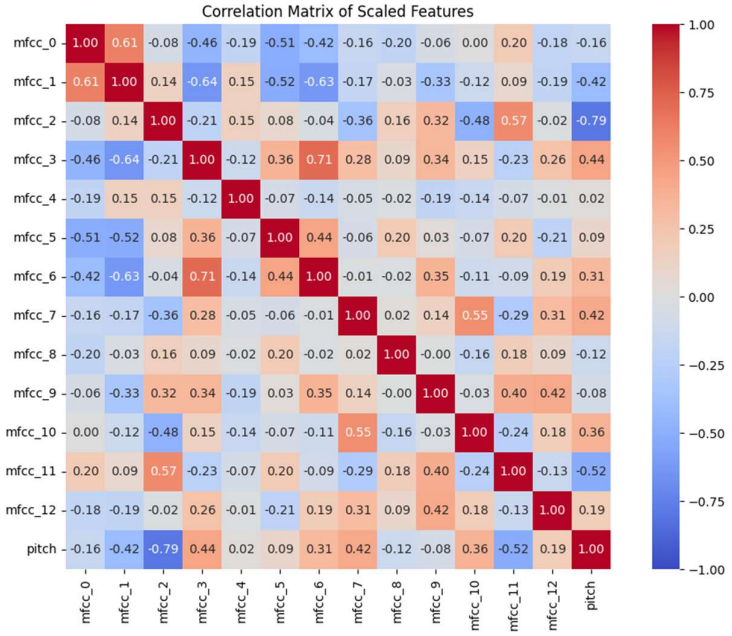


Figure 3.4: Correlation Matrix of Scaled Features

The results indicate a moderate correlation between MFCC coefficients but a weak correlation between MFCCs and pitch, reinforcing their complementary role in speaker identification.

After feature extraction, the data was organized into a pandas DataFrame to make manipulation and analysis easier. This DataFrame had the following columns:

- **Indices 0 to 12:** MFCC coefficients, representing spectral features of the audio.

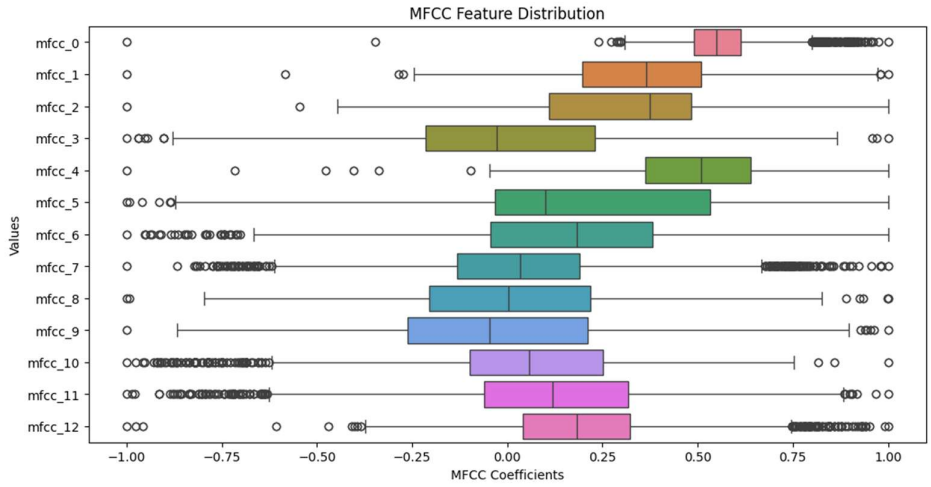


Figure 3.5: MFCC Feature Distribution

- **Index 13:** Pitch, representing the tonal feature of the audio.

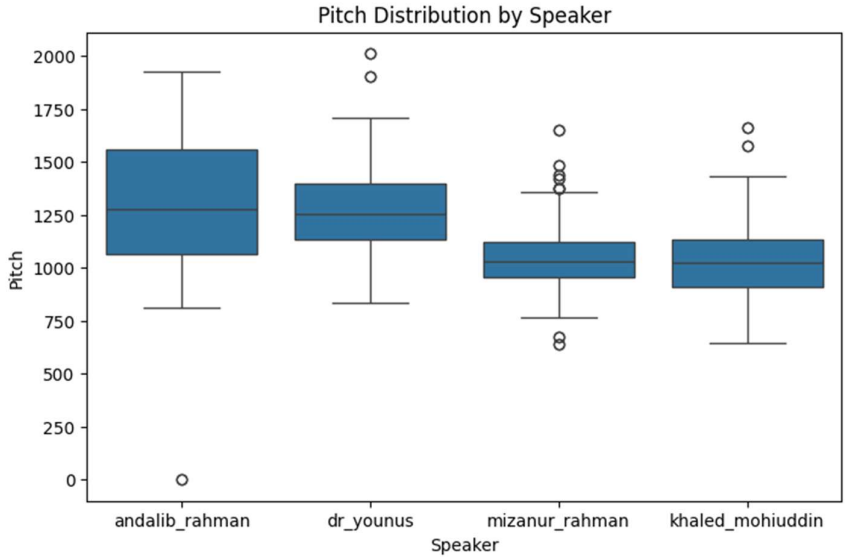


Figure 3.6: Pitch Distribution by Speaker

- **Index 14:** Data label, representing the class or category of the audio sample.

In this way, the data was prepared for further analysis or modeling, where machine learning techniques could be applied to build predictive models based on the extracted features.

