ensit
Ecole Nationale Supérieure d'Ingénieurs de Tunis

Republic of Tunisia
Ministry of Higher Education
and scientific research
University of Tunis
The Higher National Engineering
School of Tunis

UNIVERSITÉ DE TUNIS
جامعة تونس

Department of Mathematical Engineering

End of Year Report

First year in mathematical and modeling engineering

Presented by:
Sarra Ben Doub

Speech language pathology diagnostic

publicly supported in 25 may

composition of the jury:

President: Md. Balkine KHADOUMIi
Supervisor: Mr. Sabeur ABID

Academic Year: 2023-2024

Dedication

I dedicate my work, as well as my sincere thanks and gratitude:

To my wonderful parents, who formed the foundation of my existence. I would never have made it without them. Thank you for all of your sacrifices, love, and moral support over the years.

To my sister and my brother, whom I adore and who have never stopped encouraging me, I wish them a life filled with pleasure, prosperity, and love.

To all my friends, who have always encouraged me and who have been able to give me help and support.

To everyone I love.

Thank you!

Acknowledgement

Before I deliver my end-of-year report, I’d want to thank everyone who contributed to the success of my project in various ways, especially those mentioned below.

I would like to thank **Mr Sofian Kasmi**, Director of the Department of Mathematics at The Higher National Engineering School of Tunis, as well as all of our teachers and members of the administration.

I also like to thank my academic supervisor, Mr. **Sabeur Abid**. I can only respect him expertise and method of operation. I would like to express my gratitude to him once more for him patience, generosity, and human warmth.

Finally, I would like to express a word of gratitude to all our **GMAM** professors for all their knowledge that they have been able to transmit to us and the jury members for accepting to judge this modest work.

Abstract

Dysphonia and other voice disorders are common throughout the world. The use of machine learning methods to recognize and classify voice disorders has been thoroughly investigated by researchers. However, the lack of sensitive and challenging-to-obtain medical data samples in large quantities limits the efficacy of deep learning algorithms in this field.

To overcome this challenge, To create a working model for testing using machine learning, we have gathered data from internet sources.

The project's main objective is to process speech data. Our main goal is to use machine learning techniques to create a model to classify normal people and dysphonia .

From the gathered data, we have extracted 1560 speech features for our study and have used them to train our classification model. We have used K-nearest neighbors, random forests, and support vector machines three cutting-edge classifiers to assess our model's effectiveness.

Keyword: medical signal processing, Machine learning, KNN, Random forest

Table of contents

List of figures

List of tables

Abbreviations

DL	Deep Learning
ISA	Salah Aziz Institute
ML	Machine Learning
SLP	Speech Language Pathology

General introduction

Vocal organs are essential to human communication because they enable us to convey information and express our emotions. Today, the majority of workers, including phone operators, teachers, broadcasters, and singers, rely on their voices for their primary means of expression.

A person's everyday life can be significantly impacted when their voice is compromised.

Dysphonia or dysphonic voice is often used to indicate impaired speech production.

Dysphonia is the perception of voice quality resulting from adverse changes in the phonation organs.

Then, a pathological voice may be a sign of health issues. But in order to diagnose that, a qualified specialist is needed, and a number of speech exercises must be finished.

In the context of this project, we propose the development prototype model that can distinguish between speech patterns that are typical and those that are indicative of dysphonia.

This report explores the nuances of our methodology, describing the methods used for feature extraction, data collection, and machine learning algorithm implementation. We hope to give you a thorough grasp of our methodology and how it affects the project's overall goals by clarifying the subtleties of this crucial stage.
This report is structured as follows:

This report is structured as follows

1. Introduction to the host institution and problem resolution.This chapter presents an overview of the host organization and outlines the problem addressed.
2. State of art:This chapter explores the theoretical foundations of the project.
3. The third chapter entitled "Methodology and integration" outlines the work methodology with relevant examples.

Chapter 1

Introduction to the host institution and problem resolution

1.1 Introduction

In this chapter, we present the general framework of the project, starting with the presentation of the host organization in which our work took place and where we will deal with experts in **the speech language pathology domain**.
For this, we will present **Salah Azeiz institute of Tunis**. Finally, we introduced the project's topic.

1.2 Salah Azeiz institute of Tunis

The Franco-Tunisian collaboration resulted in the founding of the National Institute of Oncology in 1969.
The institute was built next to the National Institute of Nutrition and the Ministry of Public Health, across from Bab Saadoun's historic gate [2]. The institution was renamed the ”Salah Azaïz Institute” in remembrance of the physician and surgeon Salah Azaïz (1911–1953).

Figure 1.1: Salah Azeiz institute

The ISA was originally divided into three departments: radiography, surgery, and anatomopathology. The institute was reorganized in 1989, adding a nuclear medicine department and purchasing radiotherapy equipment as part of the changes.
The International Union Against Cancer and the World Health Organization have designated the ISA as a regional reference center for the diagnosis and treatment of breast and cervical cancers, a recognition that is a result of the research efforts of numerous generations of foreign and Tunisian doctors and surgeons.
Since 1998, autologous peripheral stem cell transplantation has been a ground-breaking development in Tunisia's fight against cancer[2].

1.2.1 Who is Salah Azaïz

Born in Soliman on November 24, 1911, and passing away in Tunis on July 23, 1953, Salah Azaïz is regarded as the first surgeon in modern Tunisian history and the father of cancer surgery in the country [2].

Figure 1.2: Salah Azaïz

1.3 Problematic

As was already mentioned, Salah Azeiz Institute is well known for its skill in performing cancer surgeries, with a focus on **laryngectomy** procedures.

laryngectomy is the surgical procedure in which the larynx is totally removed and the airway is interrupted, respiration being performed through a tracheal stoma resulting from bringing the trachea to the skin in the lower, anterior, cervical area. This provides a complete and permanent separation of the superior part of the airway from the inferior one, resulting in voice and smell loss [4].

1.3.1 Larynx

In order of functional priority, the larynx shares 3 basic functions in airway protection, respiration, and phonation. Most importantly, the larynx protects the airway from swallowed matter through several mechanisms. **Figure 1.3**
It coordinates and optimizes the airway with respiration. Finally, the larynx provides controlled phonation, which, in conjunction with the pharynx, oral cavity, and nose, allows for detailed vocal communication [10].

