DETECTION OF GENDER FROM VOICE USING DATA ANALYTICS

A PROJECT REPORT

Submitted by

ARAVINDHAN G	**(190501017)**
ARJUNRAJ N	**(190501019)**
HAREE J	**(190501040)**

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

SRI VENKATESWARA COLLEGE OF ENGINEERING

(An Autonomous Institution; Affiliated to Anna University, Chennai-600025)

ANNA UNIVERSITY :: CHENNAI 600 025

DECEMBER 2022

SRI VENKATESWARA COLLEGE OF ENGINEERING

(An Autonomous Institution; Affiliated to Anna University, Chennai-600025)

ANNA UNIVERSITY, CHENNAI - 600 025

BONAFIDE CERTIFICATE

Certified that this project report **"DETECTION OF GENDER FROM VOICE USING DATA ANALYTICS"** is the bonafide work of **"ARAVINDHAN G (190501017), ARJUNRAJ N (190501019) and HAREE J (190501040)"** who carried out the project work under my supervision.

SIGNATURE	**SIGNATURE**
Dr. R. ANITHA	**Dr. M. S. GIRIJA**
HEAD OF THE DEPARTMENT	**SUPERVISOR**
	ASSISTANT PROFESSOR
COMPUTER SCIENCE & ENGG	**COMPUTER SCIENCE & ENGG**

Submitted for the project viva-voce examination held on ………………..

INTERNAL EXAMINER	**EXTERNAL EXAMINER**

ABSTRACT

The idea is to develop a Machine Learning Model on detection of Gender from voice using Data Analytics. Gender detection from the speech are essential in machine and human interaction. Sometimes it is required to predict age, gender, and emotion from audio clips for investigation purposes. Gender detection from the speech are essential in machine and human interaction. The audio can be analyzed, and the generated prediction models for age, gender, and emotion can be used to predict customer demography. In predicting gender, CatBoost performs best among all predictive models with 96.4% test accuracy. On the other hand, Random Forest performs best for predicting age among all predicting models with 70.4% test accuracy. For emotion prediction, XGBoost performs best with 66.1% test accuracy. Gender identification by voice is useful in speech-based recognition systems which employ gender-dependent models. We have created a web interface where users can record and upload their voice samples for prediction and have displayed the results in each of the learning models used. Gender differentiation help to improve automatic emotion recognition from speech. Classifying speaker's gender is an important task in the context. From our project we have observed that Random forest decision tree and Support Vector Machine predict the results with best accuracy scores.

ACKNOWLEDGEMENT

We thank our Principal **Dr. S. Ganesh Vaidyanathan,** Sri Venkateswara College of Engineering for being the source of inspiration throughout our study in this college.

We express our sincere thanks to **Dr. R. Anitha,** Head of the Department, Computer Science and Engineering for her encouragement accorded to carry this project.

With profound respect, we express our deep sense of gratitude and sincere thanks to our guide **Dr. M. S. Girija**, **Assistant Professor**, for her valuable guidance and suggestions throughout this project.

We are also thankful to our project coordinators **Dr. R. Jayabhaduri, Professor, Dr. N. Revathi, Associate Professor and Ms. V. Rajalakshmi, Associate Professor** for their continual support and assistance.

We thank our family and friends for their support and encouragement throughout the course of our graduate studies.

ARAVINDHAN G
ARJUNRAJ N
HAREE J

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	**iii**
	LIST OF FIGURES	**vii**
	LIST OF ABBREVIATION	**ix**
1	**INTRODUCTION**	**1**
`	1.1 MACHINE LEARNING	1
	1.2 DEFINITONS	2
	1.3 DATA ANALYTICS	4
	1.4 EXISTING PROBLEM	6
	1.5 PROBLEM STATEMENT	7
2	**LITERATURE REVIEW**	**9**
3	**PROPOSED WORK**	**16**
	3.1 OBJECTIVE	16
	3.2 EXISTING SYSTEM	16
	3.3 PROPOSED ARCHITECTURE	17
4	**REQUIREMENT SPECIFICATION**	**22**
	4.1 FUNCTIONAL REQUIREMENT	22
	4.2 NON-FUNCTIONAL REQUIREMENT	22
	4.3 HARDWARE REQUIREMENT	23
	4.4 SOFTWARE REQUIREMENT	23
5	**IMPLEMENTATION MODULES**	**24**

	5.1 DATA VISUALISATION	24
	5.2 DATA PRE-PROCESSING	28
	5.3 MODULES OF THE PROJECT	28
6	**RESULTS AND DISCUSSIONS**	**30**
	6.1 SNAPSHOTS AND MODULES	30
7	**CONCLUSION AND FUTURE WORK**	**40**
	REFERENCE	**42**

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE NO.
3.3	Proposed Architecture	**17**
3.3.1	DFD-Level 1	**18**
3.3.2	DFD-Level 2	**18**
3.3.3	DFD-Level 3	**19**
3.3.4	Class Diagram	**20**
3.3.5	Activity Diagram	**21**
5.1.1	Male Dataset	**25**
5.1.2	Male Dataset	**26**
5.1.3	Female Dataset	**26**
5.1.4	Visualization Diagram	**27**
5.2	Data Pre-Processing	**28**
6.1.1	Feature Importance	**30**
6.1.2	Prediction Accuracy Age	**30**
6.1.3	Prediction Accuracy Emotion and Gender	**31**
6.1.4	Audio file stored in the Database	**32**
6.1.5	Home Page of the web interface	**33**

6.1.6	Audio Upload Page, View-1	**34**
6.1.7	Audio Upload Page, View-2	**35**
6.1.8	System Training Page	**36**
6.1.9	Testing and Result Page, View-1	**37**
6.1.10	Testing and Result Page, View-2	**38**
6.1.11	System Detail	**39**

LIST OF ABBREVIATION

ML	Machine Learning
DA	Data Analytic
SVM	Support Vector Machine
KNN	k-Nearest-Neighbor Classifier
WAV	Waveform Audio File
WMA	Windows Media Audio
ASR	Automated Speech Recognition
VUI	Voice User Interface
DFD	Data Flow Diagrams
CSV	Comma Seperated Values
XGBoost	eXtreme Gradient Boosting

CHAPTER 1

INTRODUCTION

1.1 MACHINE LEARNING

Machine learning is the science of getting computers to act without being explicitly programmed and it is a subset of Artificial Intelligence involving the application of computer algorithms that employs computation processes that can improve automatically through experience by the use of data. The primary intention of machine learning is to allow machines to learn autonomously without any human intervention or assistance and adjust its functions accordingly to them. Machine learning has given us self- driving cars, practical speech recognition, effective web search, and a vastlyimproved understanding of the human genome. Machine learning is sopervasive today that you probably use it dozens of times a day without knowing it. Many researchers also think it is the best way to make progress towards human-level AI.

Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs. Any technology user today has benefitted from machine learning. Facial recognition technology allows social media platforms to help users tag and share photos of friends. Recommendation engines, powered by machine learning, suggest what movies or television shows to watch next based on user preferences. Speech recognition is the ability of a machine or program to identify words and phrases in spoken language and convert them to a machine-readable format.

Speech Recognition Algorithm works with human sound in a normal environment. Technically, this environment is referred to as an analogenvironment. A computer can’t work with analog data; it needs digital data. This is why the first piece of equipment needed is an analog to digital converter.

1.2 DEFINITIONS

Machine Learning: It is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

Gender Identity: It is defined as a personal conception of oneself as male or female (or rarely others). This concept is intimately related to the concept of gender role, which is defined as the outward manifestations of personality that reflect the gender identity. Gender identity, in nearly all instances, is self-identified, as a result of a combination of inherent and extrinsic or environmental factors.

Age Identity: It is defined as a personal conception of oneself as young or matured. This concept is intimately related to the concept of age role, which is defined as the outward manifestations of personality that reflect the age identity.

Emotion Identity: It is defined as a personal conception of oneself as happy or sad. This concept is intimately related to the concept of emotion role, which is defined as the outward manifestations of personality that reflect the emotion identity.

Audio Features: An abstract representation of pieces of digital music. Audio features are computed from the raw audio signal. Simple features are the number of zero crossings of the audio signal or its centroid. More sophisticated approaches, such as MP3-based features, rhythm Patterns, Rhythm Histograms, or statistical Spectrum Descriptors take into account, for instance, findings from psycho-acoustics.

Naïve Bayes Classifier: These are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naïve) independence assumptions between the features. They are among the simplest Bayesian network models.

Support Vector Machine (SVM): It is a machine learning algorithm that analyzes data for classification and regression analysis. SVM is a supervised learning method that looks at data and sorts it into one of two categories. An SVM outputs a map of the sorted data with the margins between the two as far apart as possible.

K-Nearest-Neighbor Classifier (KNN): It is an approach to data classification that estimates how likely a data point is to be a member of one group or the other depending on what group the data points nearest to it are in.

Random Forest Classifiers or Random Decision Forests: These are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees

WarbleR: It is a package designed to streamline analysis of animal acoustic signals in R. This package allows users to collect open-access avian vocalizations data or input their own data into a workflow that facilitates spectrographic visualization and measurement of acoustic parameters. warbler makes fundamental sound analysis tools from the R package see wave, as well as new tools not yet offered in the R environment.

1.3 DATA ANALYTICS

Data analytics (DA) is the process of examining data sets in order to find trends and draw conclusions about the information they contain. Increasingly, data analytics is done with the aid of specialized systems and software. Data analytics technologies and techniques are widely used in commercial industries to enable organizations to make more-informed business decisions. Scientists and researchers also use analytics tools to verify or disprove scientific models, theories and hypotheses.

Data analytics can do much more than point out bottlenecks in production. Gaming companies use data analytics to set reward schedules for players that keep the majority of players active in the game. Content companies use many of the same data analytics to keep you clicking, watching, or re-organizing content to get another view or another click.

Data analytics initiatives can help businesses increase revenue, improve operational efficiency, optimize marketing campaigns and bolster customer service efforts. Analytics also enable organizations to respond quickly to emerging market trends and gain a competitive edge over business rivals. The ultimate goal of data analytics, however, is boosting business performance. Depending on the particular application, the data that's analyzed can consist of either historical records or new information that has been processed for real-time analytics. In addition, it can come from a mix of internal systems and external data sources.

Using the concept of big data, huge chunks of data is collected, integrated and stored for future use. This data can be sorted or unsorted, structured or unstructured dependingon the type of collection and need. Audio too is collected similarly. A huge amount of audio files and clips are collected and stored. The various ways in which audio is storedon machines and electronic devices are —

- **mp3 (MPEG-1 Audio Layer 3) —** It is a coding format for digital audio
- **wav (Waveform Audio File) —** Developed by Microsoft and IBM, it is an audio file format standard to store audio bit streams on PCs.
- **WMA (Windows Media Audio) —** Developed by Microsoft, it is a sequence of audio codes and their resultant audio coding formats.

Data Analysis Steps

The process involved in data analysis involves several different steps:

1. The first step is to determine the data requirements or how the data is grouped. Data may be separated by age, demographic, income, or gender. Data values may be numerical or be divided by category.

2. The second step in data analytics is the process of collecting it. This can be done through a variety of sources such as computers, online sources, cameras, environmental sources, or through personnel.

3. Once the data is collected, it must be organized so it can be analyzed. This may take place on a spreadsheet or other form of software that can take statistical data.

4. The data is then cleaned up before analysis. This means it is scrubbed and checked to ensure there is no duplication or error, and that it is not incomplete. This step helps correct any errors before it goes on to a data analyst to be analyzed.

1.4 EXISTING PROBLEM

During the last few years, speech recognition has improved a lot. This can mainly be attributed to the rise of graphics processing and cloud computing, as these have made large data sets widely distributable. During the last few years, speech recognition has improved a lot. This can mainly be attributed to the rise of graphics processing and cloud computing, as these have made large data sets widely distributable. The current challenges of speech recognition are caused by two major factors – reach and loud environments. This calls for even more precise systems that can tackle the most ambitious ASR use-cases. Think about live interviews, speech recognition at a loud family dinner or meetings with various people. These are the upcoming challenges to be solved for next-gen voice recognition. Speech recognition software isn't always able to interpret spoken words correctly. This is due to computers not being on par with humans in understanding the contextual relation of words and sentences, causing misinterpretations of what the speaker meant to say or achieve.

VUIs are oftentimes challenged when voice inputs divert too much from the average. Especially accents can pose a big challenge. While systems are getting better there's still a big difference in their ability to understand American or Scottish English for example. Even a simple cold can be a reason for voice commands not to work as well as usual.

To make the most of VUIs a quiet environment helps a lot. Whenever there is too much background noise speech recognition will be challenged. Making it especially hard to use them effectively in the urban outdoors or large public spaces/offices. With the use of specific microphones or headsets, the limitations can be decreased but it requires an additional device, which is never desirable.

Therefore, a great challenge of voice recognition lies in making data input available for AI, but still acknowledge the need for data privacy and security.

1.5 PROBLEM STATEMENT

One of the vital speech analysis applications is to predict gender, age, and emotion from speech. Detecting gender from the speech will help a lot to enhance human-machine interaction. Audios can be classified by age and gender after forecasting properly. Age, gender, and emotion can significantly help the relevant investigations. These predictions can be beneficial to most tele- communication companies. The audio calls can be analyzed, and the generated prediction models for age, gender, and emotion can be used to predict customer demography. They can recommend offers based on that. Several researchers have focused on detecting gender, age, and emotion from different types of sources. But according to the best of our knowledge, none of them use a single type of source to detect all of them.

Gender detection from the speech are essential in machine and human interaction. Sometimes it is required to categorize audios by age and gender from speech. Sometimes it is required to predict age, gender, and emotion from audio clips for investigation purposes.

We aim to reduce this barrier via our project, which was designed and developed to achieve systems in particular cases to provide significant help so people can share information by operating a computer using voice input as Voice recognition and the classification of the age, gender and emotion using voice are the interesting field of research in this area along with a friendly user interface for gender identification.

This project keeps that factor in mind, and an effort is made to ensure our project is able to recognize speech and convert input audiointo text;especially hard to use them effectively in the urban outdoors or large public spaces/offices. it also enables a user to perform file operations like Save, Open, or Exit fromvoice-only input. We aim to design a friendly user interface for gender recognition by identifying the gender of the speaker from various machine learning algorithms and over come the gender bias in case of voice sample containing sounds like crying or yelling which is the drawback of the existing system.

CHAPTER 2

LITERATURE REVIEW

Bhagya Vijayan et al. (2019) used Categorical emotion recognition with Acoustic and Pitch related features. They investigated Gaussian Mixture Models (GMM), Multi-Layer Perceptron (MLP), and Decision Tree Classifiers. A combination of Piece wise Gaussian Modelling features and pitch-related features with a set of Neural Networks was shown toperform better than any individual classifier.