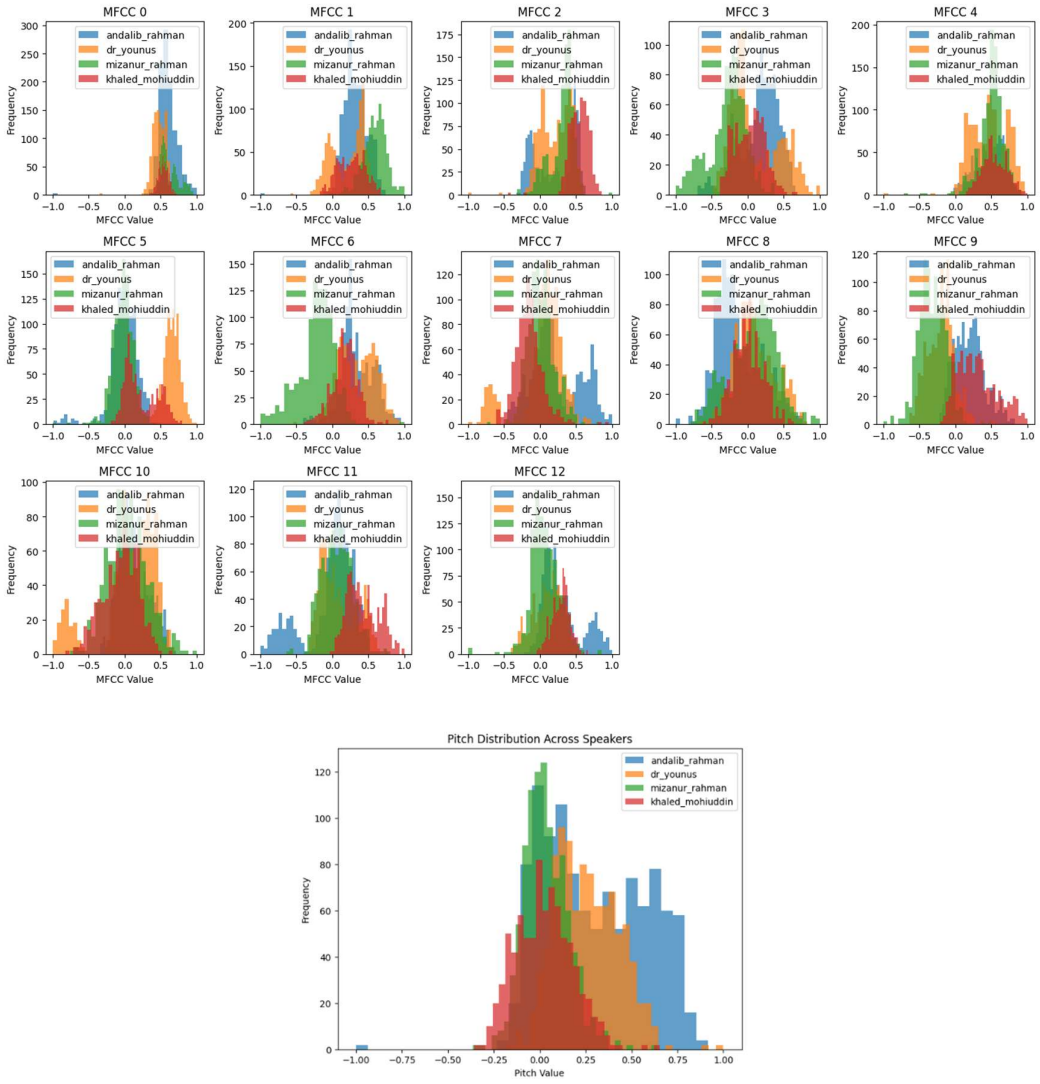


Figure 3.7: Plot MFCC and Pitch distributions for each speaker

3.4 Machine Learning Models

To understand how different machine learning models, perform in speaker identification, we applied various techniques to classify speakers based on MFCC and pitch features extracted from the audio dataset. Each model's performance was evaluated to analyze their ability to capture complex relationships within the features and find out the correct speaker.

Selection of Models

Following models have been implemented and compared using machine learning:

Random Forests: Known for handling feature interactions and noise, this ensemble model was used to classify speakers effectively.

Support Vector Machines (SVM): Explored for its ability to find an optimal hyperplane for classification with high accuracy.

Gradient Boosting: Leveraged to capture non-linear feature relationships through sequential learning.

K-Nearest Neighbors (KNN): Evaluated for its simplicity and effectiveness in identifying speaker clusters based on feature similarity.

AdaBoost: Used to improve classification performance by combining weak classifiers and focusing on misclassified data points, enhancing model accuracy. **LightGBM:** Leveraged for its high efficiency and scalability in handling large datasets, providing fast and accurate gradient boosting for classification tasks. **CatBoost:** Specialized in handling categorical features, CatBoost reduces overfitting while offering strong performance on complex datasets with minimal preprocessing

This multi-model approach helped in identifying the best-performing algorithm for speaker identification and provided insight into feature relevance and model strengths.

3.4.1 Random Forest Classifier for Speaker Identification

For this speaker identification task, a Random Forest algorithm was employed for being robust and handling nonlinear data relations. Random Forest is an ensemble learning technique that builds multiple decision trees during training and outputs the class that is the mode of the classes predicted by the individual trees. This makes it very strong in classification tasks, especially for scenarios with diverse feature sets such as MFCCs and pitch.

Model Training

Scaled Random Forest model was implemented here bye using RandomForestClassifier with a minimal number of estimators (n_estimators=2) for simplicity and quick evaluation. The model was trained on scaled training data (scaled_X_train_rf and scaled_y_train_rf) to ensure features were normalized, enhancing performance and interpretability in a controlled experimental setup.

```
scaled_rf_model=RandomForestClassifier(n_estimators=2,  random_state=42)
scaled_rf_model.fit(scaled_X_train_rf, scaled_y_train_rf)
```

Training and Predictions

The model was predicted by this line:

scaled_y_pred_rf = scaled_rf_model.predict(scaled_X_test_rf)

Model Evaluation

Precision, recall, and F1-score were almost perfect for both classes, reflecting the model's ability to minimize false positives and negatives.

The detailed metrics are summarized as follows:

	precision	recall	f1-score	support
andalib_rahman	0.92	1.00	0.95	213
dr_younus	0.95	0.99	0.97	193
khaled_mohiuddin	0.97	0.89	0.93	169
mizanur_rahman	1.00	0.94	0.97	197
accuracy			0.96	772
macro avg	0.96	0.95	0.96	772
weighted avg	0.96	0.96	0.96	772

Figure 3.8: Classification result of Random Forest.