Following this procedure, the patient can develop **dysphonia**, a disorder marked by difficulties speaking or an impairment in voice quality.

Consequently, they would seek assistance from a specialized healthcare professional known as a speech-language pathologist.

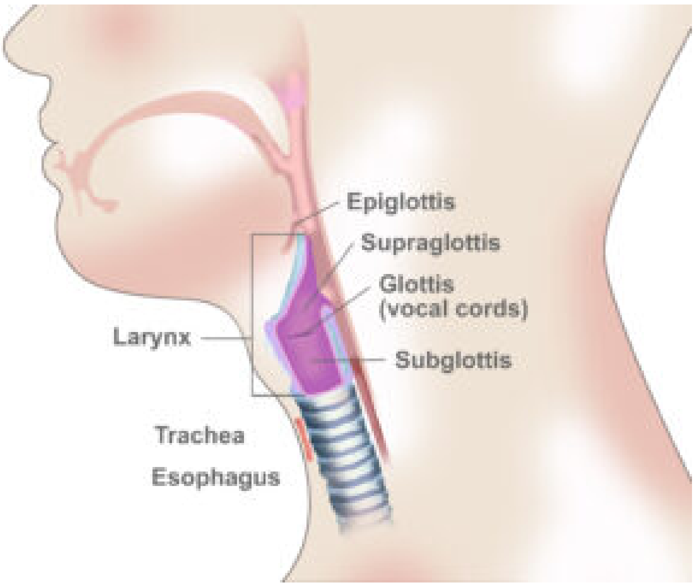


Figure 1.3: Larynx

Long sessions are often required in Tunisia to diagnose dysphonia, and even then, experts may find it difficult to assess whether a patient's condition is getting better.

1.3.2 Proposed Solution

Using machine learning and deep learning algorithms, the proposed solution is to create a model capable of extracting key features that signify the progress and improvement of patients over time.

These observations support professionals in keeping an eye on patients' improvements, but they also give patients a sense of agency by providing concrete proof of their development, which enhances their general wellbeing and self-assurance in their recuperation process.

Then we can integrate it in a website or mobile application.
As this project is complicated we devise it into small steps .the in this first step here we want to crate a model to classify normal people and dysphonia.

It’s a wise move to divide a large project into smaller, more achievable tasks in order to guarantee progress and success. During this first stage, we are concentrating on developing a model that can distinguish between speech patterns that are typical and those that indicate dysphonia.
This initial phase of the project establishes the framework for more complex analyses and advancements.

1.4 Conclusion

This chapter allowed us to present host organization as well as the overall background of this project. In the next chapter, we will represent the main concepts that we need to understand the problem of the project and the creation of the model.

Chapter 2

State of art

2.1 Introduction

This chapter defines some notions related to problem as well as the model.

2.2 What is Speech language pathologist

the SLP is called upon to provide rehabilitation services to those with varying neurological, oncological or other disease processes that may impact the person's communication, cognition and/or swallowing abilities.‘
It has been argued that patients with similar diagnoses who are nearing end of life (EOL) would also benefit from speech-language pathology input.
While such services would largely be governed by the physical, social and psychological status of the individual , nevertheless it has also been argued that within a palliative care approach, SLPs can utilise skills and strategies to ameliorate limitations in communication and/or swallowing status. This can help specifically to ensure comfort and in response to a person's change in health status [5].

Figure 2.1: Swallowing

2.2.1 What conditions do speech-language pathologists treat

Speech-language pathologists can treat:

- **Articulation disorders:**When speaking, people with articulation disorders find it difficult to use their muscles to make sounds correctly.
- **Cognitive-communication disorders:**When attention, memory, thought organization, or other brain functions related to information processing are compromised, people with cognitive-communication disorders find it difficult to communicate.
- **Phonological disorders:**Although the sounds produced by the muscles of those suffering from phonological disorders are correct, they do not conform to the standards of natural speech.
- **Social communication disorders:** Individuals who suffer from a social communication disorder find it difficult to interact with others. They might have trouble reading social cues, which makes it difficult for them to relate to people through speech and body language.
- **Swallowing disorders (dysphagia):** Dysphagia patients struggle to regulate their muscles and other body parts, which makes it difficult for them to safely swallow food and liquids.
- **Voice disorders (dysphonia):** Your vocal cords may be affected by a variety of conditions, which can make it difficult to produce sounds. Vocal cord paralysis, lesions, and dysfunction are examples of voice disorders.

Our project is related to dysphonia problem then let is unerdstand it.

2.3 What is Dysphonia

Voice evaluation requires assessment of the respiratory system, the larynx, and the resonance capabilities of the upper airway. Normal phonation requires adequate breath support, approximation of the vocal folds, vocal fold pliability, and control of vocal fold length and tension [9]. Impairment of these the patient feels as disorder in his voice like Trouble with the voice when trying to talk, including hoarseness and change in pitch or quality or voice it is named Dysphonia.

The public health impact of vocal dysfunction is becoming increasingly recognized. Dysphonia adversely impacts communication, with physical, social, and workrelated effects. Patients experience social isolation, depression, impaired disease-specific and general quality of life, and work absenteeism.1–4 Therefore, voice disorders negatively impact individuals and burden society[6].

2.3.1 Causes of dysphonia

Voice impairment can have a variety of causes. They are separated based on where the disorder first appeared.

Functional dysphonia: connected to a voice gesture disruption.

This kind of dysphonia can occur in situations where professionals (teachers, singers, salespeople, telephone operators, etc.) use their voices a lot, in stressful situations, or when there are age-related changes to their voice.

Organic dysphonia: are connected to injuries to the vocal cords. These blemishes could be malignant or benign.

2.3.2 Treatement of dysphonia

Treatment involves speech therapy administered by a speech therapist or voice hygiene advice (hydration, vocal rest) it depends on the cause.

2.4 Artificial intelligence

Artificial intelligence (IA) where it applies Machine Learning (ML) and Deep Learning (DL) methodologies, we will explore more in detail these three fundamental components of AI.

Artificial Intelligence (AI) is revolutionizing various industries and transforming the way we live and work IA is an intelligent devices that behave like humans.The benefit of intelligent machine is that can complete tasks faster and more efficiently than humans.

Artificiel intelligence is a large topic that includes ML and DL.