N. V. Chawla et al. (2002) utilized to extract audio features out of sentence recordings. He trained on Naïve Bayes, Discriminant Analysis (DA), Support Vector Machine (SVM) with Linear kernel, k-Nearest Neighbor (KNN) and Classification Tree (CT) classifiers. The DA classifier is most performance in terms of test error rate and precision. The high bias problem implies the set of all available features he is considering does not capture enough gender specific characteristics of voice. Logistic Regression Linear Regression Random Forest and AdaBoost available from Python's scikit-learn library, on summary statistics, Random forest yields the best model, achieving an accuracy of 85.0% when using all features,83.3% with onlyf0 features, and 84.1% when using the f0 + energy + voice quality feature set. And also used a Support Vector Machine (SVM) based gender identification is applied on discriminative weight training. The performances of gender classification system have been evaluated on the conditions of clean speech, with gender classification accuracy of at most 98% and remains 95% for most noisy speech.

T Jayashankar et al. (2017) trained model using Support Vector Machine (SVM), Decision Trees, Gradient Tree Boosting, Random forests and accuracy is calculated. Some unexpected behavior is in the peaks at very low frequencies (<50 Hz). This can be due to the presence of noise in the audio recordings. The Random Forest gave the best accuracy followed by the Gradient Boosting SVM and Decision Tree to produces best outcomes with accuries.

Hadi Harb and Liming Chen et al. (2005) used five different algorithms. They are Linear Discriminant Analysis (LDA), k-NearestNeighbor (KNN), Classification and Regression Trees (CART), Random Forest and Support Vector Machine (SVM) on basis of eight different metrics. The result shows that SVM algorithm performs better in classification and with reduced error rate and conducted a study to predict gender from speech using the Hidden Markov Model (HMM) and the Support vector machine (SVM). The accuracy of the predictive model using HMM is 48.1% to 70.7%, and the accuracy using SVM is 53.1% to 72.6%. Similarly, a study to predict gender from speech using Gaussian Mixture Modelling (GMM). The predictive model's accuracy is 95% for children, 99% for males, and 98% for females. A study used a statistical approach to analyze audio and detect gender using machine learning concluded with an accuracy of 97% [in SVM training 96.6% and testing 97%] by using a dataset from Kaggle. Similarly, conducted a study to predict gender from speech using a set of neural networks as classifiers. The accuracy of the predictive model is 91.72%. Also, conducted a study to predict gender from speech using GMM, Multilayer Perceptron (MLP), Vector Quantization (VQ), and Learning Vector Quantization (LVQ). The accuracy of the predictive model is 96.4% using the IViE corpus dataset. Similarly, conducted a study to predict gender from speech using Weighted Supervised Non negative Matrix Factorization (WSNMF) and age with Generalized Regression Neural Network (GRNN). The accuracy of the predictive model 96% for gender using the Dutch database. And also a study to predict gender from Pitch's speech using Auto-Correlation, Signal Energy, Mel Frequency Cepstral Coefficients (MFCC), and SVM classifiers. The predictive model's accuracy is 69.23% for MFCC, 57.14% for Pitch, 55.81% for Energy using the TIMIT data set. Also a study to predict gender from speech using Deep Neural Networks. The predictive model's accuracy for gender with age is 90.56% to 92.72% using 17,408 real-traffic and Microsoft spoken dialogue system.

K. Vinothkumar et al. (2017) conducted a study to predict age from speech using HMM and SVM. The accuracy of the predictive model using HMM is 48.1% to 70.7%, and the accuracy using SVM is 53.1% to 72.6%. Also conducted a study to predict age from speech using GMM to a study to predict age with GRNN and conducted a study to predict age from speech using Deep Neural Networks. The predictive model's accuracy for gender with age is 90.56%to 92.72% using 17408 real-traffic Mandarin utterances collected from a Microsoft spoken dialogue system. Age has been estimated using voice or from auditory information using different types of learning machines. The speech signal also contains age information and can be used for this purpose. Age estimation from speech recordings has been approached several times in the last two decades. In speech signals sampled at 16kHz, and PCM coded with 16 bits were used to calculate MFCC, ΔΔMFCC and Power coefficients, to be applied to Linear Discriminant Analyzers (LDA), and Neural Networks to estimate speakers' age.

S. Jadav et al. (2018) conducted a study to predict emotion from speech using HMM and SVM. The accuracy of the predictive model using HMM is 48.1% to 70.7%, and the accuracy using SVM is 53.1% to 72.6%. The accuracy of the predictive model is 74%by using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset. Also conducted a study to predict emotion from speech using Deep Neural Networks. The accuracy of the predictive model for emotion is 59.40% to 63.20% using 17408 real-traffic Mandarin utterances collected from a Microsoft spoken dialogue system. Several studies have been conducted to predict gender, age, and emotion from speech using different approaches. The frequency spectrum analysis of speech can predict gender, age, and emotion. There is no study in the literature to compare all the predictive models using supervised machine learning algorithms to measure comparative performance. In most studies, all the approaches followed for predicting gender, age, and emotion from the speech are not the same. Our study focused on following a single approach of predicting gender, age, and emotion using machine learning by extracting statistical features from frequency spectrum analysis of speech and getting the best predictive model among all models through a comprehensive evaluation.

Remna R. Nair et al. (2019) made an excessive investigation has been completed to recognize emotions by using speech statistics. Also proposed a ranking SVM method for synthesize information about emotion recognition to solve the problem of binary classification. This ranking method, instruct SVM algorithms for particular emotions, treating data from every utter as a distinct query then mixed all predictions from rankers to apply multi-class prediction. Ranking SVM achieves two advantage, first, for training and testing steps in speaker- independent it obtains speaker specific data. Second, it considers theintuition that each speaker may express mixed of emotion to recognize the dominant emotion. Ranking approaches achieves substantial gain in terms of accuracy compare to conventional SVM in two public datasets of acted emotional speech, Berlin and LDC. In both acted data and the spontaneous data, which comprises neutral intense emotional utterances, ranking-based SVM achieved higher 020105-2 accuracy in recognizing emotional utterances than conventional SVM methods. Unweight average (UA) or Balance accuracy achieved 44.4%.

Taniya Mishra et al. (2016) described the enhancement of ASR from the perspective of the speaker's problems and the microphone that captures the speaker's voice. Usually, the results of the ASR are good when the training and testing data are matched. However, the results are much worse when they differ in the number and arrangement of microphones. Also suggested an unsupervised spatial clustering approach to microphone array processing. This approach, known as Model-based EM Source Separation and Localization (MESSL). While using MESSL's outputs for spatial covariance estimates of the noise improved ASR performance compared to a standard baseline. Also proposed a method that used multichannel non negative matrix factorization (MNMF) to estimate the spatial covariance matrix (SCM) of speech and noise in an unsupervised manner and generated an enhanced speech signal with beam forming. They found that the proposed methods were more robust in an unknown environment than the state-of-the-art beam forming method with DNN-based mask estimation. Moreover, proposed a target speaker extraction network (TEnet) to isolate the speech of a specific speaker. They relied on the auxiliary speaker characteristics provided by an anchor (a clean audio sample of the target speaker). They demonstrated that the proposed TEnet can outperform the single short anchor baseline by about 22.5% on WER and poor- quality microphone installed on the sensors. Also provided an adapting DNN-based 15.5% on the SDR. Sensitivity to recording conditions can be caused by a high level of background noise and a mediocre or poor-quality microphone installed on the sensors and provided an adapting DNN-based acoustic model. They used an audio database recorded by wireless sensors to train an accurate model for the actual speech processing application. They found that joint training was not significantly better than training on the sensor-recorded noisy database subset, while the DNN adaptation turned out to perform significantly better. In the authors provided an ASR system that employs various methods to address noisy acoustic scenes in public environments using an NMF with VB technique to separate the target speaker's voice from background sources and a time-varying minimum variance distortion less response (MVDR) to detect failure in the microphone channel. They use the AMFB that implicates prior information of speech to analyze its temporal dynamics. The proposed system achieved an absolute WER of 5.67% on the real evaluation test data. Also, in the authors suggested extending an existing attention-based encoder-decoder framework to address the challenging noisy ASR tasks using a neural beam former. In addition, they proposed an architecture of multiple channels in end-to-end ASR that allows the deduction of recognizing multichannel speech toenhance it based on an ASR objective. Their comprehensive framework works effectively with a noisy background. They found that the suggested framework results exceeded the end-to-end baseline with noisy input. Furthermore, successful learning was achieved by the beam former of the noisy suppression. To improve the prediction accuracy of speakers, the authors of [1] proposed a hybrid method for automatic speaker identification using an ANN. The recognition is performed using Bayesian regularization and MLPs. The features are extracted using the mel frequency cepstral coefficient (MFCC). They found that the proposed method provides the best discrimination and has a high accuracy of 93.33%.Speech overlapping means that several people are talking at the same time. Researchers observed a vital degradation in the performance of ASR systems when speech contained cross-talk. A few recent articles have addressed speech overlapping in ASR. In [9], the authors suggested a target speaker extraction network (TEnet) that identifies and isolates a specific speaker's speech based on a clean voice sample of the speaker to address the problem of multiple people speaking at the same time. They concluded that the proposed method has a high performance with a word error rate (WER) of 22.5% and signal-to-distortion rate (SDR) of 15.5%.

V. Michel et al.(2011) proposed a model that divides the interfering speech recognition problem in one channel into three parts: translation, speaker tracking, and speech recognition. They find that it improves by 30% the rate of speech errors. In combined approaches to address the cross-talk problem called deep clustering (DPCL) by creating a hybrid acoustic model. They obtained a WER of 16.5% on the wsj0- 2mix data set, which is the best performance reported so far. However, one of the challenges of automatic speech recognition is identifying children's speech in bilateral interactions because children have weaker communication ability.

A. Raahul et al. (2017) suggested two methods: semantic response generation and lexical repetition. They concluded that it improved children's speech recognition and was applicable. In addition, in to address the problems of integrating multimedia features and frame alignment between two data streams, the authors proposed Wave Net with a mutual interest for automatic voice and visual speech recognition. It improved performance as it reduced Tibetan singular syllable error by 4.5% and 39.8% on English word error in speech. To improve the prediction accuracy of speakers, the authors of proposed a hybrid method for automatic speaker identification using an ANN. They found that the proposed methods were more robust in an unknown environment than the state-of-the-art beam forming method with DNN-based mask estimation and also predict gender from speech using a set of neural networks as classifiers. The accuracy of the predictive model is 91.72%. Also, conducted a study to predict gender from speech using GMM Moreover, proposed a target speaker extraction network (TEnet) to isolate the speech of a specific speaker and The recognition is performed using Bayesian regularization and MLPs. The features are extracted using the mel frequency cepstral coefficient. The high bias problem implies the set of all available features he is considering does not capture enough gender specific characteristics of voice.

CHAPTER 3

PROPOSED WORK

3.1 OBJECTIVE

Gender identification by voice is useful in speech-based recognition systems which employ dependent models. A friendly user interface for Gender differentiation help to improve automatic emotion recognition from speech. Classifying speaker's gender is an important task in the context of multimedia indexing. Gender identification can improve the prediction of other speaker traits such as age and emotion, either by jointly modelling gender with age (or emotion) or in a pipe lined manner. Automatic gender detection also useful in some cases of a mobile healthcare system i.e., there are some pathologies, such as vocal fold cyst. For detecting feeling like male sad, female anger, etc. Differentiating audios and videos using tags. Spontaneous salutations. Helping personal assistants to answer questions with gender-specific results etc.

To create a learning model based on the data set, and train it using various machine learning algorithms. To create a web interface where user can upload their voice samples. As a result, to display the prediction with the results in each of the learning models Used.

3.2 EXISTING SYSTEM

Algorithms used for classifying:

- Linear Discriminate Analysis
- K-Nearest Neighbour
- Classification and Regression Trees
- Random Forest

- Support Vector Machine
- Naive Bayes
- Gradient Tree Boosting and Mel-frequency cepstral coefficients summary statistics to provide the output.

3.3 PROPOSED ARCHITECTURE

System architecture in Figure 3.3 include the representation, connection and arrangement of components that are used in the project.

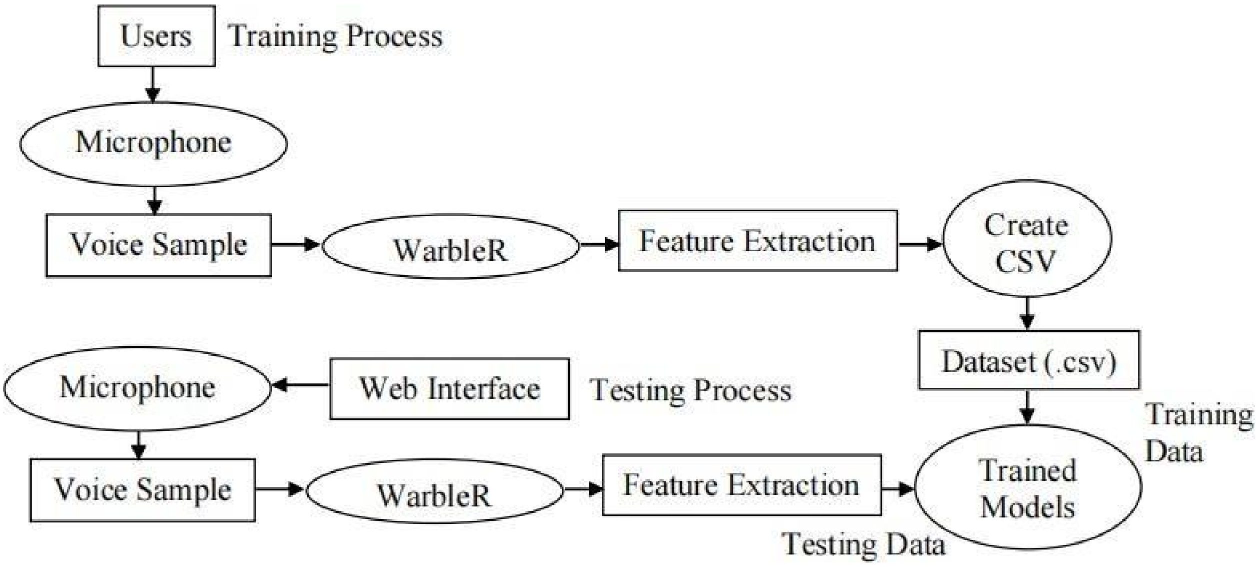


Figure 3.3 Proposed Architecture

3.3.1 DATA FLOW DIAGRAMS

A DFD in Figure 3.3.1, Figure 3.3.2, Figure 3.3.3 is a logical model of the system. The model does not depend on the hardware software and data-structures of file organization. Ittends to be easy for a given non-technical users to understand and thus serves as anexcellent communication tool. A DFD is a logical model of the system. The model does not depend on the hardware software and data structures of file organization. It tends to be easy for a given non-technical users to understand and thus serves as an excellent communication tool.