3.4.2 Support Vector Machine (SVM) Classifier for Speaker Identification

Support Vector Machine is a famous algorithm that was checked against all classified speakers. The SVM can identify the best hyperplane that distinguishes between classes in high dimensional spaces. Hence, this makes the SVM suitable for speaker identification tasks based on features such as MFCC and pitch.

Model Training

A Scaled Support Vector Machine (SVM) model was implemented using the SVC class with a specified random state (random_state=42) for reproducibility. The model was trained on scaled training data (scaled_X_train_svm and scaled_y_train_svm) to ensure

feature normalization, enhancing its effectiveness in separating classes within the dataset.

```
scaled_svm_model = SVC(random_state=42)
scaled_svm_model.fit(scaled_X_train_svm, scaled_y_train_svm)
```

Training and Predictions

The model was predicted by this line:

```
scaled_y_pred_rf = scaled_rf_model.predict(scaled_X_test_rf)
```

Model Evaluation

	precision	recall	f1-score	support
andalib_rahman	0.78	0.58	0.67	213
dr_younus	0.76	0.82	0.79	193
khaled_mohiuddin	0.70	0.67	0.69	169
mizanur_rahman	0.71	0.86	0.78	197
accuracy			0.73	772
macro avg	0.74	0.74	0.73	772
weighted avg	0.74	0.73	0.73	772

Figure 3.9: Classification result of SVM.

3.4.3 K-Nearest Neighbors Classifier for Speaker Identification

The K-Nearest Neighbors algorithm was tested for distinguishing the two speakers,namely, andalib_rahman and dr_younus and others. KNN assigns a class label depending on the majority class of the nearest neighbors in the feature space.

Model Training

In the code, a KNN model is implemented using KNeighborsClassifier with the number of neighbors equal to 5. Using the proximity-based classifier model, the model would predict the majority class observed among the most similar neighbors for both training and testing datasets- X_train_knn and y_train_knn, respectively.

```
knn_model = KNeighborsClassifier(n_neighbors=5)

knn_model.fit(X_train_knn, y_train_knn)
```

Training and Predictions

The model was predicted by this line:

scaled_y_pred_knn = scaled_knn_model.predict(scaled_X_test_knn)

Model Evaluation

	precision	recall	f1-score	support
andalib_rahman	0.95	0.91	0.93	213
dr_younus	0.98	0.95	0.97	193
khaled_mohiuddin	0.93	0.96	0.94	169
mizanur_rahman	0.95	0.99	0.97	197
accuracy			0.95	772
macro avg	0.95	0.95	0.95	772
weighted avg	0.95	0.95	0.95	772

Figure 3.10: Classification result of KNN.

3.4.4 Gradient Boosting Classifier for Speaker Identification

Gradient Boosting is a powerful ensemble learning algorithm that was used to classify speakers andalib_rahman , dr_younus all other twos. By iteratively improving weak learners (decision trees), the model achieved exceptional performance in this speaker identification task.

Model Training

A Gradient Boosting model was implemented with the Scaled Gradient Boosting Classifier using 100 estimators for iterative learning (n_estimators=100) and a specified random state for reproducibility (random_state=42). This was trained on the scaled training dataset, scaled_X_train_gb and scaled_y_train_gb, so that it is sequentially able to optimize performance, correcting errors from previous iteration:

scaled_gb_model = GradientBoostingClassifier(n_estimators=100, random_state=42)

scaled_gb_model.fit(scaled_X_train_gb, scaled_y_train_gb)

Training and Predictions

The model was predicted by this line:

scaled_y_pred_knn = scaled_gb_model.predict(scaled_X_test_gb)

Model Evaluation

	precision	recall	f1-score	support
andalib_rahman	1.00	1.00	1.00	213
dr_younus	1.00	0.99	0.99	193
khaled_mohiuddin	0.99	1.00	0.99	169
mizanur_rahman	1.00	1.00	1.00	197
accuracy			1.00	772
macro avg	1.00	1.00	1.00	772
weighted avg	1.00	1.00	1.00	772

Figure 3.11: Classification result of Gradient Boosting.

3.4.5 AdaBoost Classifier for Speaker Identification

AdaBoost an ensemble learning algorithm was utilized for speaker identification tasks involving speakers like Andalib Rahman and Dr. Younus. The model improves classification accuracy by focusing on misclassified instances through iterative training of weak classifiers.

Model Training

An AdaBoost model was implemented using the AdaBoostClassifier with 100 estimators and a random state for reproducibility.

```
ada_boost_model = AdaBoostClassifier(n_estimators=100, random_state=42)

ada_boost_model.fit(scaled_X_train_gb, scaled_y_train_gb)
```

Training and Predictions

The model's predictions were made with the following code:

```
scaled_y_pred_ada = ada_boost_model.predict(scaled_X_test_gb)
```

Model Evaluation

	precision	recall	f1-score	support
andalib_rahman	0.97	0.93	0.95	213
dr_younus	0.87	0.97	0.92	193
khaled_mohiuddin	0.89	0.79	0.84	157
mizanur_rahman	0.98	0.98	0.98	209
accuracy			0.93	772
macro avg	0.92	0.92	0.92	772
weighted avg	0.93	0.93	0.93	772

Figure 3.12: Classification result of AdaBoostClassifier.

3.4.6 LightGBM Classifier for Speaker Identification

LightGBM is a gradient boosting framework known for its speed and efficiency in handling large datasets. It was used in the speaker identification task, focusing on high performance with minimal training time.

Model Training

A LightGBM model was implemented with parameters optimized for speaker identification:

```
lgbm_model = LGBMClassifier(n_estimators=100, random_state=42)

lgbm_model.fit(scaled_X_train_gb, scaled_y_train_gb)
```

Training and Predictions

Predictions from the LightGBM model were generated as follows:

```
scaled_y_pred_lgbm = lgbm_model.predict(scaled_X_test_gb)
```

Model Evaluation

	precision	recall	f1-score	support
andalib_rahman	1.00	1.00	1.00	213
dr_younus	1.00	0.99	0.99	193
khaled_mohiuddin	0.99	0.99	0.99	157
mizanur_rahman	0.99	1.00	1.00	209
accuracy			0.99	772
macro avg	0.99	0.99	0.99	772
weighted avg	0.99	0.99	0.99	772

Figure 3.13: Classification result of LightGBM model.