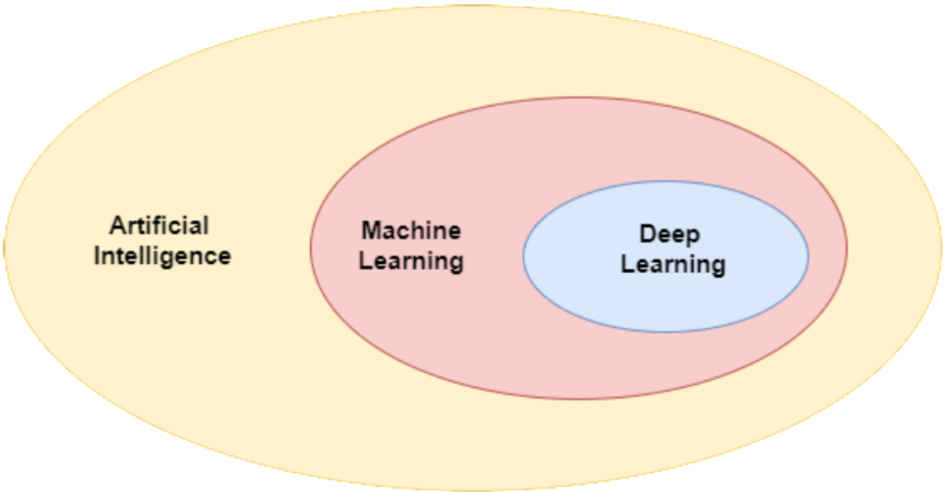


Figure 2.2: Artificial intelligence

2.5 Machine learning

ML is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

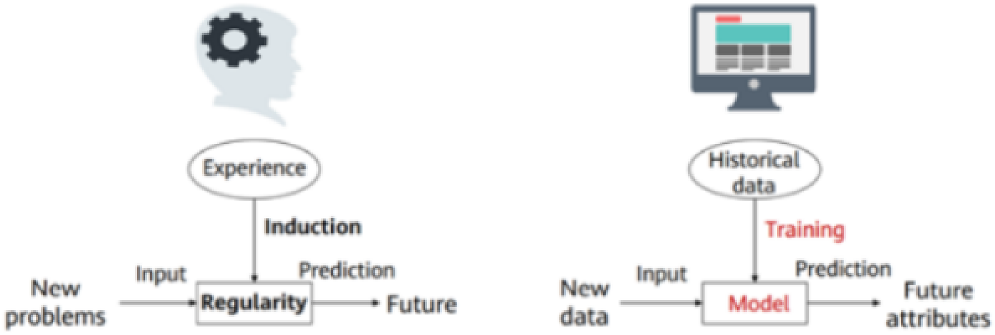


Figure 2.3: difference between machine learning and human learning

Human learning is a complex process that begins with sensory input (which includes information received through our five senses) and involves the integration of new information with past knowledge and experience. However, ML algorithms are often designed to automatically process massive volumes of data and discover patterns and relationships.

Machine learning and human learning both entail the acquisition of knowledge and abilities via experience. They differ in their learning process, creativity, and adaptability

Learning process: human learning needs to acquire new knowledge, skills, or behaviors through experience, study, or instruction, however, processes large amounts of data.

Adaptability: humans are very flexible and may quickly learn new abilities, in contrast, machine learning algorithms are often intended to accomplish specific tasks and may fail to adapt to new or unexpected conditions.

Creativity: humans are capable of creative thinking and invention, while, machine learning algorithms can recognize patterns and make predictions based on existing data.

2.5.1 Types of machine learning

Supervised learning: the algorithm is trained and learns to make predictions on labeled data and can then be used to make predictions on new unseen data.

Unsupervised learning: the algorithm is trained on unlabeled data and must identify patterns and relationships within the data on its own.

Reinforcement learning: learning training strategy that rewards desired behaviors while penalizing undesirable ones. A reinforcement learning agent, in general, can detect and comprehend its surroundings, act, and learn through trial and error.

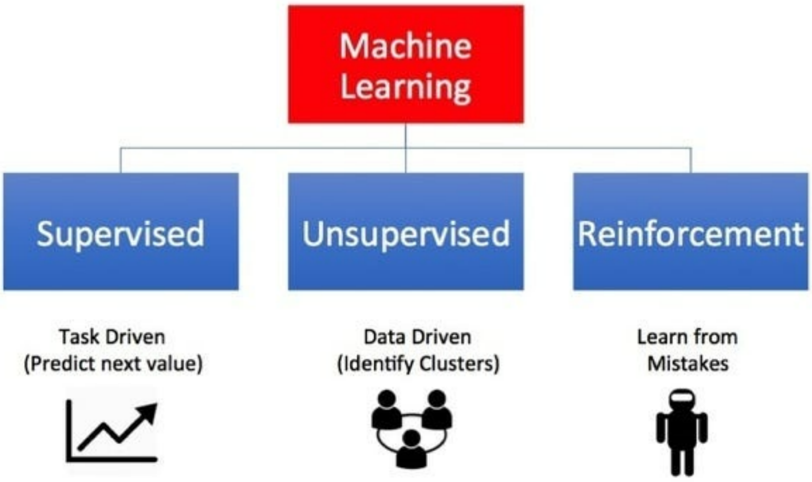


Figure 2.4: Types of machine learning

2.5.2 Neural network

The neural network is a model inspired by the structure and function of the human brain. Like the neurons in our brain, the circles above represent a node. Green circles represent the input layer (input layers), red circles represent hidden layers or intermediate layers and the blue circle represents the output layer. That work together to process information.

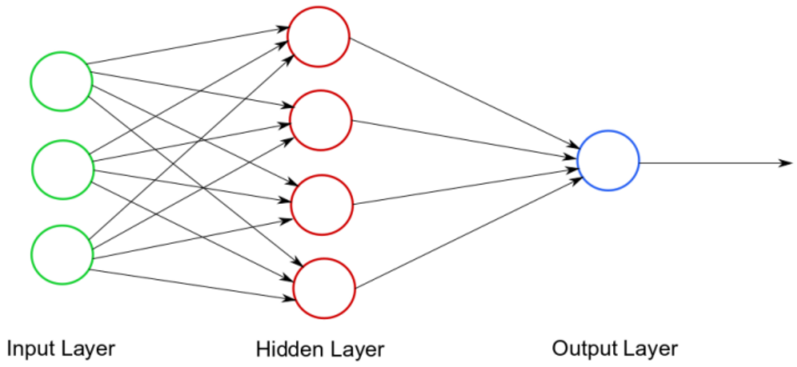


Figure 2.5: Neural Network and its functionality

Each neuron takes input from other neurons and uses this information to produce an output value. The output of each neuron is then passed on to other neurons on the network, allowing information to be processed and sent throughout the network.