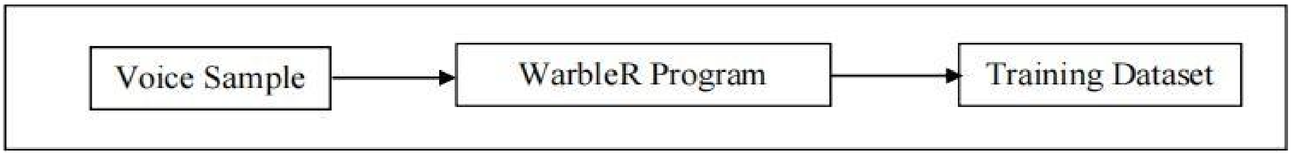


Figure 3.3.1 DFD-Level 1

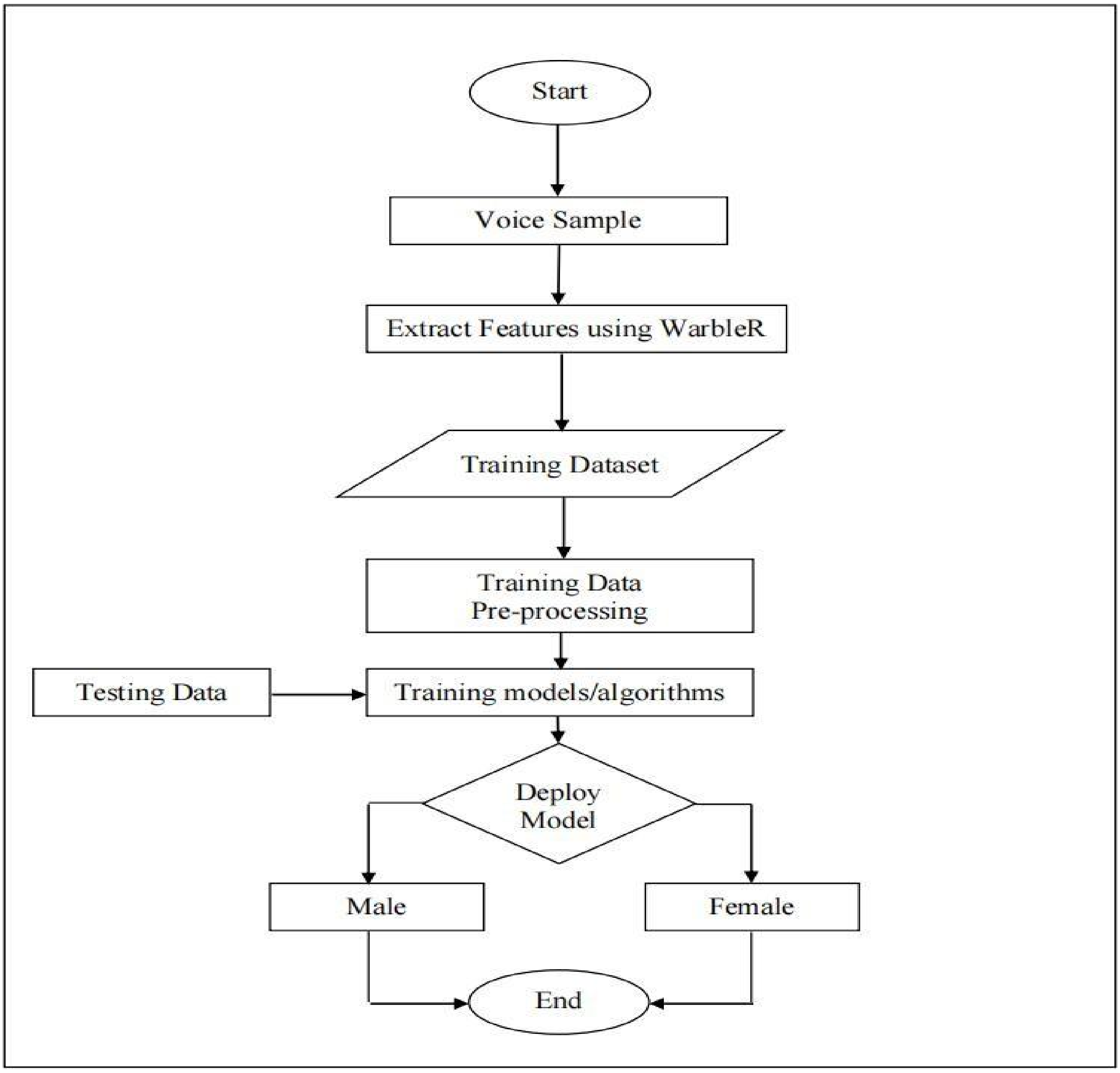


Figure 3.3.2 DFD-Level 2

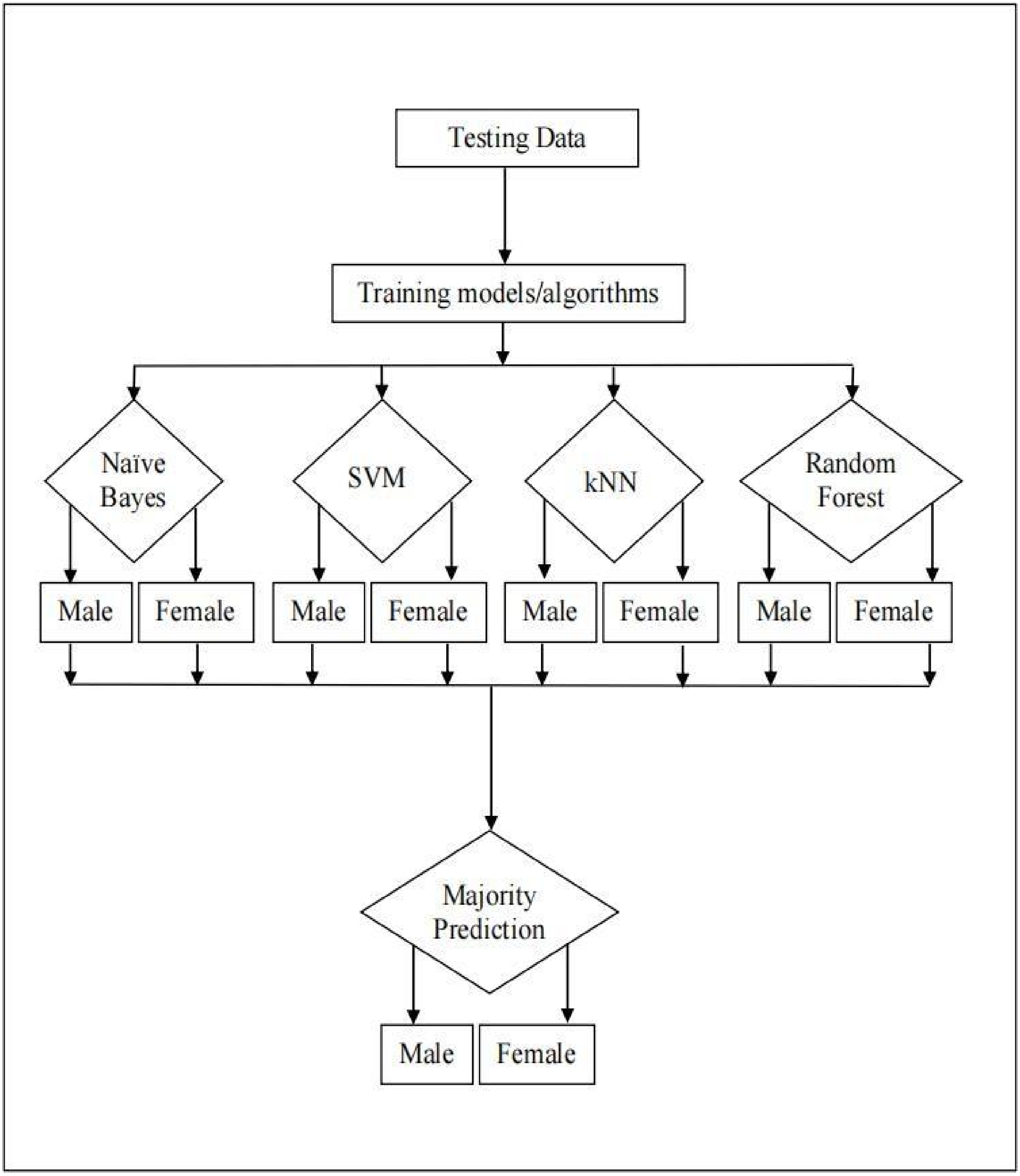


Figure 3.3.3 DFD-Level 3

3.3.4 CLASS DIAGRAMS

Class diagram in Figure 3.3.4 is a type of static structure diagram that describes thestructure of a system and the relationship of the objects.

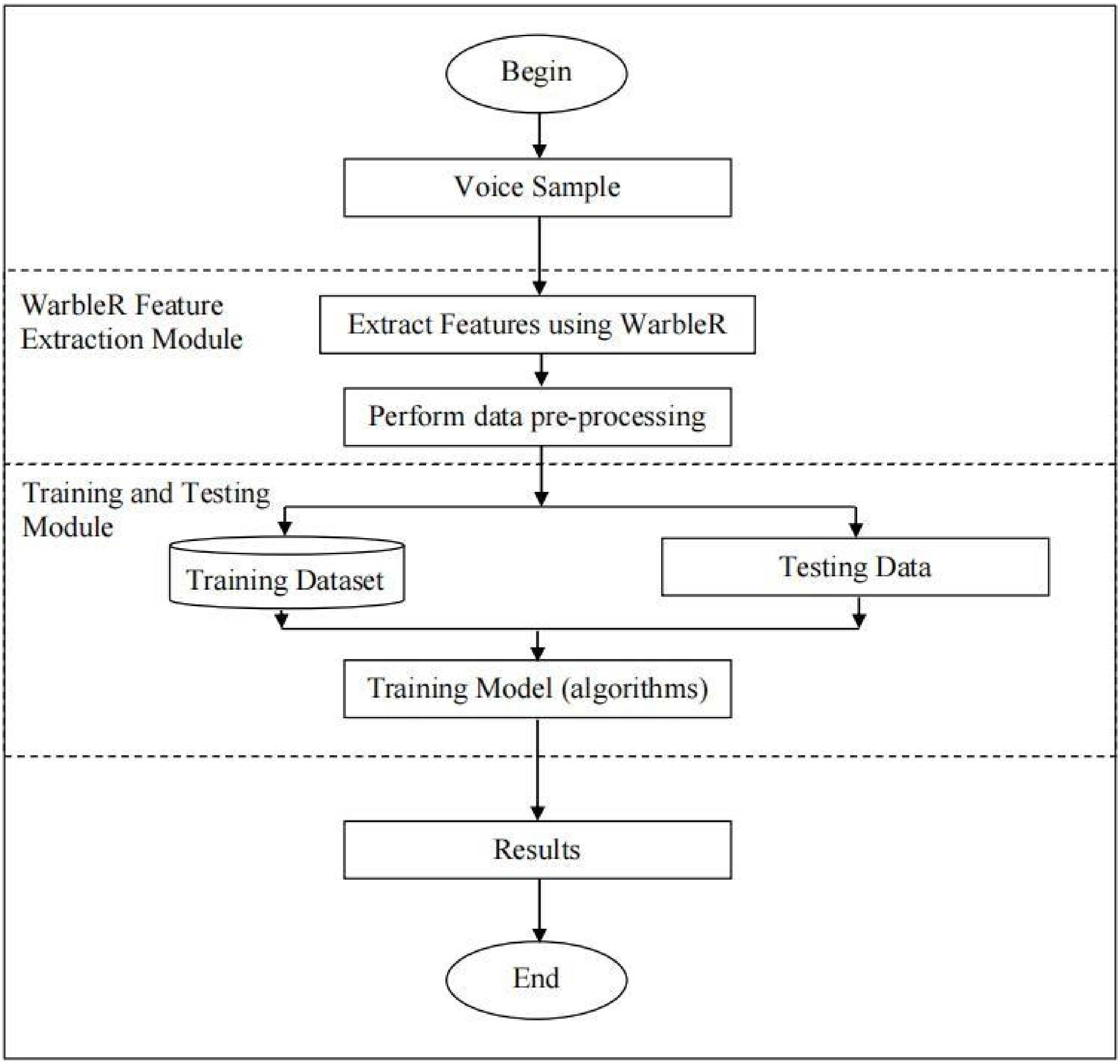

Begin
Voice Sample
WarbleR Feature Extraction Module
Extract Features using WarbleR
Perform data pre-processing
Training and Testing Module
Training Dataset
Testing Data
Training Model (algorithms)
Results
End

Figure 3.3.4 Class Diagram

3.3.5 ACTIVITY DIAGRAMS

Activity diagram in Figure 3.3.5 is a representation of workflows of step wise activities and actions with support for choice, interaction and competency. Activity diagrams are intended to model both computational and organizational process (workflows), as well as the data flows intercepting with relating activities.

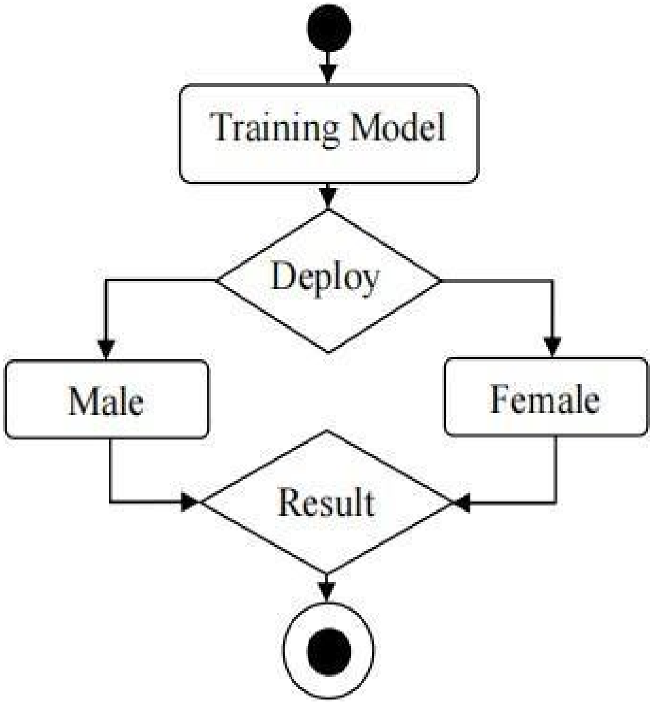


Figure 3.3.5 Activity Diagram

CHAPTER 4

REQUIREMENT SPECIFICATION

4.1 FUNCTIONAL REQUIREMENTS

The functional requirements for a system describe what the system should do. These requirements depend on the type of software being developed. The general approach taken by the organization when writing requirements. The functional system requirements describe the system function in detail, its inputs and outputs, exceptions and so on.

The Functional requirements are as follows:

- Develop a robust system which can predict the gender of a recorded voice sample.
- Train the model against enough variety in dataset to overcome gender bias.
- Data pre-processing on the dataset to increase the accuracy prediction.
- Provide web interface features for the users to get indicate the accuracies on various models.

4.2 NON-FUNCTIONAL REQUIREMENTS

Non-functional requirements, as the name suggests, are requirements that are notdirectly concerned with the specific functions delivered by the system. They may relate to emergent system properties such as reliability, response time and store occupancy. Alternatively, they may define constraints on the system such as capabilities of I/O devices and the data representations used in system interfaces.

The non-functional requirements are as follows:

- Make the system in such a way that after uploading an audio file, it takes minimum amount of time to processes and display the result.

- Record the voice sample from the microphone in a confined space and as quite as possible to obtain best prediction. The algorithm should never fail in any of the testcases.