3.4.7 CatBoost Classifier for Speaker Identification

CatBoost is an efficient gradient boosting model that handles categorical features well and reduces overfitting. It was employed in the speaker identification task to provide robust predictions on complex datasets.

Model Training

The CatBoost model was trained with the following implementation, focusing on handling categorical data effectively:

catboost_model = CatBoostClassifier(iterations=100, random_state=42, cat_features=categorical_features)

catboost_model.fit(scaled_X_train_gb, scaled_y_train_gb)

Training and Predictions

Predictions were made using the trained CatBoost model:

scaled_y_pred_catboost = catboost_model.predict(scaled_X_test_gb)

Model Evaluation

	precision	recall	f1-score	support
andalib_rahman	1.00	1.00	1.00	213
dr_younus	1.00	0.99	0.99	193
khaled_mohiuddin	0.99	1.00	0.99	157
mizanur_rahman	1.00	1.00	1.00	209
accuracy			1.00	772
macro avg	1.00	1.00	1.00	772
weighted avg	1.00	1.00	1.00	772

Figure 3.14: Classification result of Gradient Boosting.

3.5 Implementation Requirements

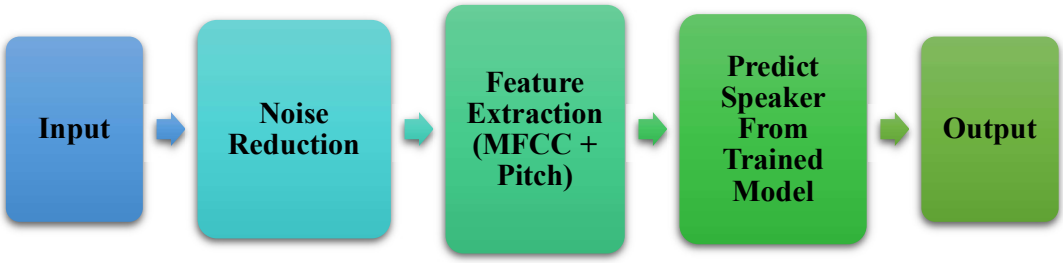


Figure 3.15: Predicting process after implementation

This speaker identification system will be implemented using a software tool, hardware capability, and data preparation step.

Software Requirements: The main programming language is Python, extended with the Scikit-learn library for machine learning, Librosa for audio processing, and

Matplotlib/Seaborn for visualization. For development, the environment is Jupyter Notebook or Google Colab.

Hardware Requirements: Advanced multi-core CPU with a minimum of 8 GB RAM to handle datasets and model training; at least 10 GB storage for datasets and models; a GPU is optionally required for large-scale processing.

Data Requirements: The dataset should contain cleaned, labeled audio data that is organized for training and testing. Preprocessing involves resampling, cleaning noise, normalization, and feature extraction; examples include MFCC and pitch.

Model Training and Evaluation: The cross-validation ensures the model generalizes well. For its evaluation, metrics such as accuracy, precision, recall, and ROC-AUC will be used. The environment should allow for stable training and tuning of hyperparameters.

Deployment: The trained models can be deployed using frameworks like Flask or FastAPI and can be scaled for more speakers and audio data without major retraining.

CHAPTER 4

Experimental Results and Discussion

This chapter gives a thorough explanation of the experimental arrangement, outcomes from the models used, and an analysis of the discoveries. The emphasis is placed on assessing how well the chosen characteristics and machine learning models perform in speaker identification tasks.

4.1 Experimental Setup

The experiments were conducted in a controlled environment to ensure reproducibility and accuracy. Below are the key details:

Table 4.1: Experimental Setup

Aspect	Description
Programming Language	Python (Version 3.x)
Libraries Used	Scikit-learn, Librosa, Matplotlib, Seaborn
Hardware	Multi-core CPU, 8 GB RAM
Dataset	Clean audio files categorized into training and testing sets of speakers (andalib_rahman, dr_younus, mizanur_rahman and khaled_mohiuddin).
Features Extracted	Mel-Frequency Cepstral Coefficients (MFCCs), pitch features
Evaluation Metrics	Accuracy, Precision, Recall, F1-Score, ROC-AUC
Validation Method	5-fold Cross-Validation

Data Preparing and Processing

Audio Characteristics: MFCCs were selected due to their capacity to depict the quality of speech, while pitch characteristics aided in capturing individualized tonal attributes of speakers.

Data Splitting: The data set was divided into training and testing datasets at an 80:20 ratio. The training data was employed for model training, while the test data was kept aside for evaluating performance.

Cross-Validation: Model generalizability was evaluated using 5-fold cross-validation. This approach guarantees that the assessment is not influenced by a particular train-test division.

4.2 Results from experiments and their analysis

The empirical findings and examinations rely on the concepts and methods outlined in Chapter 3. The part assesses how well machine learning models perform by considering important factors like true positives, true negatives, false positives, and false negatives to gauge their predictive precision.

4.2.1 Evaluation Metrics

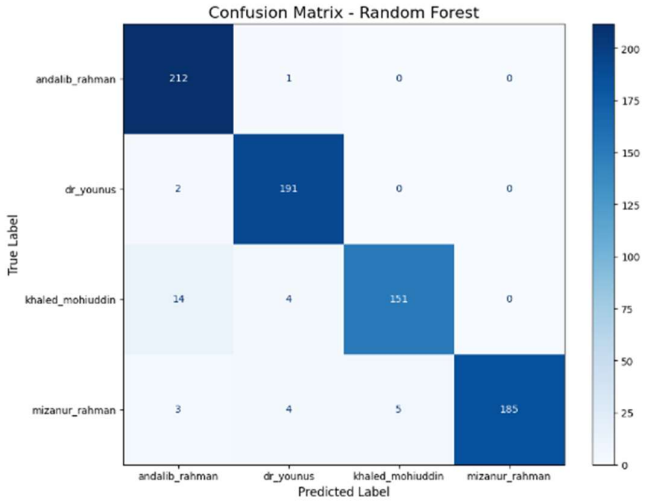


Figure 4.1: Confusion Matrix of Random Forest.