Neural networks have been effectively used in a variety of applications, including image recognition, audio recognition, natural language preprocessing,autonomous vehicle...

2.5.3 Deep learning

DL is a subset of machine learning, which is essentially a neural network with three or more layers.

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structures in large data sets by using the back propagation algorithm to indicate how a machine should change its internal parameters These parameters are used to compute the representation in each layer based on the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shone a light on sequential data such as text and speech[8].

2.6 Signal (Md Balqine cours)

A signal is a feature that carries information about the state or behavior of a signal .

2.6.1 Signal's type (Md Balqine cours)

-Monodimention:one-dimensional signals capture information along a single dimension, usually time, and are used in many fields, including communication, medicine, and environmental sciences.
-Bidimention two-dimensional signals capture information in two dimensions and are used in many areas, including computer vision, mapping,

2.6.2 Signal processing (Md Balqine cours)

A discipline that develops and studies the techniques of signal processing, analysis and processing in order to extract the maximum of useful information.

2.6.3 Deterministic signal and random signal (Md Balqine cours)

A deterministic signal is a signal represented by a temporal function the sine signal, square... however aleatory signal is decreed by its statistical properties.

2.6.4 Voice signal

Vocal cord vibrations produced by speech or singing cause changes in air pressure. These oscillations travel through the atmosphere as sound waves, which microphones can pick up and transform into electrical signals for examination.
In disciplines like speech pathology, linguistics, and communication studies, voice signal analysis is a useful tool because these features can shed light on the nature, health, and qualities of the voice.

2.7 Conclusion

In this chapter, we discussed the basic concepts necessary for the development of our project as AI, ML, DL and the signal Next chapter, we will explain more clearly about the methodologyto create the moedel.

Chapter 3

Methedology

3.1 Introducion

In this chapter, we will present the working environment as well as focus on the methodology of work and we will explain with examples through our project.

3.2 Working environment

In this section, we will present the hardware configuration and software environment used.

3.2.1 Hardware configuration

During this project we used the following machine configuration:

- **Brand:** DELL.
- **RAM Memory:** 16 Go.
- **Processor:** Intel Core i5-1135G7 processor, up to 4.2 GHz, 8 MB cache.
- **Graphics card:** Nvidea GeForce MX350, 2 GB dedicated memory.
- **Hard disk :** SSD 256 Go

.

- **Overleaf**

Overleaf is a free online platform allowing you to edit LATEX text without any application download. In addition, it offers the possibility to write documents in a collaborative way, to offer its documents directly to different editors (IEEE Journal, Springer, etc.) or open archive platforms (arXiv, engrxiv, etc.) for a possible publication.

- **Colab**

Colab is a hosted Jupyter Notebook service that requires no setup to use and provides free access to computing resources, including GPUs and TPUs. Colab is especially well suited to machine learning, data science, and education.

- **Kaggle**

Kaggle is an interactive web platform that offers machine learning competitions in data science. The platform provides free datasets, notebooks and tutorials that data scientists need to complete their machine learning projects.

3.2.2 Development language

Python

Python is the open source programming language most used by informants. It is used in machine learning and data science, the Python language also imposes itself in other sectors of activity thanks to its simplicity and compatibility.

Labraries

One of the great strengths of Python is that it contains a very large number of biblio-libraries:

- **Pandas:** pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.
- **librosa** iis a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems.
- **matplotlib** Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.
- **seaborn** is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
- **os** This module provides a portable way of using operating system dependent functionality. If you just want to read or write a file , if you want to manipulate paths,...
- **NumPy** is a library for Python programming language, intended to manipulate matrices or multidimensional arrays as well as mathematical functions operating on these arrays.

- **Scikit-Learn** Scikit-learn is a free Python library for machine learning. It is developed by numerous contributors, notably in the academic world by French higher education and research institutes such as Inria.

3.3 Methodology of work

3.3.1 Understanding the business

The CRISP-DM (CRoss Industry Standard Process for Data Mining) project proposed a comprehensive process model for carrying out data mining projects. The process model is independent of both the industry sector and the technology used[11].

CRISP–DM is a hierarchical process consisting of 5 main steps and a final optional step as shown in Figure 3.1. We focus on meeting the special needs of people with dysphonia in the field of healthcare, particularly in the speech-language pathology sector.