4.3 HARDWARE REQUIREMENTS

Processor	:	Intel i5 7th gen or above
RAM	:	8GB
Hard Disk	:	256 GB
Peripherals	:	Monitor, Keyboard, Mouse, Microphone, Speaker

4.4 SOFTWARE REQUIREMENTS

Operating System	:	Windows 10 (64bit)
Development Software	:	Anaconda Jupyter Notebook, R Studio, Visual Studio and Wamp Server
Web Browser	:	Anything which supports HTML 5
Programming Environment	:	Python version 3.8.2 R version 3.6.3 Flask version 1.1.2

CHAPTER 5

IMPLEMENTATION MODULES

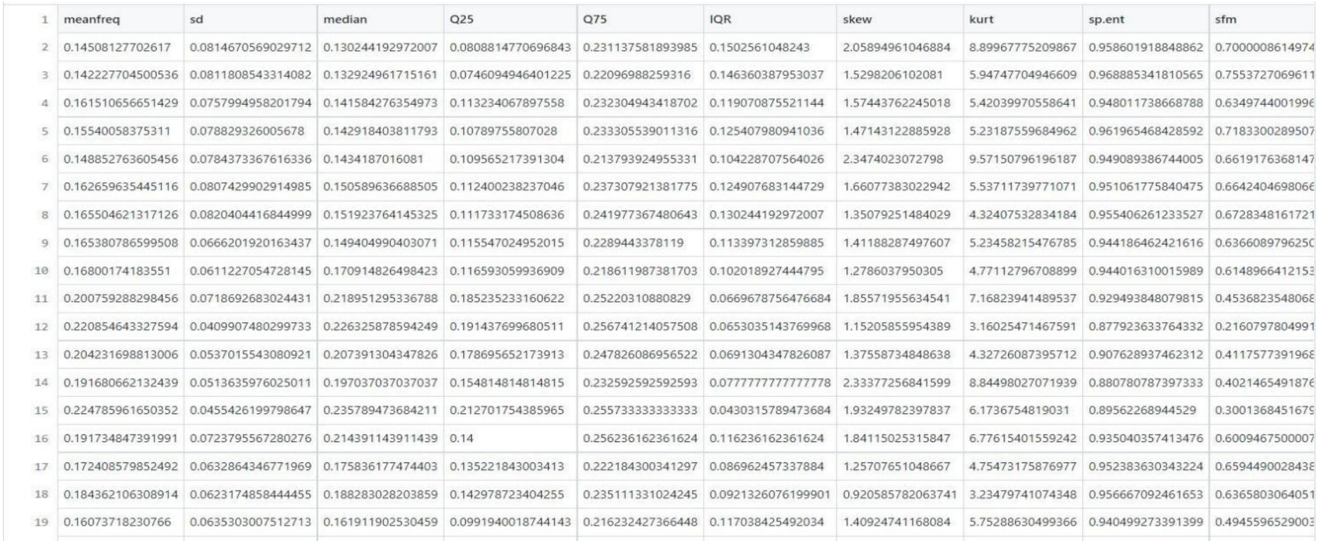
5.1 DATA VISUALISATION

In order to analyze gender by voice and speech, a training database was required. Adatabase was built using thousands of samples of male and female voices in Figure 5.1.1, Figure 5.1.2, Figure 5.1.3 each labeled by their gender of male or female. Each voice sampleis stored as a .WAV file, which is then pre-processed for acoustic analysis using thefunction from the Warble R package. It can measures 22 acoustic parameters onacoustic signals for which the start and end times are provided. The output from the pre-processed WAV files were saved into a CSV file, containing 3168 rows and 21columns (20 columns for each feature and one label column for the classification of male or female). You can download the pre-processed data set in CSV format, usingthe link above. Figure 5.1.4 represents the visualization diagram.

The following acoustic properties of each voice are measured:

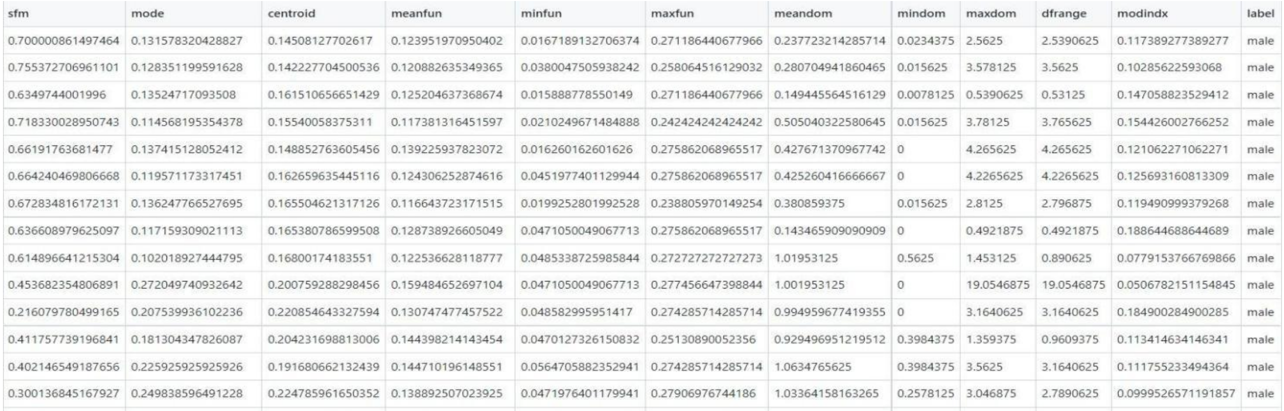
- **duration**: length of signal
- **meanfreq**: mean frequency (in kHz)
- **sd**: standard deviation of frequency
- **median**: median frequency (in kHz)
- **Q25**: first quantile (in kHz)
- **Q75**: third quantile (in kHz)
- **IQR**: interquantile range (in kHz)
- **skew**: skewness (see note in specprop description)
- **kurt**: kurtosis (see note in specprop description)
- **sp.ent**: spectral entropy
- **sfm**: spectral flatness

- **mode**: mode frequency
- **centroid**: frequency centroid (see specprop)
- **peakf**: peak frequency (frequency with highest energy)
- **meanfun**: average of fundamental frequency measured across acoustic signal
- **minfun**: minimum fundamental frequency measured across acoustic signal
- **maxfun**: maximum fundamental frequency measured across acoustic signal
- **meandom**: average of dominant frequency measured across acoustic signal
- **mindom**: minimum of dominant frequency measured across acoustic signal
- **maxdom**: maximum of dominant frequency measured across acoustic signal
- **dfrange**: range of dominant frequency measured across acoustic signal
- **modindx**: modulation index. Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range



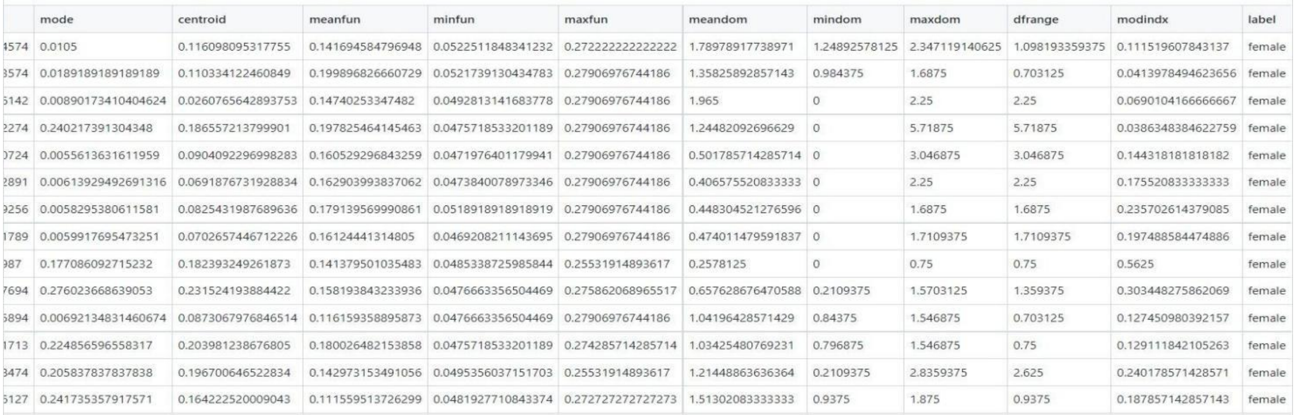
1	meanfreq	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm
2	0.14508127702617	0.0814670569029712	0.130244192972007	0.0808814770696843	0.231137581893985	0.1502561048243	2.05894961046884	8.89967775209867	0.958601918848862	0.7000008614974
3	0.142227704500536	0.0811808543314082	0.132924961715161	0.0746094946401225	0.22096988259316	0.146360387953037	1.5298206102081	5.94747704946609	0.968885341810565	0.7553727069611
4	0.161510656651429	0.0757994958201794	0.141584276354973	0.113234067897558	0.232304943418702	0.119070875521144	1.57443762245018	5.42039970558641	0.948011738668788	0.6349744001996
5	0.15540058375311	0.078829326005678	0.142918403811793	0.10789755807028	0.233305539011316	0.125407980941036	1.47143122885928	5.23187559684962	0.961965468428592	0.7183300289507
6	0.148852763605456	0.0784373367616336	0.1434187016081	0.109565217391304	0.213793924955331	0.104228707564026	2.3474023072798	9.57150796196187	0.949089386744005	0.6619176368147
7	0.162659635445116	0.0807429902914985	0.150589636688505	0.112400238237046	0.237307921381775	0.124907683144729	1.66077383022942	5.53711739771071	0.951061775840475	0.6642404698066
8	0.165504621317126	0.0820404416844999	0.151923764145325	0.111733174508636	0.241977367480643	0.130244192972007	1.35079251484029	4.32407532834184	0.955406261233527	0.6728348161721
9	0.165380786599508	0.0666201920163437	0.149404990403071	0.115547024952015	0.2289443378119	0.113397312859885	1.41188287497607	5.23458215476785	0.944186462421616	0.6366089796250
10	0.16800174183551	0.0611227054728145	0.170914826498423	0.116593059936909	0.218611987381703	0.102018927444795	1.2786037950305	4.77112796708899	0.944016310015989	0.6148966412153
11	0.200759288298456	0.0718692683024431	0.218951295336788	0.185235233160622	0.25220310880829	0.0669678756476684	1.85571955634541	7.16823941489537	0.929493848079815	0.4536823548068
12	0.220854643327594	0.0409907480299733	0.226325878594249	0.191437699680511	0.256741214057508	0.0653035143769968	1.15205855954389	3.16025471467591	0.877923633764332	0.2160797804991
13	0.204231698813006	0.0537015543080921	0.207391304347826	0.178695652173913	0.247826086956522	0.0691304347826087	1.37558734848638	4.32726087395712	0.907628937462312	0.4117577391968
14	0.191680662132439	0.0513635976025011	0.197037037037037	0.154814814814815	0.232592592592593	0.0777777777777778	2.33377256841599	8.84498027071939	0.880780787397333	0.4021465491876
15	0.224785961650352	0.0455426199798647	0.235789473684211	0.212701754385965	0.255733333333333	0.0430315789473684	1.93249782397837	6.1736754819031	0.89562268944529	0.3001368451679
16	0.191734847391991	0.0723795567280276	0.214391143911439	0.14	0.256236162361624	0.116236162361624	1.84115025315847	6.77615401559242	0.935040357413476	0.6009467500007
17	0.172408579852492	0.0632864346771969	0.175836177474403	0.135221843003413	0.222184300341297	0.086962457337884	1.25707651048667	4.75473175876977	0.952383630343224	0.6594490028438
18	0.184362106308914	0.0623174858444455	0.188283028203859	0.142978723404255	0.235111331024245	0.0921326076199901	0.920585782063741	3.23479741074348	0.956667092461653	0.6365803064051
19	0.16073718230766	0.0635303007512713	0.161911902530459	0.0991940018744143	0.216232427366448	0.117038425492034	1.40924741168084	5.75288630499366	0.940499273391399	0.4945596529003

Figure 5.1.1 Male Dataset



sfm	mode	centroid	meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange	modindx	label
0.700000861497464	0.131578320428827	0.14508127702617	0.123951970950402	0.0167189132706374	0.271186440677966	0.237723214285714	0.0234375	2.5625	2.5390625	0.117389277389277	male
0.755372706961101	0.128351199591628	0.142227704500536	0.120882635349365	0.0380047505938242	0.258064516129032	0.280704941860465	0.015625	3.578125	3.5625	0.10285622593068	male
0.6349744001996	0.13524717093508	0.161510656651429	0.125204637368674	0.015888778550149	0.271186440677966	0.149445564516129	0.0078125	0.5390625	0.53125	0.147058823529412	male
0.718330028950743	0.114568195354378	0.15540058375311	0.117381316451597	0.0210249671484888	0.242424242424242	0.505040322580645	0.015625	3.78125	3.765625	0.154426002766252	male
0.66191763681477	0.137415128052412	0.148852763605456	0.139225937823072	0.016260162601626	0.275862068965517	0.427671370967742	0	4.265625	4.265625	0.121062271062271	male
0.664240469806668	0.119571173317451	0.162659635445116	0.124306252874616	0.0451977401129944	0.275862068965517	0.425260416666667	0	4.2265625	4.2265625	0.125693160813309	male
0.672834816172131	0.136247766527695	0.165504621317126	0.116643723171515	0.0199252801992528	0.238805970149254	0.380859375	0.015625	2.8125	2.796875	0.119490999379268	male
0.636608979625097	0.117159309021113	0.165380786599508	0.128738926605049	0.0471050049067713	0.275862068965517	0.143465909090909	0	0.4921875	0.4921875	0.188644688644689	male
0.614896641215304	0.102018927444795	0.16800174183551	0.122536628118777	0.0485338725985844	0.272727272727273	1.01953125	0.5625	1.453125	0.890625	0.0779153766769866	male
0.453682354806891	0.272049740932642	0.200759288298456	0.159484652697104	0.0471050049067713	0.277456647398844	1.001953125	0	19.0546875	19.0546875	0.0506782151154845	male
0.216079780499165	0.207539936102236	0.220854643327594	0.130747477457522	0.048582995951417	0.274285714285714	0.994959677419355	0	3.1640625	3.1640625	0.184900284900285	male
0.411757739196841	0.181304347826087	0.204231698813006	0.144398214143454	0.0470127326150832	0.25130890052356	0.929496951219512	0.3984375	1.359375	0.9609375	0.113414634146341	male
0.402146549187656	0.225925925925926	0.191680662132439	0.144710196148551	0.0564705882352941	0.274285714285714	1.0634765625	0.3984375	3.5625	3.1640625	0.111755233494364	male
0.300136845167927	0.249838596491228	0.224785961650352	0.138892507023925	0.0471976401179941	0.27906976744186	1.03364158163265	0.2578125	3.046875	2.7890625	0.0999526571191857	male