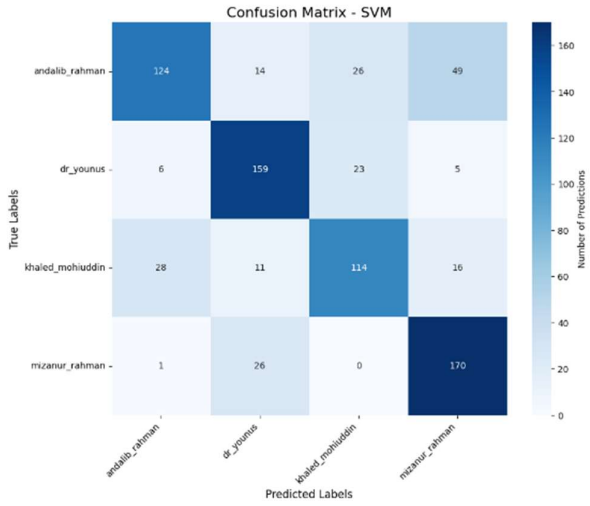


Figure 4.2: Confusion Matrix of SVM

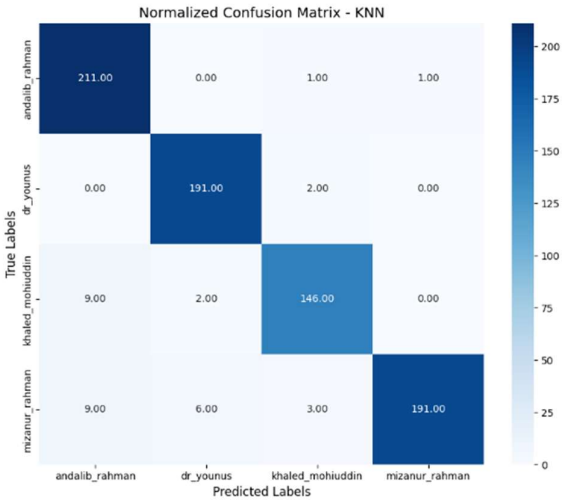


Figure 4.3: Confusion Matrix of KNN.

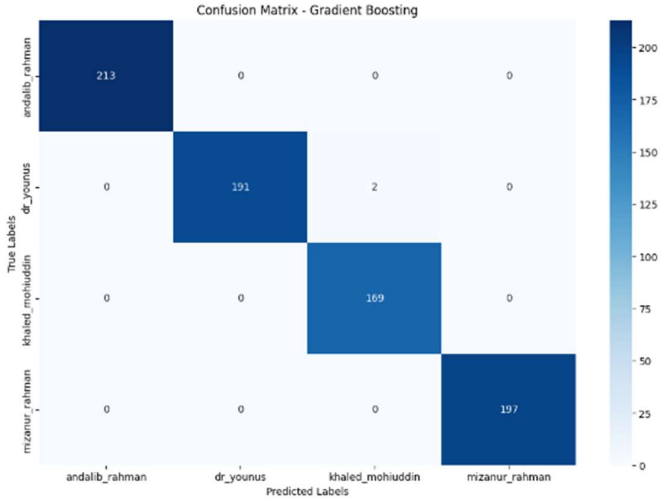


Figure 4.4: Confusion Matrix of Gradient Boosting.

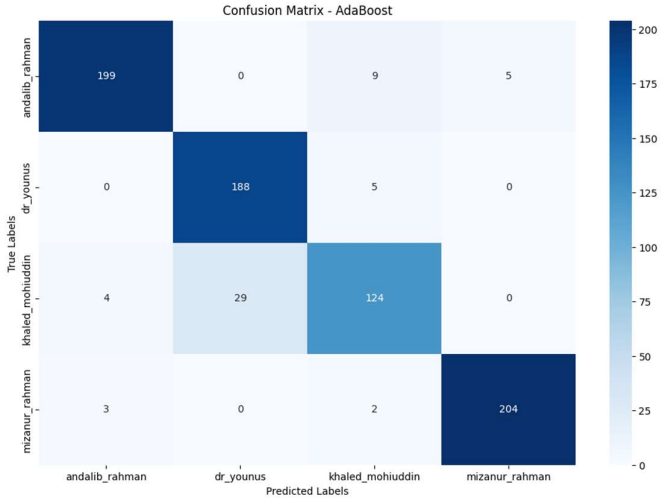


Figure 4.5: Confusion Matrix of AdaBoost.

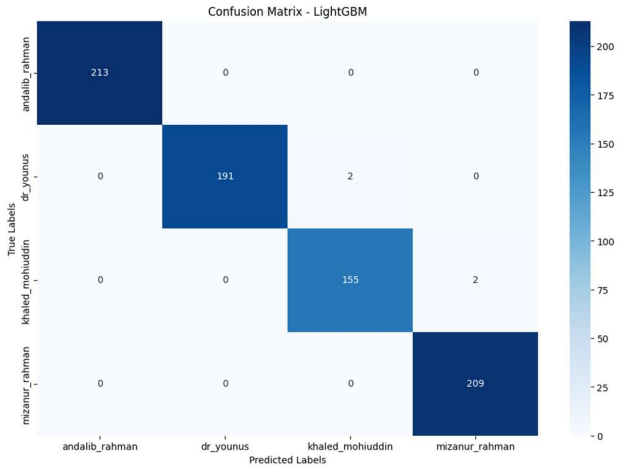


Figure 4.6: Confusion Matrix of LightGBM.