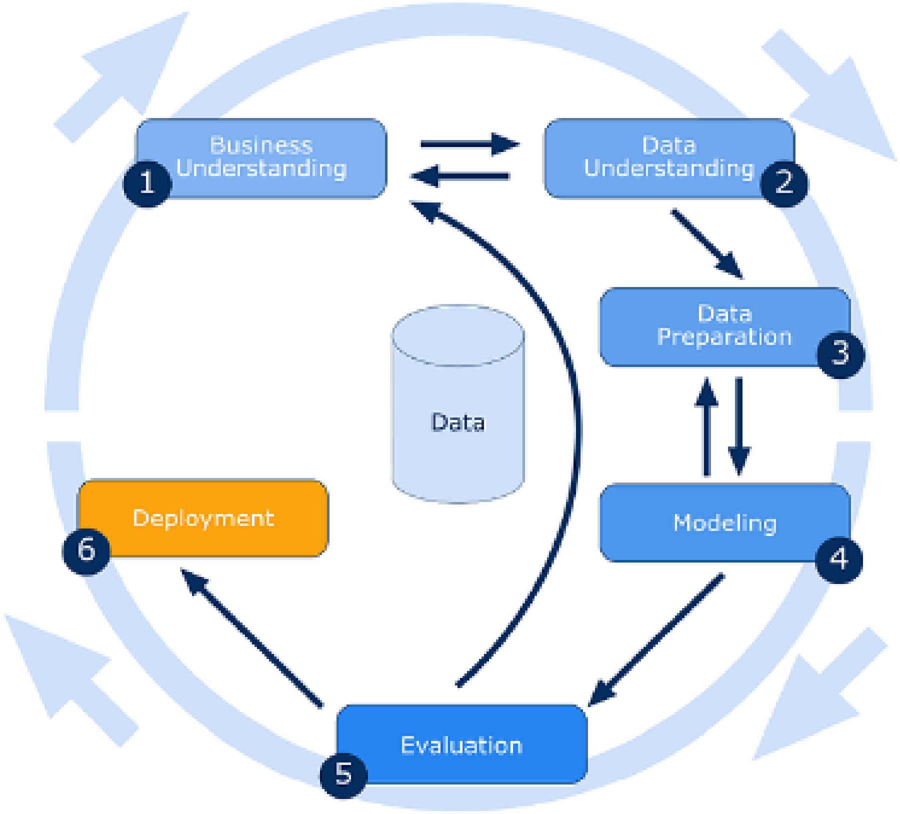


Figure 3.1: Processus CRISP-DM

Our goals are to guarantee the provision of high-quality care and to enhance communication accessibility.

3.3.2 Data collection & Understanding data

We employed the Saarbrucken Voice Database, which is open to the public [1].
It is an assortment of voice recordings in which samples from one subject are recorded producing the vowels /a/, /i/, and /u/ at normal, high and low pitches.

3.3.3 Data Preparation

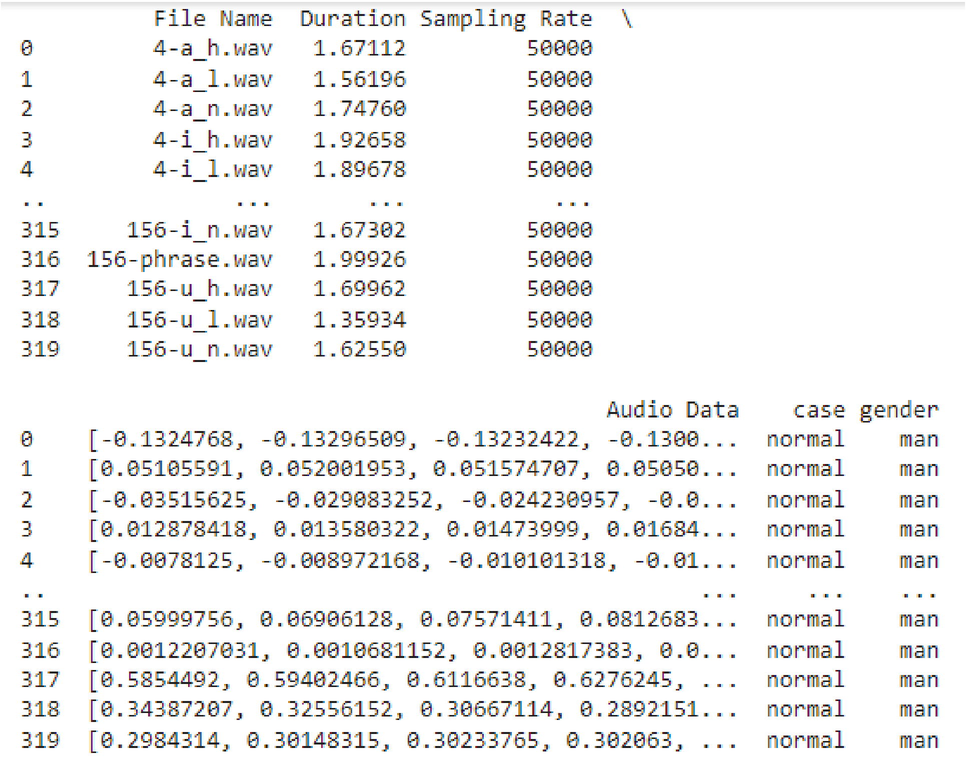
First, we load normal male and female audio files separately **Figure 3.2**. Then, we create a dataframe containing the following information for each audio file :

File name: This column stores the name of the audio file.

Duration: The audio file's duration is listed in this column as seconds. It is computed by dividing the audio data's length by the sampling rate.

Sample Rating: The number of audio samples carried per second, expressed in Hertz (Hz), is known as the sampling rate.

Audio loading: represent the audio content loaded from each file.

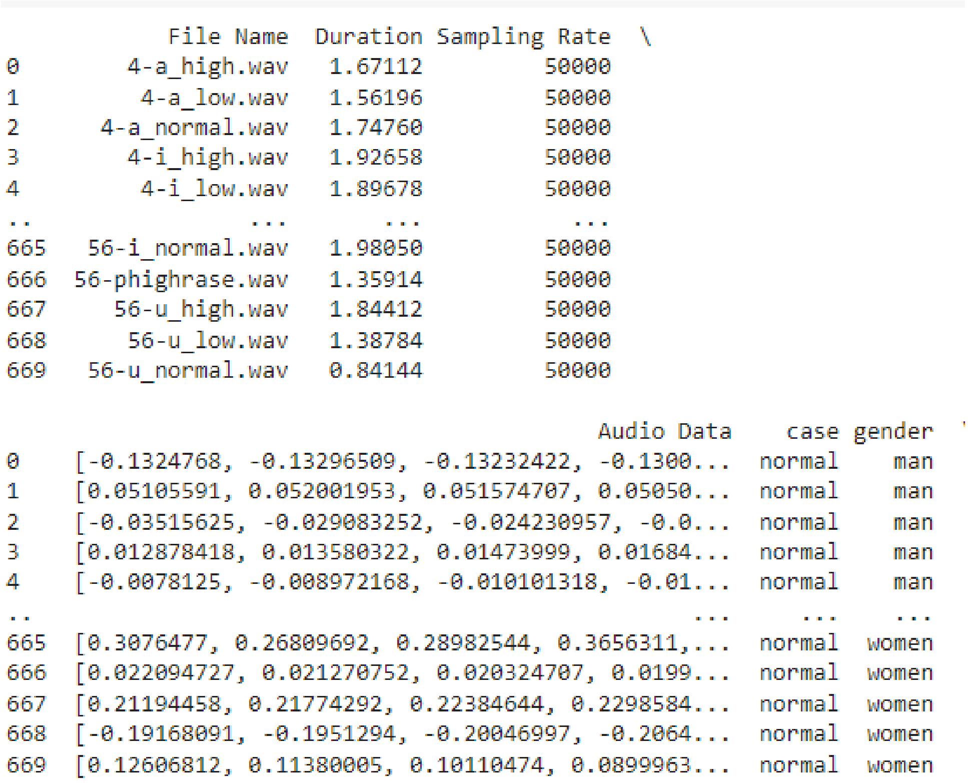


```
          File Name  Duration  Sampling Rate  \
0         4-a_h.wav   1.67112          50000
1         4-a_l.wav   1.56196          50000
2         4-a_n.wav   1.74760          50000
3         4-i_h.wav   1.92658          50000
4         4-i_l.wav   1.89678          50000
..              ...       ...            ...
315     156-i_n.wav   1.67302          50000
316  156-phrase.wav   1.99926          50000
317     156-u_h.wav   1.69962          50000
318     156-u_l.wav   1.35934          50000
319     156-u_n.wav   1.62550          50000

                                            Audio Data    case gender
0    [-0.1324768, -0.13296509, -0.13232422, -0.1300...  normal    man
1    [0.05105591, 0.052001953, 0.051574707, 0.05050...  normal    man
2    [-0.03515625, -0.029083252, -0.024230957, -0.0...  normal    man
3    [0.012878418, 0.013580322, 0.01473999, 0.01684...  normal    man
4    [-0.0078125, -0.008972168, -0.010101318, -0.01...  normal    man
..                                                 ...     ...    ...
315  [0.05999756, 0.06906128, 0.07571411, 0.0812683...  normal    man
316  [0.0012207031, 0.0010681152, 0.0012817383, 0.0...  normal    man
317  [0.5854492, 0.59402466, 0.6116638, 0.6276245, ...  normal    man
318  [0.34387207, 0.32556152, 0.30667114, 0.2892151...  normal    man
319  [0.2984314, 0.30148315, 0.30233765, 0.302063, ...  normal    man
```