Figure 5.1.2 Male Dataset



	mode	centroid	meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange	modindx	label
4574	0.0105	0.116098095317755	0.141694584796948	0.0522511848341232	0.272222222222222	1.78978917738971	1.24892578125	2.347119140625	1.098193359375	0.111519607843137	female
3574	0.0189189189189189	0.110334122460849	0.199896826660729	0.0521739130434783	0.27906976744186	1.35825892857143	0.984375	1.6875	0.703125	0.0413978494623656	female
5142	0.00890173410404624	0.0260765642893753	0.14740253347482	0.0492813141683778	0.27906976744186	1.965	0	2.25	2.25	0.0690104166666667	female
2274	0.240217391304348	0.186557213799901	0.197825464145463	0.0475718533201189	0.27906976744186	1.24482092696629	0	5.71875	5.71875	0.0386348384622759	female
0724	0.0055613631611959	0.0904092296998283	0.160529296843259	0.0471976401179941	0.27906976744186	0.501785714285714	0	3.046875	3.046875	0.144318181818182	female
2891	0.00613929492691316	0.0691876731928834	0.162903993837062	0.0473840078973346	0.27906976744186	0.406575520833333	0	2.25	2.25	0.175520833333333	female
9256	0.0058295380611581	0.0825431987689636	0.179139569990861	0.0518918918918919	0.27906976744186	0.448304521276596	0	1.6875	1.6875	0.235702614379085	female
1789	0.0059917695473251	0.0702657446712226	0.16124441314805	0.0469208211143695	0.27906976744186	0.474011479591837	0	1.7109375	1.7109375	0.197488584474886	female
987	0.177086092715232	0.182393249261873	0.141379501035483	0.0485338725985844	0.25531914893617	0.2578125	0	0.75	0.75	0.5625	female
7694	0.276023668639053	0.231524193884422	0.158193843233936	0.0476663356504469	0.275862068965517	0.657628676470588	0.2109375	1.5703125	1.359375	0.303448275862069	female
5894	0.00692134831460674	0.0873067976846514	0.116159358895873	0.0476663356504469	0.27906976744186	1.04196428571429	0.84375	1.546875	0.703125	0.127450980392157	female
1713	0.224856596558317	0.203981238676805	0.180026482153858	0.0475718533201189	0.274285714285714	1.03425480769231	0.796875	1.546875	0.75	0.129111842105263	female
3474	0.205837837837838	0.196700646522834	0.142973153491056	0.0495356037151703	0.25531914893617	1.21448863636364	0.2109375	2.8359375	2.625	0.240178571428571	female
5127	0.241735357917571	0.164222520009043	0.111559513726299	0.0481927710843374	0.272727272727273	1.51302083333333	0.9375	1.875	0.9375	0.187857142857143	female

Figure 5.1.3 Female Dataset

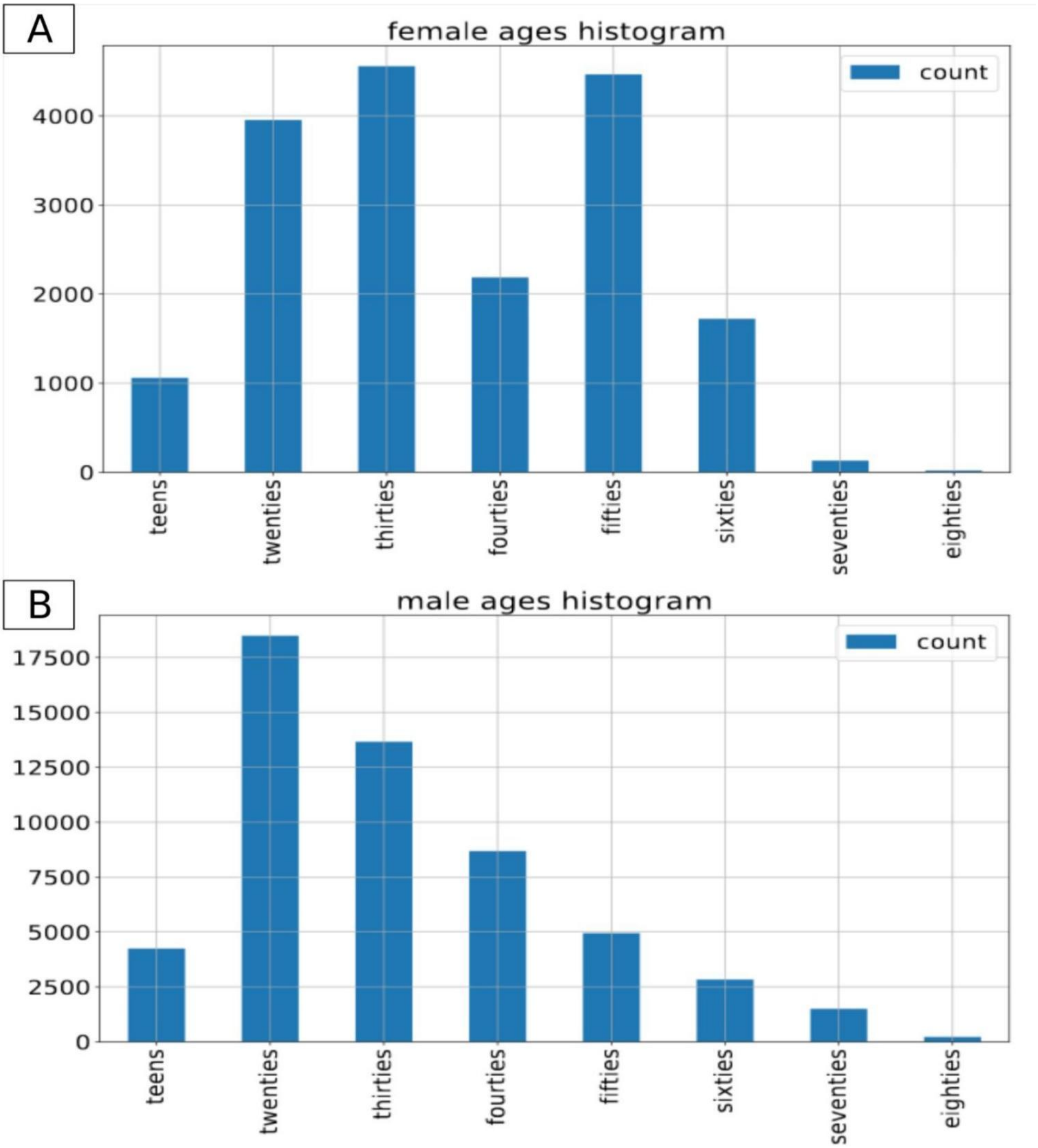
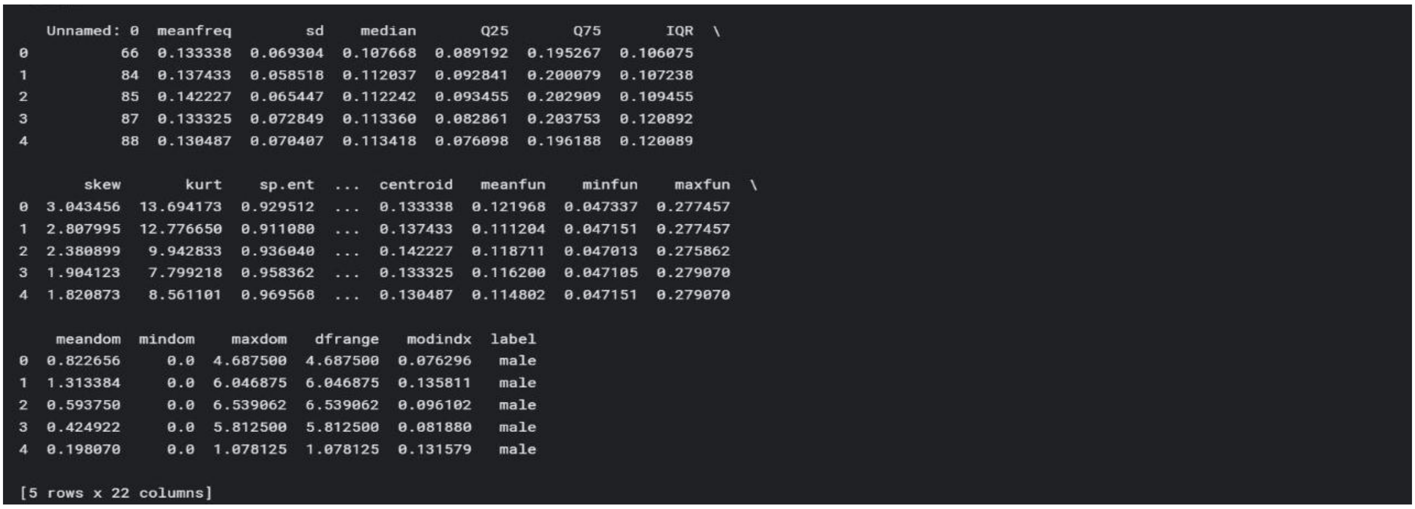


Figure 5.1.4 Visualization Diagram

5.2 DATA PRE-PROCESSING

Here we will preprocess in Figure 5.2 the data, as some algorithms such as K-Nearest- Neighbor Classifier (KNN), Naive Bayes Classifier, Random Forest and SVM tend to perform better with scaled data. In addition, we will also split the full data set into training and test datasets.



```
   Unnamed: 0  meanfreq        sd    median       Q25       Q75       IQR  \
0          66  0.133338  0.069304  0.107668  0.089192  0.195267  0.106075
1          84  0.137433  0.058518  0.112037  0.092841  0.200079  0.107238
2          85  0.142227  0.065447  0.112242  0.093455  0.202909  0.109455
3          87  0.133325  0.072849  0.113360  0.082861  0.203753  0.120892
4          88  0.130487  0.070407  0.113418  0.076098  0.196188  0.120089

       skew       kurt    sp.ent  ...  centroid   meanfun    minfun    maxfun  \
0  3.043456  13.694173  0.929512  ...  0.133338  0.121968  0.047337  0.277457
1  2.807995  12.776650  0.911080  ...  0.137433  0.111204  0.047151  0.277457
2  2.380899   9.942833  0.936040  ...  0.142227  0.118711  0.047013  0.275862
3  1.904123   7.799218  0.958362  ...  0.133325  0.116200  0.047105  0.279070
4  1.820873   8.561101  0.969568  ...  0.130487  0.114802  0.047151  0.279070

    meandom  mindom    maxdom   dfrange   modindx  label
0  0.822656     0.0  4.687500  4.687500  0.076296   male
1  1.313384     0.0  6.046875  6.046875  0.135811   male
2  0.593750     0.0  6.539062  6.539062  0.096102   male
3  0.424922     0.0  5.812500  5.812500  0.081880   male
4  0.198070     0.0  1.078125  1.078125  0.131579   male

[5 rows x 22 columns]
```

Figure 5.2 Data Pre-Processing

5.3 MODULES OF THE PROJECT

Module 1

Module Name: Training Module

Functionality: The audio features extracted using the WarbleR program is feed to the training model in the form of CSV file. Pre-processing is done on the input dataset. Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Pr-eprocessing is a technique that is used to convert the raw data into aclean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. After this, the dataset is trained using Naive Bayes classifier, Support Vector Machine, k-Nearest Neighbors classifier and Random Forest Classifier.

Module 2

Module Name: WarbleR Feature Extraction Module

Functionality: R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. warbleR is a package designed to streamline analysis of acoustic signals in R. This package allows users to collect open-access avian vocalizations data or input their own data into a workflow that facilitates spectrographic visualization and measurement of acoustic parameters. warbleR makes fundamental sound analysis tools from the R package see wave, as well as new toolsnot yet offered in the R environment, readily available for batch process analysis.

Module 3

Module Name: Testing Module

Functionality: After the training is done, the system is ready for predictions. The user now records a new voice sample, this voice sample is feed to the WarbleR program to extract the audio features necessary and classify the given sample as either male or female.

CHAPTER 6

RESULTS AND DISCUSSIONS

6.1 SNAPSHOTS OF MODULES

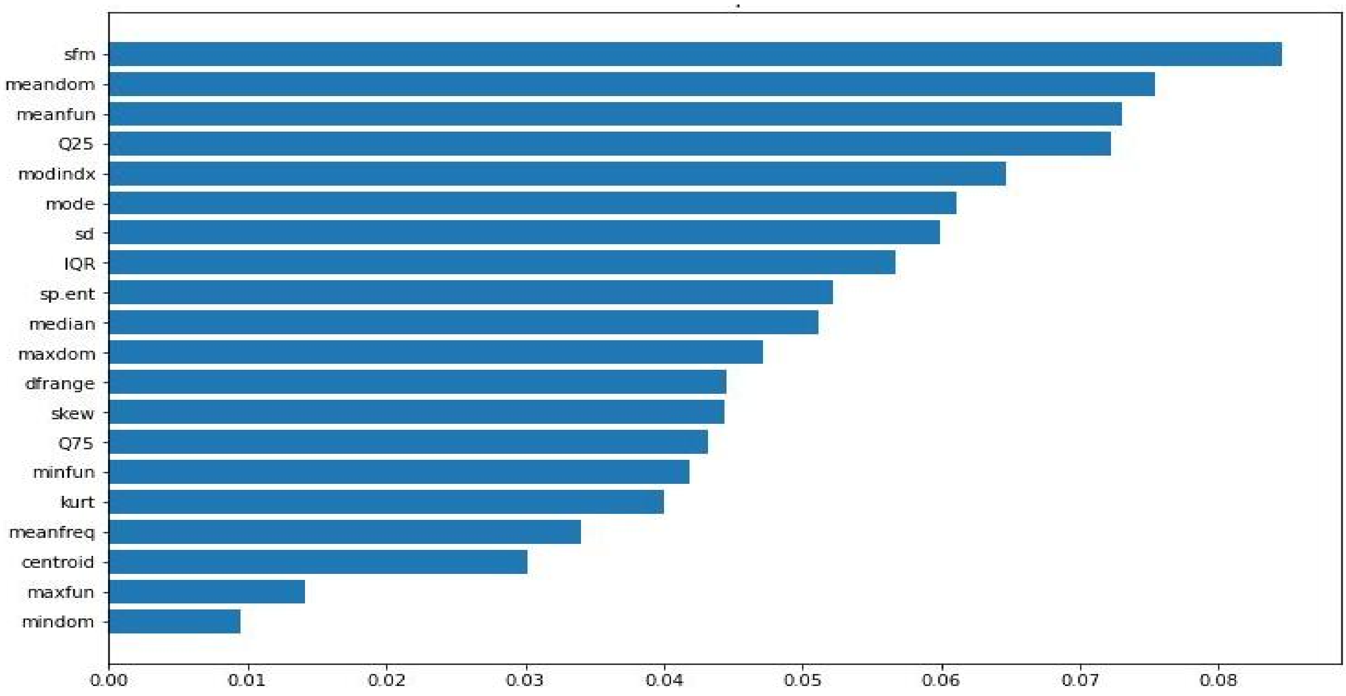
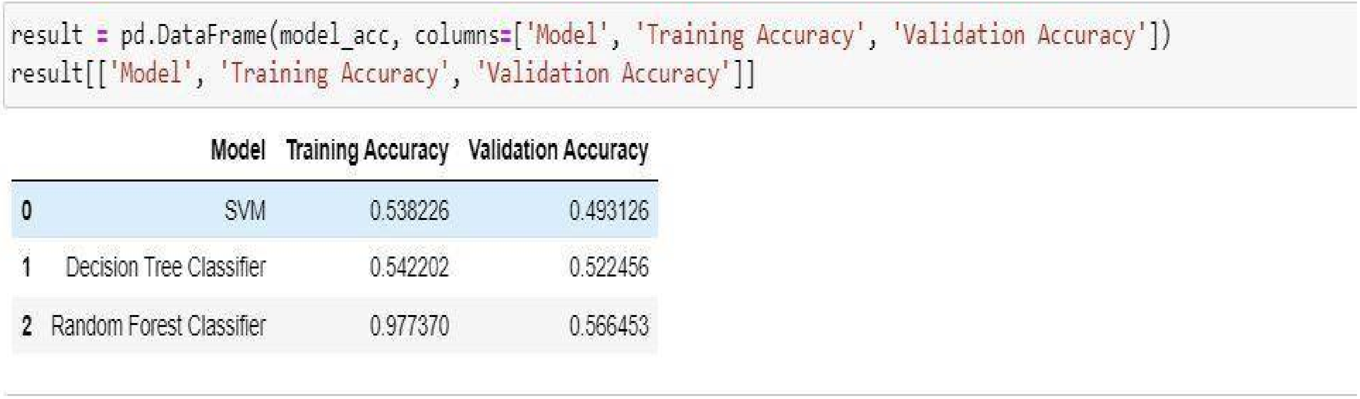