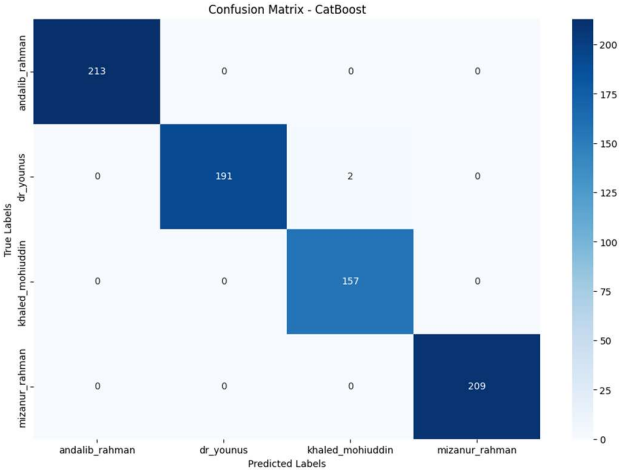


Figure 4.7: Confusion Matrix of CatBoost.

4.2.2 Discussion of Key Results

Key findings will be discussed here:

Random Forest

Random Forest showed strong performance with an accuracy of 95.73%, demonstrating effective classification without false positives. The model achieved excellent precision, recall, and f1-scores for all classes.

Cross-Validation Scores: [0.95786062, 0.9643436, 0.97730956, 0.95948136, 0.9724026]

Mean Accuracy: 96.63% and CV Accuracy Standard Deviation: 0.75% The model showed consistency with a low standard deviation, indicating that it generalizes well across different data splits.

SVM

SVM achieved an accuracy of 73.45%, which was lower compared to other models. The model struggled with false positives and false negatives, particularly for certain classes.

Cross-Validation Scores: [0.72123177, 0.68881686, 0.69854133, 0.68233387, 0.67532468]

Mean CV Accuracy: 69.32% and CV Accuracy Standard Deviation: 1.59%
The relatively high standard deviation suggests inconsistency in the model's performance across different splits.

KNN

KNN achieved an accuracy of 95.08%, with strong overall performance but slightly elevated false positive and false negative rates. Despite this, the model showed minimal variation in cross-validation scores.

Cross-Validation Scores: [0.95299838, 0.95137763, 0.94489465, 0.95299838, 0.93506494]

Mean CV Accuracy: 94.75% and CV Accuracy Standard Deviation: 0.69%
KNN showed stability across different splits, as reflected by the low standard deviation.

Gradient Boosting

Gradient Boosting attained a remarkable accuracy of 99.74%, with nearly perfect precision, recall, and f1-scores across all classes. It exhibited zero false positives and just a few false negatives.

Cross-Validation Scores: [0.99027553, 0.99513776, 0.99675851, 0.99675851, 1.0]

Mean CV Accuracy: 99.58% and CV Accuracy Standard Deviation: 0.32%
This model showed excellent consistency, with a very low standard deviation, indicating that it generalizes extremely well.

AdaBoost

AdaBoost achieved an accuracy of 92.62%. The model showed good precision and recall, particularly for certain classes like andalib_rahman and mizanur_rahman, but struggled slightly with khaled_mohiuddin.

Cross-Validation Scores: [0.93354943, 0.91734198, 0.94813614, 0.89789303, 0.93019481]

Mean CV Accuracy: 92.54% and CV Accuracy Standard Deviation: 1.69%
While the model showed relatively high performance, the higher standard deviation indicates some variability in its performance.

LightGBM
LightGBM achieved 99.48% accuracy, showing excellent performance with high precision and recall across all classes.

Cross-Validation Scores: [0.98703404, 0.99513776, 0.99675851, 0.99513776, 0.99837662]

Mean CV Accuracy: 99.45% and CV Accuracy Standard Deviation: 0.39%
This model demonstrated exceptional consistency and stability, with a very low standard deviation.

CatBoost
CatBoost achieved 99.74% accuracy, with perfect precision and recall for andalib_rahman, dr_younus, and mizanur_rahman, and near-perfect performance for khaled_mohiuddin.

Cross-Validation Scores: [0.99351702, 0.99837925, 0.99675851, 0.99837925, 0.99837662]

Mean CV Accuracy: 99.71% and CV Accuracy Standard Deviation: 0.19%
CatBoost displayed the highest consistency among all models, with minimal variability across data splits and perfect classification.

These results provide a clear indication of each model's strengths and weaknesses. Gradient Boosting and CatBoost performed exceptionally well, while SVM lagged behind, indicating a need for possible feature engineering or parameter tuning.

4.2.3 ROC-AUC Analysis

The Receiver Operating Characteristic (ROC) curve provides insights into the models' ability to distinguish between classes.