Figure 3.2: Exemple: Normal people dataframe

Next, we add a column named ”Gender” to the dataframe. This column will contain the gender label associated with each audio file, with ”man” for entries in the normal male dataframe and ”woman” for entries in the normal female dataframe. **Figure 3.3**

After adding the ”Gender” column to the dataframe, we ensure that the two dataframes do not contain duplicate audio files.



```
            File Name  Duration Sampling Rate  \
0        4-a_high.wav   1.67112         50000
1         4-a_low.wav   1.56196         50000
2      4-a_normal.wav   1.74760         50000
3        4-i_high.wav   1.92658         50000
4         4-i_low.wav   1.89678         50000
..                ...       ...           ...
665    56-i_normal.wav  1.98050         50000
666  56-phighrase.wav   1.35914         50000
667      56-u_high.wav  1.84412         50000
668       56-u_low.wav  1.38784         50000
669    56-u_normal.wav  0.84144         50000

                                            Audio Data    case gender
0    [-0.1324768, -0.13296509, -0.13232422, -0.1300...  normal    man
1    [0.05105591, 0.052001953, 0.051574707, 0.05050...  normal    man
2    [-0.03515625, -0.029083252, -0.024230957, -0.0...  normal    man
3    [0.012878418, 0.013580322, 0.01473999, 0.01684...  normal    man
4    [-0.0078125, -0.008972168, -0.010101318, -0.01...  normal    man
..                                                 ...     ...    ...
665  [0.3076477, 0.26809692, 0.28982544, 0.3656311,...  normal  women
666  [0.022094727, 0.021270752, 0.020324707, 0.0199...  normal  women
667  [0.21194458, 0.21774292, 0.22384644, 0.2298584...  normal  women
668  [-0.19168091, -0.1951294, -0.20046997, -0.2064...  normal  women
669  [0.12606812, 0.11380005, 0.10110474, 0.0899963...  normal  women
```

Figure 3.3: Joing the datasets and adding gender’s colunm

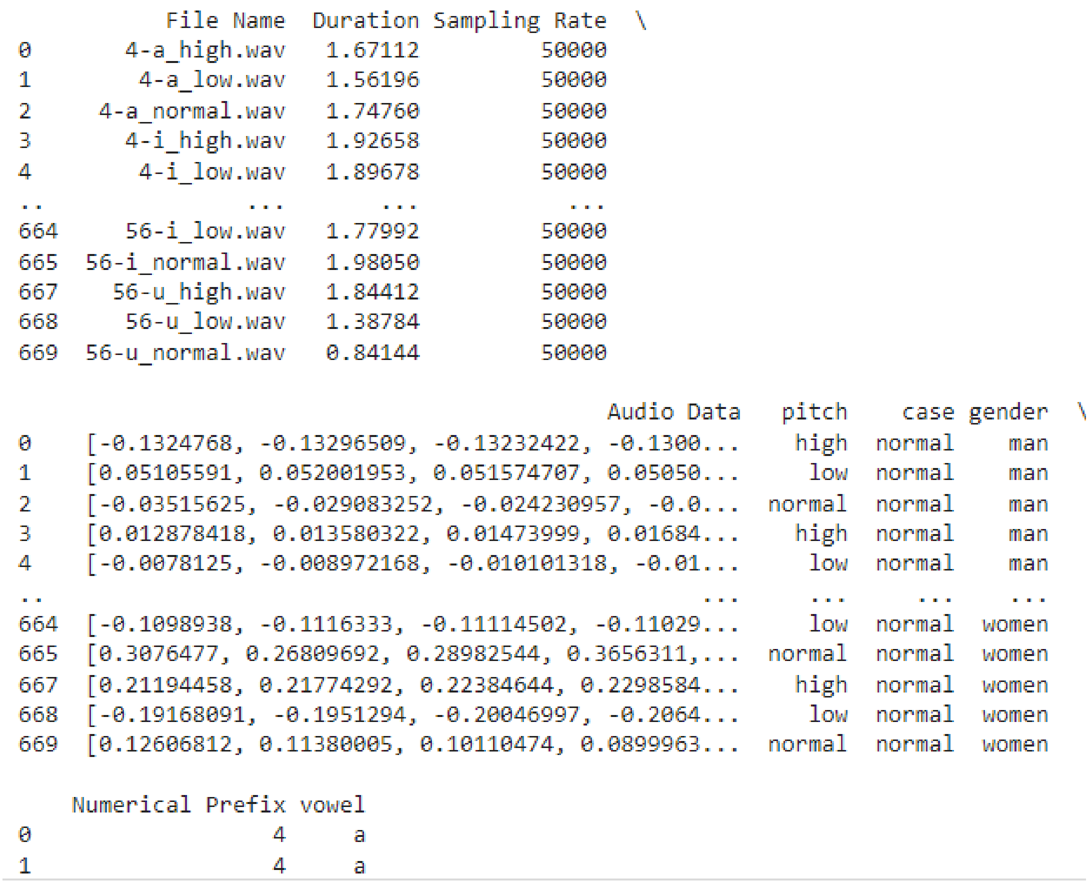
Next, we join the two dataframes and modify the filenames to replace ’h’, ’l’, and ’n’ with ’high’, ’low’, and ’normal’, respectively.
Additionally, we remove audio files that contain phrases instead of the pronunciation of vowels.
Furthermore, three columns are added to the dataframe :

- The first column indicates the pronounced vowel.