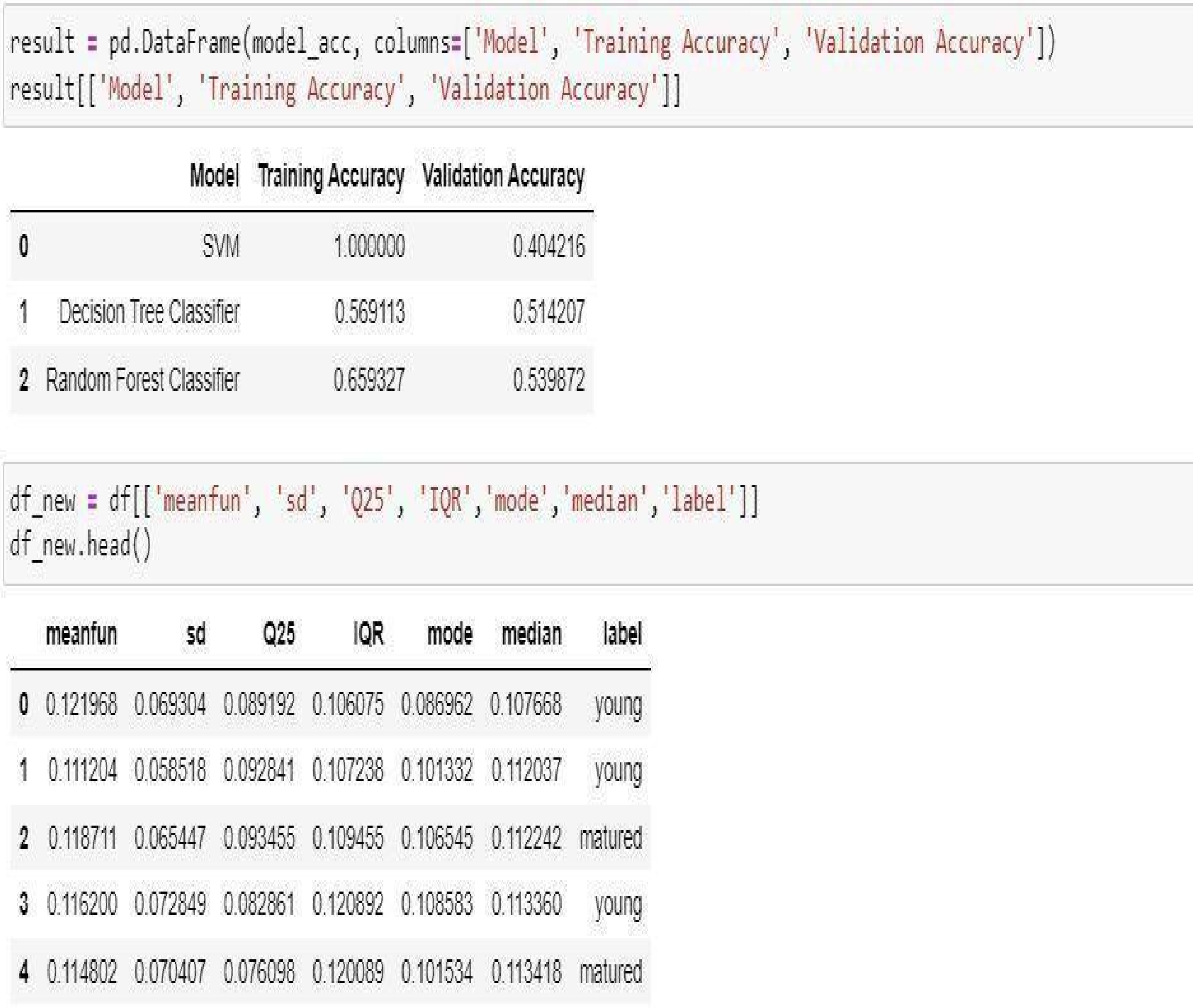
Figure 6.1.1 Feature Importance



```
result = pd.DataFrame(model_acc, columns=['Model', 'Training Accuracy', 'Validation Accuracy'])
result[['Model', 'Training Accuracy', 'Validation Accuracy']]
```

	Model	Training Accuracy	Validation Accuracy
0	SVM	0.538226	0.493126
1	Decision Tree Classifier	0.542202	0.522456
2	Random Forest Classifier	0.977370	0.566453

Figure 6.1.2 Prediction Accuracy Age



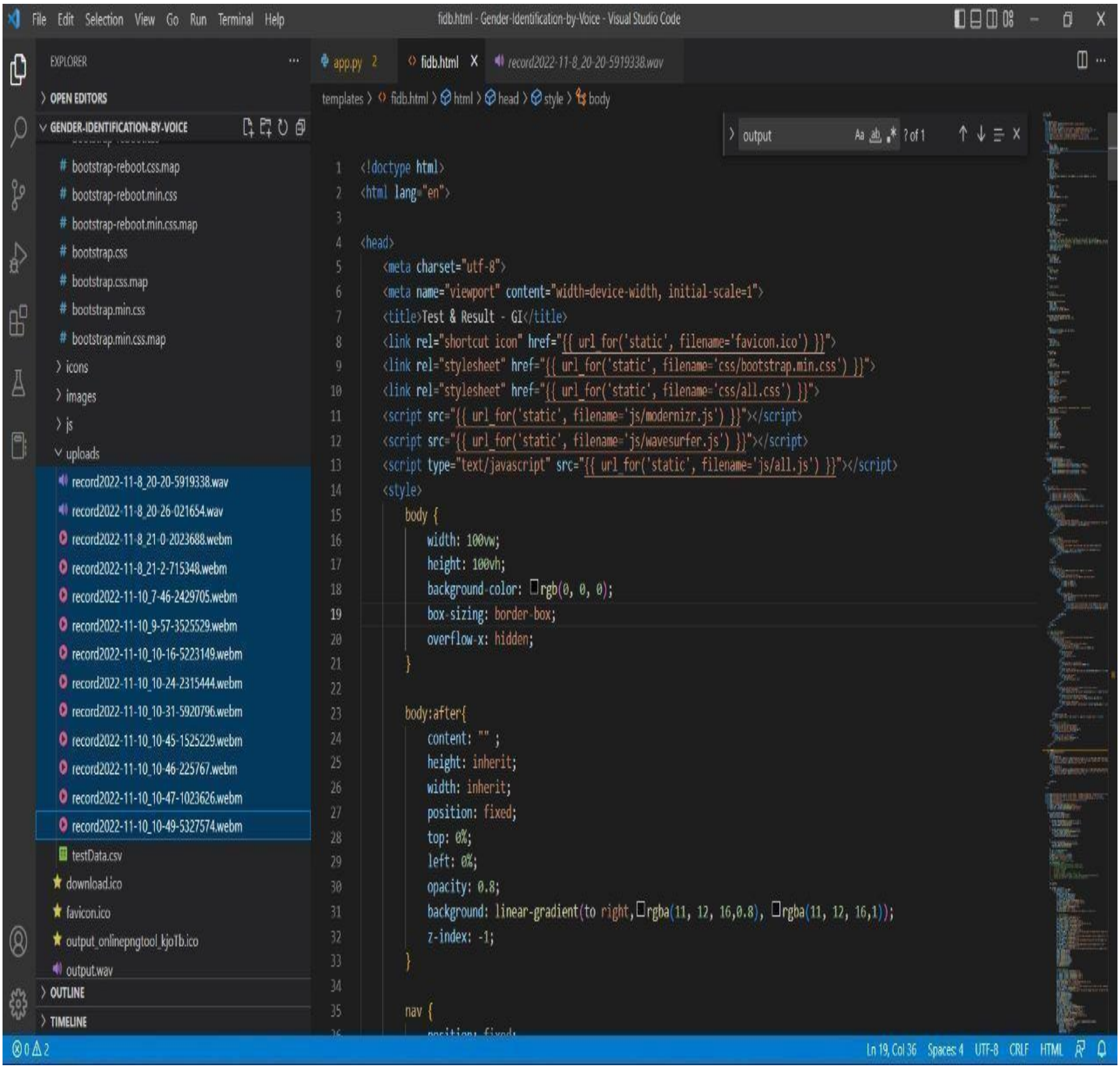
```
result = pd.DataFrame(model_acc, columns=['Model', 'Training Accuracy', 'Validation Accuracy'])
result[['Model', 'Training Accuracy', 'Validation Accuracy']]
```

	Model	Training Accuracy	Validation Accuracy
0	SVM	1.000000	0.404216
1	Decision Tree Classifier	0.569113	0.514207
2	Random Forest Classifier	0.659327	0.539872

```
df_new = df[['meanfun', 'sd', 'Q25', 'IQR','mode','median','label']]
df_new.head()
```

	meanfun	sd	Q25	IQR	mode	median	label
0	0.121968	0.069304	0.089192	0.106075	0.086962	0.107668	young
1	0.111204	0.058518	0.092841	0.107238	0.101332	0.112037	young
2	0.118711	0.065447	0.093455	0.109455	0.106545	0.112242	matured
3	0.116200	0.072849	0.082861	0.120892	0.108583	0.113360	young
4	0.114802	0.070407	0.076098	0.120089	0.101534	0.113418	matured

Figure 6.1.3 Prediction Accuracy Emotion and Gender


File
Edit
Selection
View
Go
Run
Terminal
Help
fidb.html - Gender-Identification-by-Voice - Visual Studio Code
EXPLORER
OPEN EDITORS
app.py 2
fidb.html
record2022-11-8_20-20-5919338.wav
templates > fidb.html > html > head > style > body
output
? of 1
OUTLINE
TIMELINE
0 2
Ln 19, Col 36
Spaces: 4
UTF-8
CRLF
HTML

```
GENDER-IDENTIFICATION-BY-VOICE
    # bootstrap-reboot.css.map
    # bootstrap-reboot.min.css
    # bootstrap-reboot.min.css.map
    # bootstrap.css
    # bootstrap.css.map
    # bootstrap.min.css
    # bootstrap.min.css.map
  > icons
  > images
  > js
  ∨ uploads
    record2022-11-8_20-20-5919338.wav
    record2022-11-8_20-26-021654.wav
    record2022-11-8_21-0-2023688.webm
    record2022-11-8_21-2-715348.webm
    record2022-11-10_7-46-2429705.webm
    record2022-11-10_9-57-3525529.webm
    record2022-11-10_10-16-5223149.webm
    record2022-11-10_10-24-2315444.webm
    record2022-11-10_10-31-5920796.webm
    record2022-11-10_10-45-1525229.webm
    record2022-11-10_10-46-225767.webm
    record2022-11-10_10-47-1023626.webm
    record2022-11-10_10-49-5327574.webm
    testData.csv
  download.ico
  favicon.ico
  output_onlinepngtool_kjoTb.ico
  output.wav
```

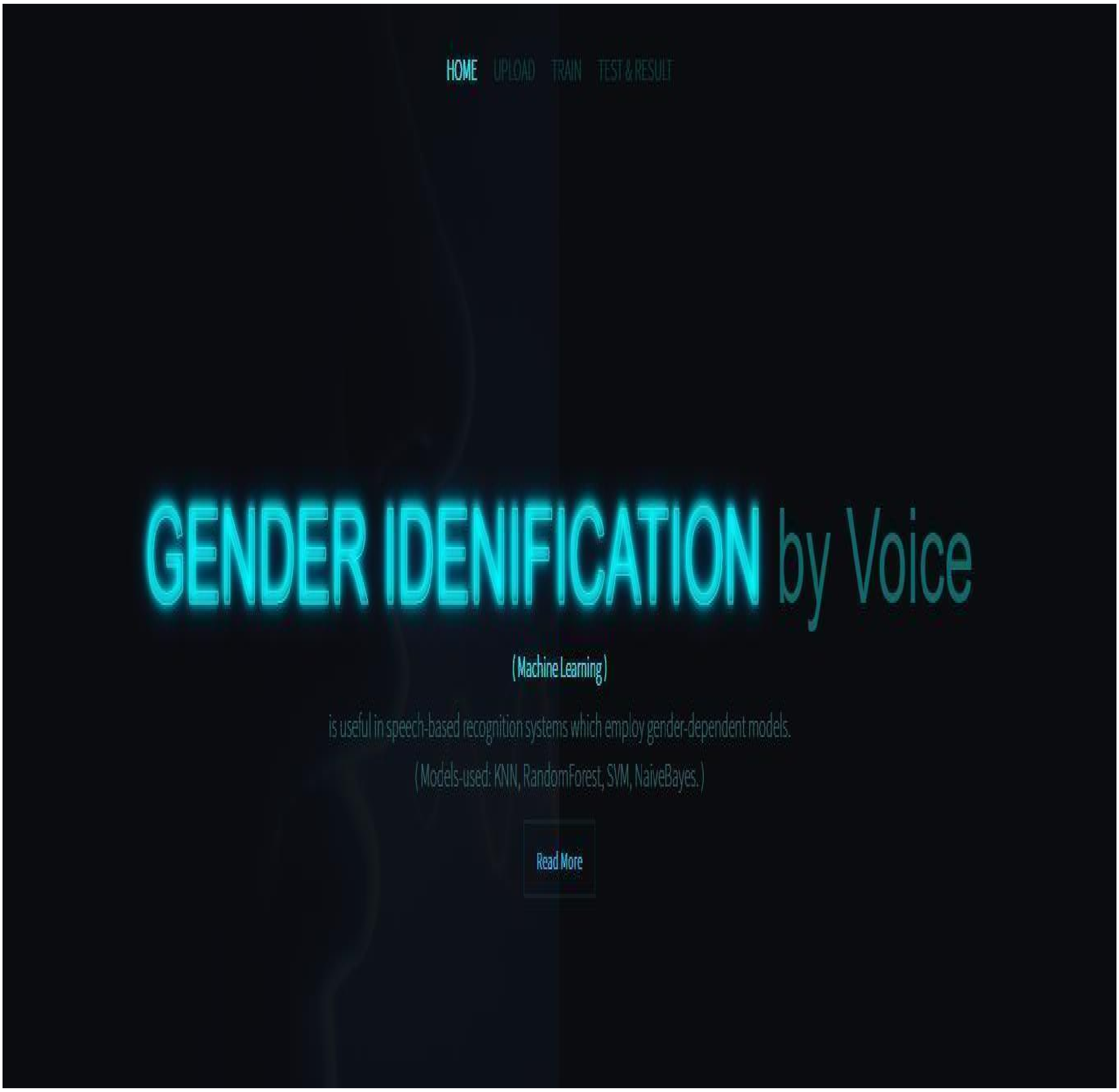
```html
<!doctype html>
<html lang="en">

<head>
    <meta charset="utf-8">
    <meta name="viewport" content="width=device-width, initial-scale=1">
    <title>Test & Result - GI</title>
    <link rel="shortcut icon" href="{{ url_for('static', filename='favicon.ico') }}">
    <link rel="stylesheet" href="{{ url_for('static', filename='css/bootstrap.min.css') }}">
    <link rel="stylesheet" href="{{ url_for('static', filename='css/all.css') }}">
    <script src="{{ url_for('static', filename='js/modernizr.js') }}"></script>
    <script src="{{ url_for('static', filename='js/wavesurfer.js') }}"></script>
    <script type="text/javascript" src="{{ url_for('static', filename='js/all.js') }}"></script>
    <style>
        body {
            width: 100vw;
            height: 100vh;
            background-color: rgb(0, 0, 0);
            box-sizing: border-box;
            overflow-x: hidden;
        }

        body:after{
            content: "" ;
            height: inherit;
            width: inherit;
            position: fixed;
            top: 0%;
            left: 0%;
            opacity: 0.8;
            background: linear-gradient(to right, rgba(11, 12, 16,0.8), rgba(11, 12, 16,1));
            z-index: -1;
        }

        nav {
            position: fixed;
```