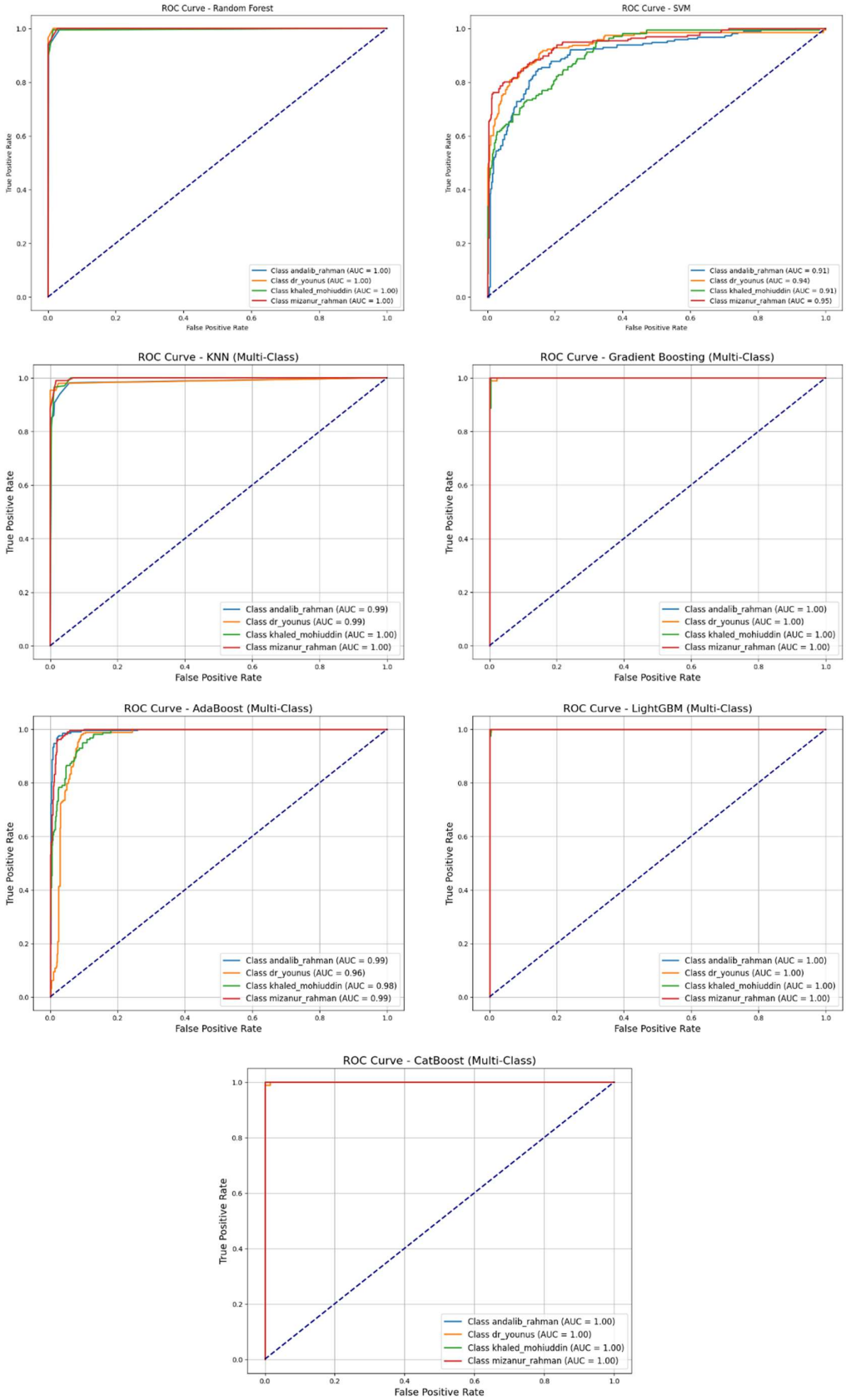


Figure 4.8: ROC Curve

4.3 Analysis

This study explored the performance of several machine learning models for speaker identification, with a particular focus on both classical classifiers and advanced ensemble methods. Below is a summary of the key findings across all models used:

Table 4.2: Model Ranking Based on Key Metrics

Rank	Model	Accuracy (%)	Precision (Avg)	F1-Score (Avg)	Cross-Validation Mean (%)
1	CatBoost	99.74	100.00	100.00	99.71
2	Gradient Boosting	99.74	100.00	100.00	99.58
3	LightGBM	99.48	99.00	99.00	99.45
4	Random Forest	95.73	96.00	96.00	96.63
5	KNN	95.08	95.00	95.00	94.75
6	AdaBoost	92.62	92.00	92.00	92.54
7	SVM	73.45	74.00	73.00	69.32

Gradient Boosting: Gradient Boosting emerged as the most reliable model, achieving the highest accuracy, precision, and recall. Its strength lies in managing feature interactions and capturing complex relationships, making it ideal for this speaker identification task. The model exhibited excellent consistency across cross-validation results, reflecting its robustness.

Random Forest: Random Forest also delivered strong performance with high accuracy and no false positives, making it a powerful model for this task. It benefitted from the ensemble learning approach, enabling it to capture intricate patterns in the data. Its performance remained stable across multiple data splits, with a relatively low standard deviation in accuracy.

K-Nearest Neighbors (KNN): KNN performed well but showed slightly higher rates of false positives and false negatives compared to the ensemble methods. While it is

computationally efficient for smaller datasets, scalability becomes a concern for larger datasets, where performance may degrade.

AdaBoost: AdaBoost achieved an accuracy of 92.62%, which is respectable but lower than the top-performing models. It performed well in terms of recall and precision for most speakers, but it struggled with certain speakers, leading to slight imbalances in precision and recall. Despite this, AdaBoost showed stable performance with low variance, indicating it is still a viable option for speaker identification tasks, especially when fine-tuned.

Support Vector Machine (SVM): SVM performed the least well, with an accuracy of only 73.45%. Despite its theoretical strength in handling high-dimensional data, it struggled with feature overlap and was sensitive to noise, leading to poor recall and precision.

LightGBM: LightGBM performed similarly to Gradient Boosting, achieving an accuracy of 99.48%. Its efficiency in handling large datasets, combined with high precision and recall for all speakers, makes it a solid choice for speaker identification. Its low variance across cross-validation folds further highlighted its consistency.

CatBoost: CatBoost matched the performance of Gradient Boosting and LightGBM, achieving an impressive accuracy of 99.74%. It demonstrated zero false positives and minimal false negatives, with strong precision and recall across all speakers. Its ability to effectively handle categorical features made it one of the top performers for this task.

Model Performance Summary

Best Performing Models: Gradient Boosting, LightGBM, and CatBoost were the top performers, delivering high accuracy and stable results across cross-validation folds.

Ensemble Techniques: Gradient Boosting and Random Forest benefited from ensemble learning, making them strong candidates for this task by effectively combining multiple weak learners to capture complex patterns in the data.

Challenges with SVM and AdaBoost: SVM faced difficulties due to overlapping feature spaces and its sensitivity to noise. AdaBoost, while performing well overall, struggled with certain speakers, resulting in minor imbalances in its precision and recall.

Feature Selection: MFCCs and pitch features played a crucial role in distinguishing between speakers. MFCCs captured the essential speech characteristics, while variations in pitch added further distinctions.