- The second column represents the pitch of pronunciation.

- The third column determines whether the audio is classified as normal or dysphonia.

The result in **Figure3.4**



```
          File Name  Duration Sampling Rate  \
0       4-a_high.wav   1.67112         50000
1        4-a_low.wav   1.56196         50000
2     4-a_normal.wav   1.74760         50000
3       4-i_high.wav   1.92658         50000
4        4-i_low.wav   1.89678         50000
..               ...       ...           ...
664     56-i_low.wav   1.77992         50000
665  56-i_normal.wav   1.98050         50000
667    56-u_high.wav   1.84412         50000
668     56-u_low.wav   1.38784         50000
669  56-u_normal.wav   0.84144         50000

                                            Audio Data   pitch    case gender  \
0    [-0.1324768, -0.13296509, -0.13232422, -0.1300...    high  normal    man
1    [0.05105591, 0.052001953, 0.051574707, 0.05050...     low  normal    man
2    [-0.03515625, -0.029083252, -0.024230957, -0.0...  normal  normal    man
3    [0.012878418, 0.013580322, 0.01473999, 0.01684...    high  normal    man
4    [-0.0078125, -0.008972168, -0.010101318, -0.01...     low  normal    man
..                                                 ...     ...     ...    ...
664  [-0.1098938, -0.1116333, -0.11114502, -0.11029...     low  normal  women
665  [0.3076477, 0.26809692, 0.28982544, 0.3656311,...  normal  normal  women
667  [0.21194458, 0.21774292, 0.22384644, 0.2298584...    high  normal  women
668  [-0.19168091, -0.1951294, -0.20046997, -0.2064...     low  normal  women
669  [0.12606812, 0.11380005, 0.10110474, 0.0899963...  normal  normal  women

    Numerical Prefix vowel
0                  4     a
1                  4     a
```

Figure 3.4: Joining the datasets and adding pitch and vowel's column

Next, we load dysphonia male and female audio files and create a dataframe containing the following information for each audio file:

File name: This column stores the name of the audio file.

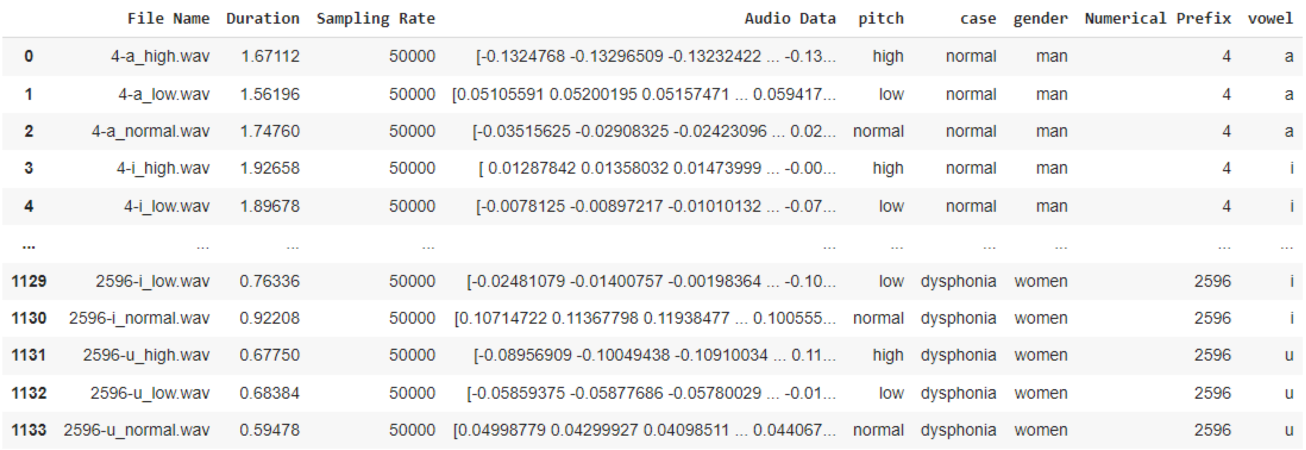
Duration: The duration of the audio file, listed in seconds. It is computed by dividing the length of the audio data by the sampling rate.

Sample Rate: The number of audio samples carried per second, expressed in Hertz (Hz), known as the sampling rate.

Audio loading: This column represents the audio content loaded from each file. The same process is applied as for normal audio files.

The resulting dataframe is presented as **Figure 3.4.**

the last step is downloading the final dataset **Figure 3.5** to use lately for data visualisation.