Figure 6.1.4 Audio file stored in the Database



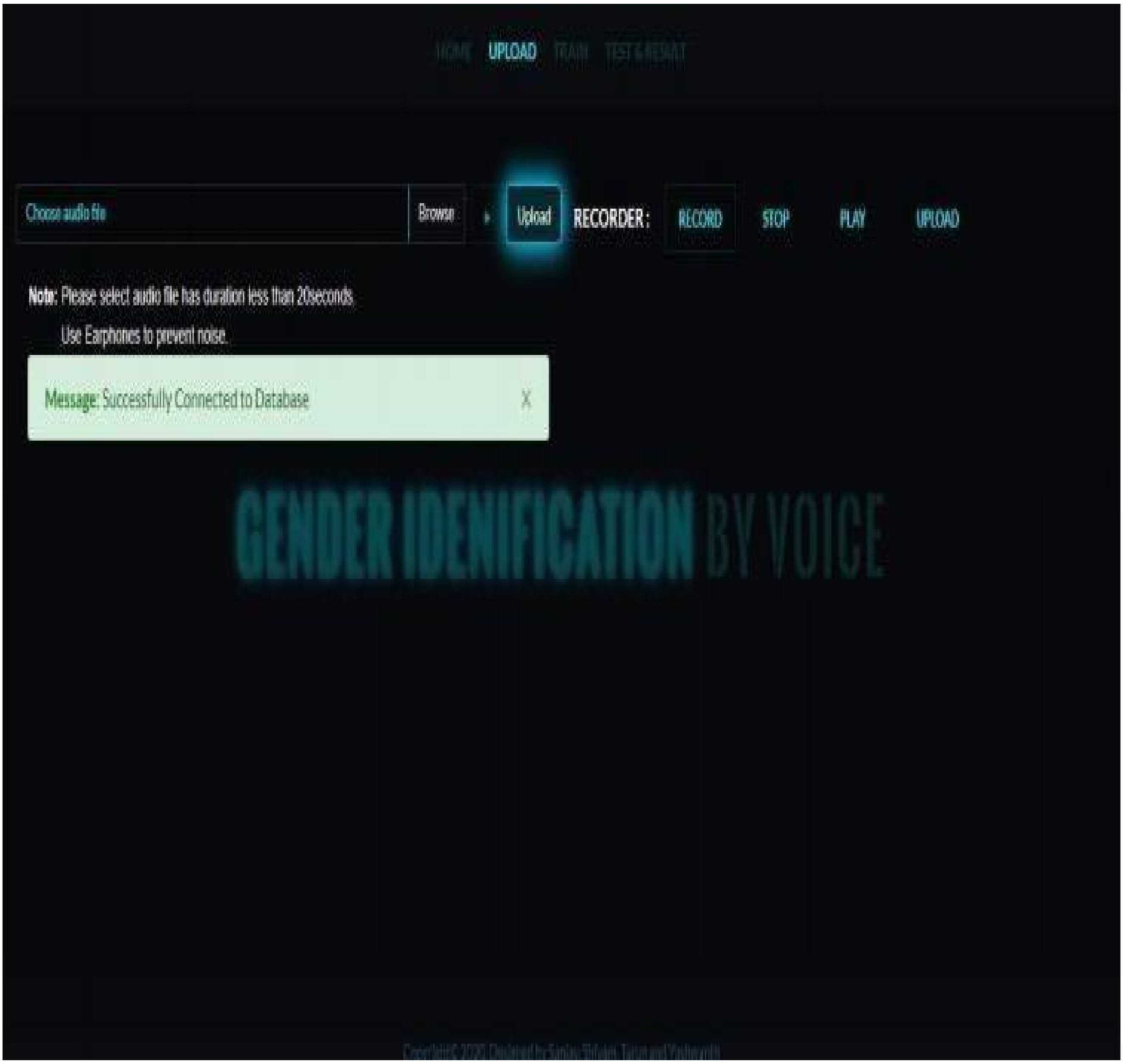
GENDER IDENIFICATION by Voice

(Machine Learning)

is useful in speech-based recognition systems which employ gender-dependent models.

(Models-used: KNN, RandomForest, SVM, NaiveBayes.)

Figure 6.1.5 Home Page of the web interface

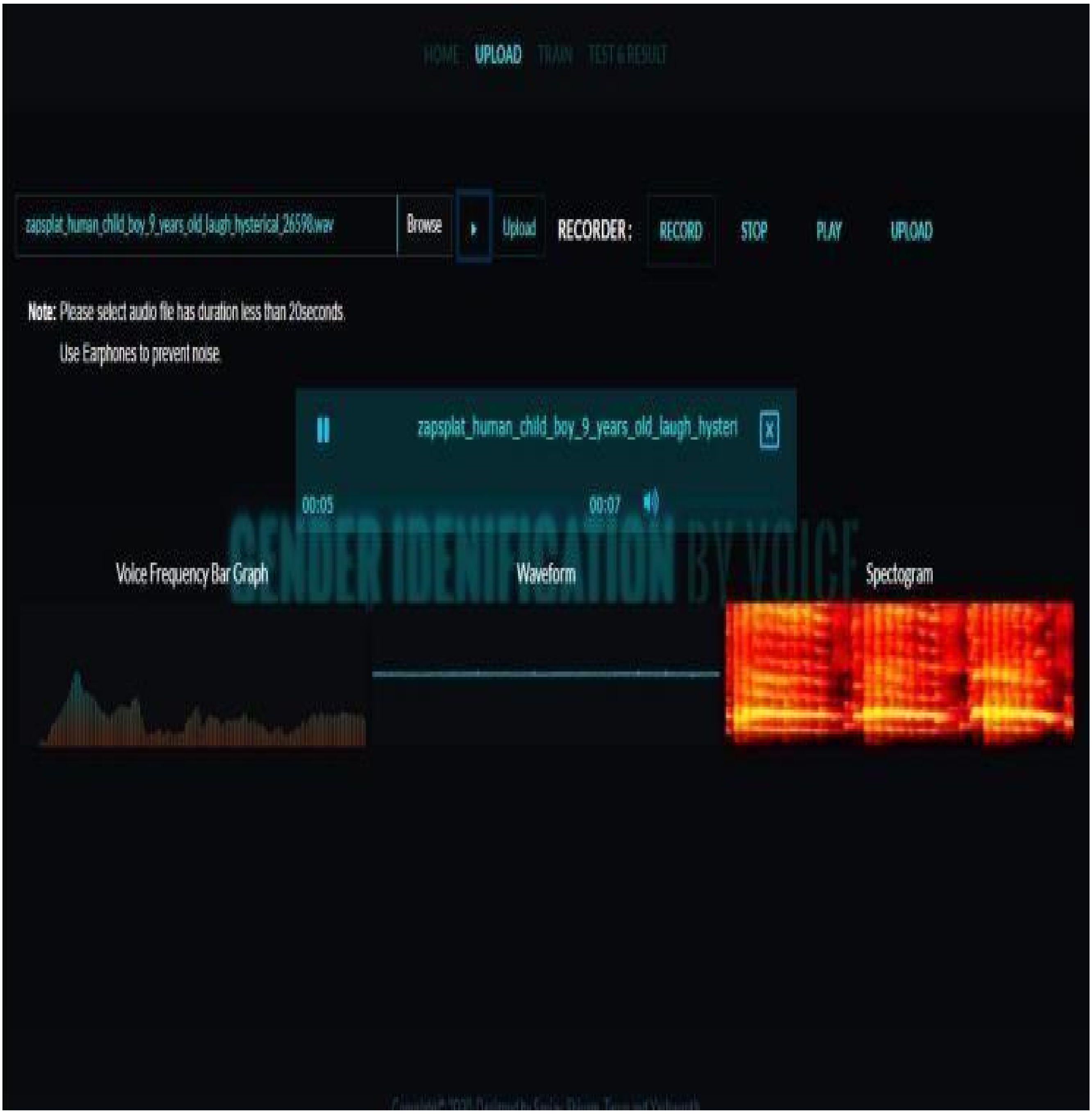


Note: Please select audio file has duration less than 20seconds.

Use Earphones to prevent noise.

Message: Successfully Connected to Database

Figure 6.1.6 Audio Upload Page, View-1



zapsplat_human_child_boy_9_years_old_laugh_hysterical_26598.wav

Note: Please select audio file has duration less than 20seconds.

Use Earphones to prevent noise.

zapsplat_human_child_boy_9_years_old_laugh_hysteri

Figure 6.1.7 Audio Upload Page, View-2

The “Train” page in Figure 6.1.8, where a user can upload the voice training dataset in form of a CSV file and as a result the accuracies in each learning model is depicted in a graph.

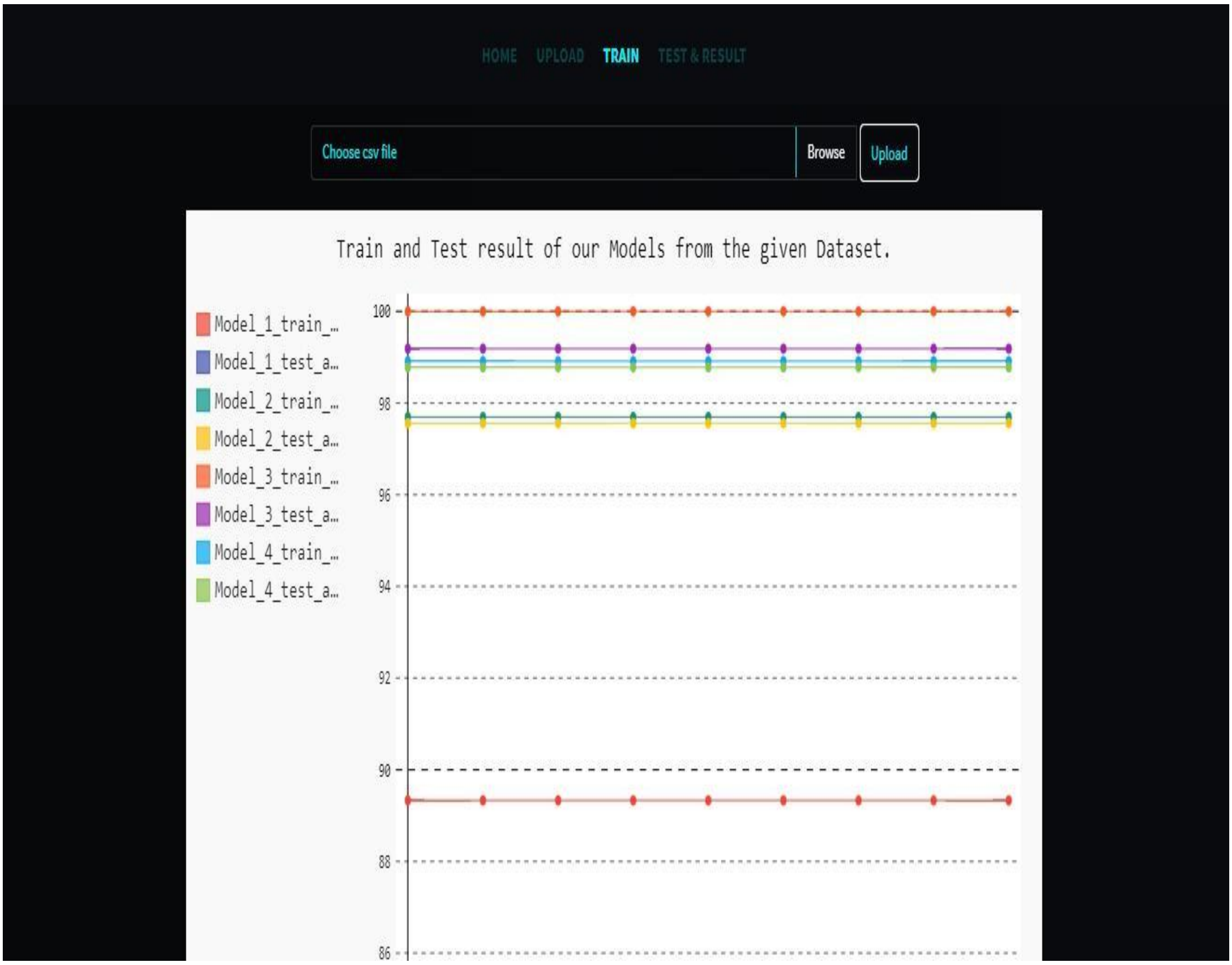
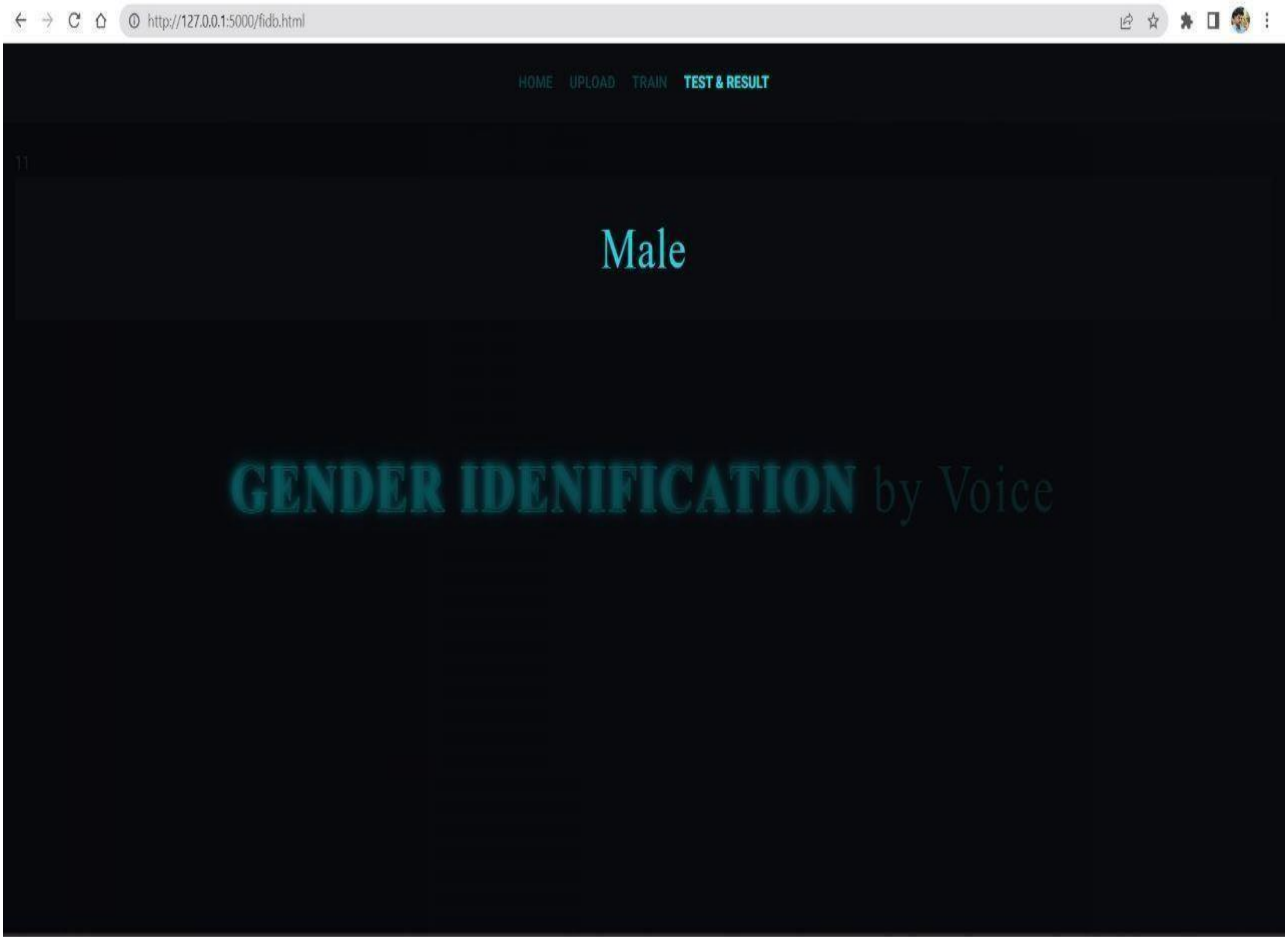