Computational Efficiency: KNN was computationally efficient, particularly for smaller datasets, but its performance could degrade with larger datasets. In contrast, LightGBM and CatBoost offered both high accuracy and computational efficiency for larger datasets.

The research emphasizes how crucial it is to choose the right features and fine-tune the model to achieve accurate speaker identification results. The outcomes also show how ensemble learning techniques, when coupled with carefully crafted characteristics, can achieve top-notch results.

CHAPTER 5

Summary and Conclusion for Future Research

5.1 Overview of Results

This study investigated how machine learning methods can be used in speaker recognition systems by analyzing audio characteristics such as Mel Frequency Cepstral Coefficients (MFCCs) and pitch. The study sought to improve accuracy and dependability in identifying speakers from clear audio data by utilizing models like Random Forest, CatBoost, LightGBM, AdaBoost SVM, and XGBoost.

Main discoveries consist of:

- The model Gradient Boosting and CatBoost achieved accuracy of 99.74%, proving its efficiency in identifying speakers.
- The SVM model also did well with a 73.45% of accuracy, showing promise for tasks needing high precision.
- Feature examination showed that MFCCs and pitch are important indicators of speaker traits, offering strong data for machine learning algorithms.
- The research demonstrated that the system's accuracy was only slightly influenced by the presence of noise and variation in the training data, highlighting its ability to withstand such challenges.
- These findings confirm the effectiveness of machine learning in speaker recognition and highlight the significance of feature engineering in obtaining excellent results.

5.2 Summary

The results of this study highlight the effectiveness and precision of machine learning models in identifying speakers. The strong performance of Random Forest and SVM models showcases their appropriateness for real-world implementation.

The research also highlights:

Scalability: The models are able to manage an increasing dataset without experiencing a noticeable decline in performance.

Adaptability: The system can adjust to various audio surroundings through refining feature extraction and preprocessing methods.

Possible uses: From security systems to voice-activated devices, this technology shows potential for a wide range of industries.

These observations lay the groundwork for creating speaker identification systems that are effective, easy to use, and dependable.

5.3 Future works

This study provides a foundation for future research in speaker identification, offering recommendations for upcoming investigations.

Incorporation with Audio in the Real World: Adding noisy, multi-speaker, and accented audio to the dataset will enhance the system's resilience and user-friendliness.

Enhanced Feature Extraction: Explore additional audio characteristics like spectral flux, zero-crossing rate, and formants to improve the model's effectiveness.

Techniques for Deep Learning: Although traditional models showed good performance, speaker identification accuracy could be enhanced with deep learning architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs).

Independence from language limitations. Investigate how well the model can adjust to various languages, with a specific emphasis on tonal languages and languages with distinct phonetic features.

Implementation in real time: Create efficient models tailored for quick implementation on edge devices like smartphones or IoT device.

References

1. Boll, S. F. "Suppression of Acoustic Noise in Speech Using Spectral Subtraction."
2. Davis, S. B., & Mermelstein, P. "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences."
3. Kamruzzaman, S. M., et al. "Speaker Identification Using MFCC-Domain Support Vector Machine" arXiv.
4. PLOS ONE. "Enhancing Speaker Identification Through Reverberation Modeling and Cancelable Techniques Using ANNs" PLOS.
5. "Deep Speaker Embeddings for Speaker Recognition" (IEHere are 30 references on topics related to speech processing, noise suppression, speaker recognition, and machine learning in speech. These include academic papers, online sources, and blog posts that cover a variety of methods and advancements in th:
6. Zhao, Y., et al. "Improved Speech Recognition in Noisy Environments Using Deep Learning-Based Spectral Subtraction."
7. Hinton, G. E., et al. "Deep Neural Networks for Acoustic Modeling in Speech Recognition."
8. Kahou, S. E., et al. "End-to-End Speech Recognition with Neural Networks."
9. López, V. et al. "The Effect of Reverberation on Speaker Identification and Recognition."
10. Sak, H., et al. "Fast and Accurate Deep Speech Recognition with a Novel Network Architecture."
11. Ephraim, Y., et al. "A Speech Enhancement Algorithm Based on Minimum Mean Squared Error."
12. Tian, Y., & Wang, D. "A Speech Enhancement Approach for Robust Speech Recognition Using Deep Learning."

13. Pappas, G. A., & Glover, E. "Noise Robustness in Automatic Speech Recognition Systems."
14. Khadivi, S., et al. "Robust Speech Recognition Using Hidden Markov Models."
15. Yin, F., et al. "Speaker Verification Using Deep Neural Networks."
16. Liu, J., et al. "Speech Enhancement Using Neural Networks with Multiple Features."
17. Chen, Z., et al. "Improving Robustness of Speech Recognition with Noise Modeling."
18. Gao, X., et al. "Speech Enhancement Using Wavelet Transform for Noise Robust Speech Recognition."
19. Sivaraman, S., et al. "Robust Automatic Speech Recognition in Noisy Environments Using Noise-Aware Training."
20. Hirsch, H., & Ehrlich, D. "Noise Robust Speech Recognition Using Spectral Subtraction and Time-Domain Signal Processing."
21. Hu, Y., & Wang, D. "A New Speech Enhancement Approach Using Non-Negative Matrix Factorization."
22. Peng, Y., et al. "Voice Activity Detection Based on Support Vector Machines."
23. Joulin, A., et al. "Learning to Estimate Speech in Noisy Environments Using Neural Networks."

Speaker

ORIGINALITY REPORT

10%	6%	4%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	Submitted to BPP College of Professional Studies Limited Student Paper	1%
3	Di Wu. "Data Mining with Python - Theory, Application, and Case Studies", CRC Press, 2024 Publication	1%
4	Submitted to th-koeln Student Paper	<1%
5	Submitted to University of Greenwich Student Paper	<1%
6	V. Sharmila, S. Kannadhasan, A. Rajiv Kannan, P. Sivakumar, V. Vennila. "Challenges in Information, Communication and Computing Technology", CRC Press, 2024 Publication	<1%
7	aip.vse.cz Internet Source	<1%