Figure 6.1.8 System Training Page

The “Test & Result” page in Figure 6.1.9, where a user can test the already uploaded audio file gender prediction. The audio upload in this case is a Male’s voice and hence the system predicts it as “Male” accurately.

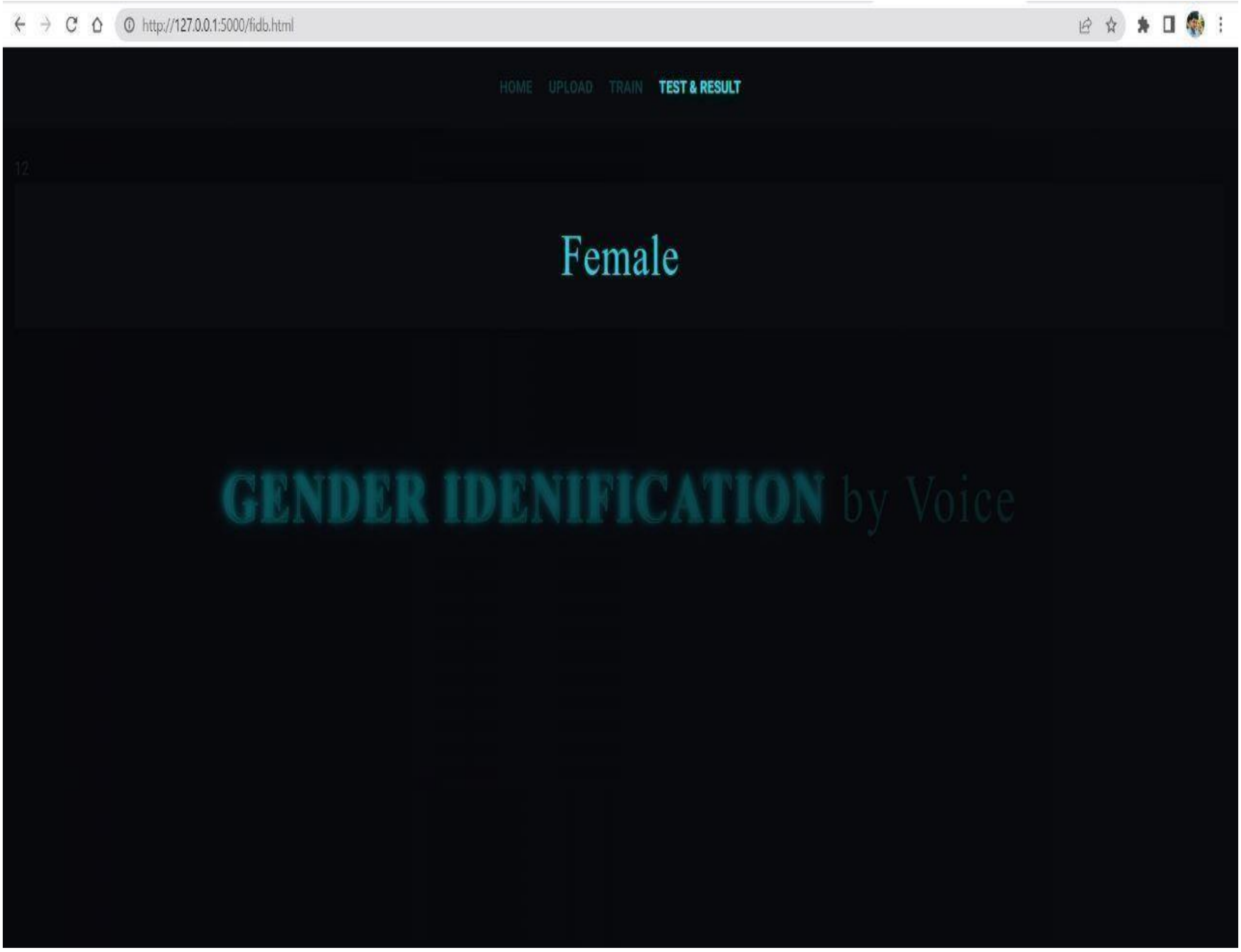


Male

GENDER IDENIFICATION by Voice

Figure 6.1.9 Testing and Result Page, View-1

The “Test & Result” page in Figure 6.1.10, where a user can test the already uploaded audio file gender prediction. The audio upload in this case is a Female’s voice and hence the system predicts it as “Female” accurately.



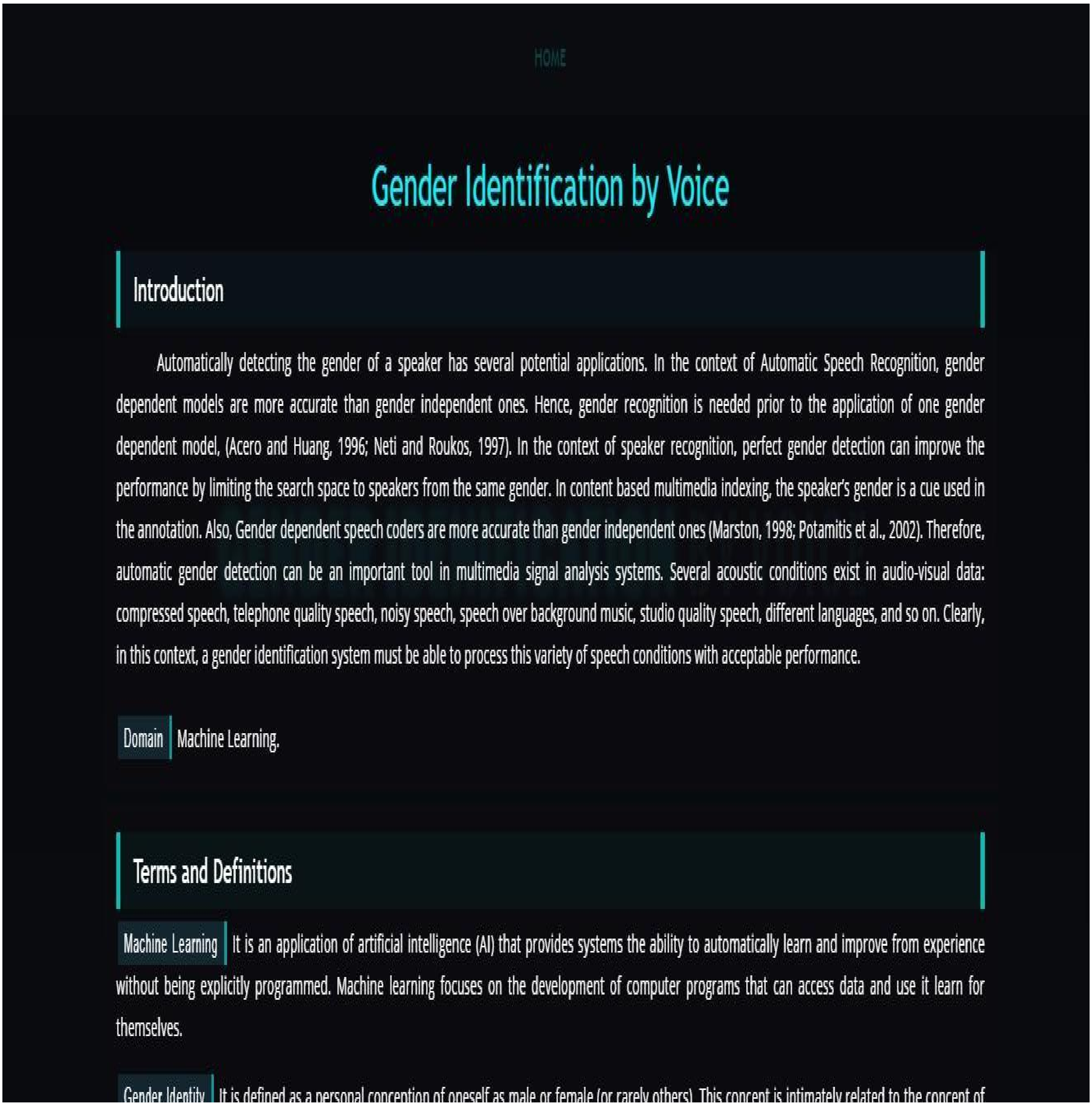
http://127.0.0.1:5000/fidb.html

12

Female

GENDER IDENIFICATION by Voice

Figure 6.1.10 Testing and Result Page, View-2



Gender Identification by Voice

Introduction

Automatically detecting the gender of a speaker has several potential applications. In the context of Automatic Speech Recognition, gender dependent models are more accurate than gender independent ones. Hence, gender recognition is needed prior to the application of one gender dependent model, (Acero and Huang, 1996; Neti and Roukos, 1997). In the context of speaker recognition, perfect gender detection can improve the performance by limiting the search space to speakers from the same gender. In content based multimedia indexing, the speaker's gender is a cue used in the annotation. Also, Gender dependent speech coders are more accurate than gender independent ones (Marston, 1998; Potamitis et al., 2002). Therefore, automatic gender detection can be an important tool in multimedia signal analysis systems. Several acoustic conditions exist in audio-visual data: compressed speech, telephone quality speech, noisy speech, speech over background music, studio quality speech, different languages, and so on. Clearly, in this context, a gender identification system must be able to process this variety of speech conditions with acceptable performance.

Domain Machine Learning.

Terms and Definitions

Machine Learning It is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

Gender Identity It is defined as a personal conception of oneself as male or female (or rarely others). This concept is intimately related to the concept of

Figure 6.1.11 System Details

CHAPTER 7

CONCLUSION AND FUTURE WORK

Recognizing the gender of human voice has been considered one of the challenging tasks because of its importance in various applications. The contributions are threefold including studying the extracted features by examining the correlation between each other, building classification models using different ML techniques from distinct families, and evaluating the natural feature selection techniques in finding the optimal subset of relevant features on classification performance.

Our experiments involve applying standard machine learning techniques such as Naïve Bayes, k-NN, Random Forest and Support Vector Machine to the voice-based gender identification problem. We can observe that Random Forest Decision Trees provides best accuracy which is up to 99% as the testing accuracy. In addition, we also conclude that general-purpose audio features may not be able to capture enough gender-specific characteristics of voice.

Automatically detecting the gender of a speaker has several potential applications. The applications of gender detection system have increased significantly due to the recent developments in speech/speaker recognition, human-computer interaction, and biometric security systems including authentication to access data, surveillance, and security. Moreover, a gender detection system can be used for automatic transfer of a phone call of a male/female to the relevant person or department. In a mobile healthcare system, gender detection can play a significant role. There are some vocal folds pathologies, which are biased to a gender; for example, vocal folds cyst can be seen particularly in female patients. If there is a mechanism to automatically detect the gender of the patient, it is easier for a healthcare professional to prescribe the appropriate treatment.

It would be interesting to introduce high order audio features to our models. In the future, more experiments can be conducted to use various feature categories, ML techniques, and other natural feature selection techniques. Furthermore, the proposed techniques can be examined on different datasets, since here only standard voice samples were used.

REFERENCES

1. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al., "Scikit-learn: Machine learning in python," the Journal of machine Learning research, vol. 12, 2011, pp. 2825–2830.

2. S. Jadav, "Voice-based gender identification using machine learning," in 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018, pp. 1–4.

3. Hadi Harb and Liming Chen, "Voice-Based Gender and age Identification in Multimedia Applications" J. Intell. Inf. Syst. 24, 2005, pp. 179-198.

4. Sarah Ita Levitan, Taniya Mishra & Srinivas Bangalore, "Automatic identification of gender and emotion from speech", 2016, pp. 84-88.

5. T. Jayasankar, K. Vinothkumar and Arputha Vijayaselvi, "Automatic Gender Identification in Speech Recognition by Genetic Algorithm", Applied Mathematics &Information Sciences 11, No. 3, 2017, pp. 907-913.

6. Remna R. Nair, Bhagya Vijayan, "Voice based Gender Recognition", International Research Journal of Engineering and Technology (IRJET), Vol. 6, No. 5, May 2019, pp. 3-45.

7. A Raahul, R Sapthagiri, K Pankaj and V Vijayarajan, "Voice based gender and age classification using machine learning", IOP Conf. Series: Materials Science and Engineering 263, pp. 6-20, 2017.

8. M.-H. Grosbras, P. D. Ross, and P. Belin, "Categorical emotion recognition from voice improves during childhood and adolescence," Scientific reports, vol. 8, no. 1, 2018 pp. 1–11.

9. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, no. 1,June 2002, pp. 321–357.

10. R. Djemili, H. Bourouba, and M. C. A. Korba, "A speech signal based gender and emotion identification system using four classifiers," in 2012 International Conference on Multimedia Computing and Systems, 2012, pp. 184–187.