	File Name	Duration	Sampling Rate	Audio Data	pitch	case	gender	Numerical Prefix	vowel
0	4-a_high.wav	1.67112	50000	[-0.1324768 -0.13296509 -0.13232422 ... -0.13...	high	normal	man	4	a
1	4-a_low.wav	1.56196	50000	[0.05105591 0.05200195 0.05157471 ... 0.059417...	low	normal	man	4	a
2	4-a_normal.wav	1.74760	50000	[-0.03515625 -0.02908325 -0.02423096 ... 0.02...	normal	normal	man	4	a
3	4-i_high.wav	1.92658	50000	[0.01287842 0.01358032 0.01473999 ... -0.00...	high	normal	man	4	i
4	4-i_low.wav	1.89678	50000	[-0.0078125 -0.00897217 -0.01010132 ... -0.07...	low	normal	man	4	i
...	...	...	...	...	...	...	...	...	...
1129	2596-i_low.wav	0.76336	50000	[-0.02481079 -0.01400757 -0.00198364 ... -0.10...	low	dysphonia	women	2596	i
1130	2596-i_normal.wav	0.92208	50000	[0.10714722 0.11367798 0.11938477 ... 0.100555...	normal	dysphonia	women	2596	i
1131	2596-u_high.wav	0.67750	50000	[-0.08956909 -0.10049438 -0.10910034 ... 0.11...	high	dysphonia	women	2596	u
1132	2596-u_low.wav	0.68384	50000	[-0.05859375 -0.05877686 -0.05780029 ... -0.01...	low	dysphonia	women	2596	u
1133	2596-u_normal.wav	0.59478	50000	[0.04998779 0.04299927 0.04098511 ... 0.044067...	normal	dysphonia	women	2596	u

Figure 3.5: Final dataset

3.3.4 Data visualization

Histogram

Histogram is a visual representation of the distribution of data.

Observing **Figure 3.10**, we can discern that the distribution of vowels is equal between normal and dysphonia individuals, as well as the distribution of pitch and gender. However, it's notable that the distribution of normal individuals is higher compared to individuals with dysphonia.

Boxplot

A box plot or boxplot is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their quartiles.

We plot boxplots of the duration of normal individuals for each vowel with its pitch. Upon observation **Figure 3.11**
we find that the maximum durations are as follows: **Table 3.1,Table 3.2,Table 3.3**

Also we plot boxplots of the duration of dysphonia individuals for each vowel with its pitch. Upon observation **Figure 3.12**, we find that the maximum durations are as follows: **Table 3.4,Table 3.5,Table 3.6**

3.3.5 frequency spectrum

The frequency spectrum can provide valuable insights into the characteristics of the voice, such as pitch, formants, and harmonic content.

We plotted the frequency of a normal man articulating the alphabet 'a' with normal pitch in **Figure 3.13**, and the frequency of a dysphonic man articulating the alphabet 'a' with normal pitch in **Figure 3.14**.

Additionally, we plotted the frequency of a normal woman articulating the alphabet ’a’ with normal pitch in **Figure 3.15**, and the frequency of a dysphonic woman articulating the alphabet ’a’ with normal pitch in **Figure 3.16**.

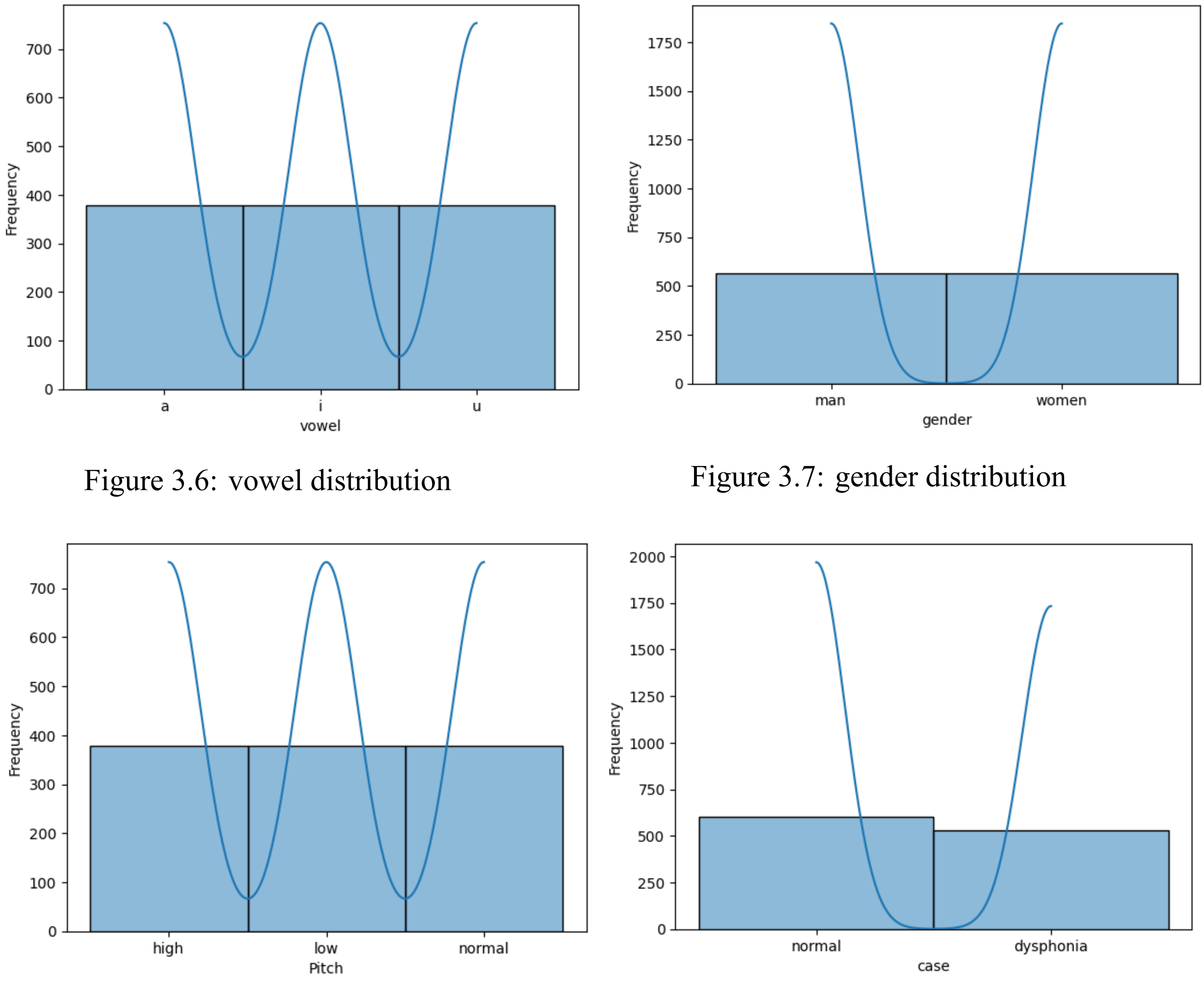


Figure 3.6: vowel distribution

Figure 3.7: gender distribution

Figure 3.8: Pitch distribution

Figure 3.9: Case distribution

Figure 3.10: Dataset’s features distribution

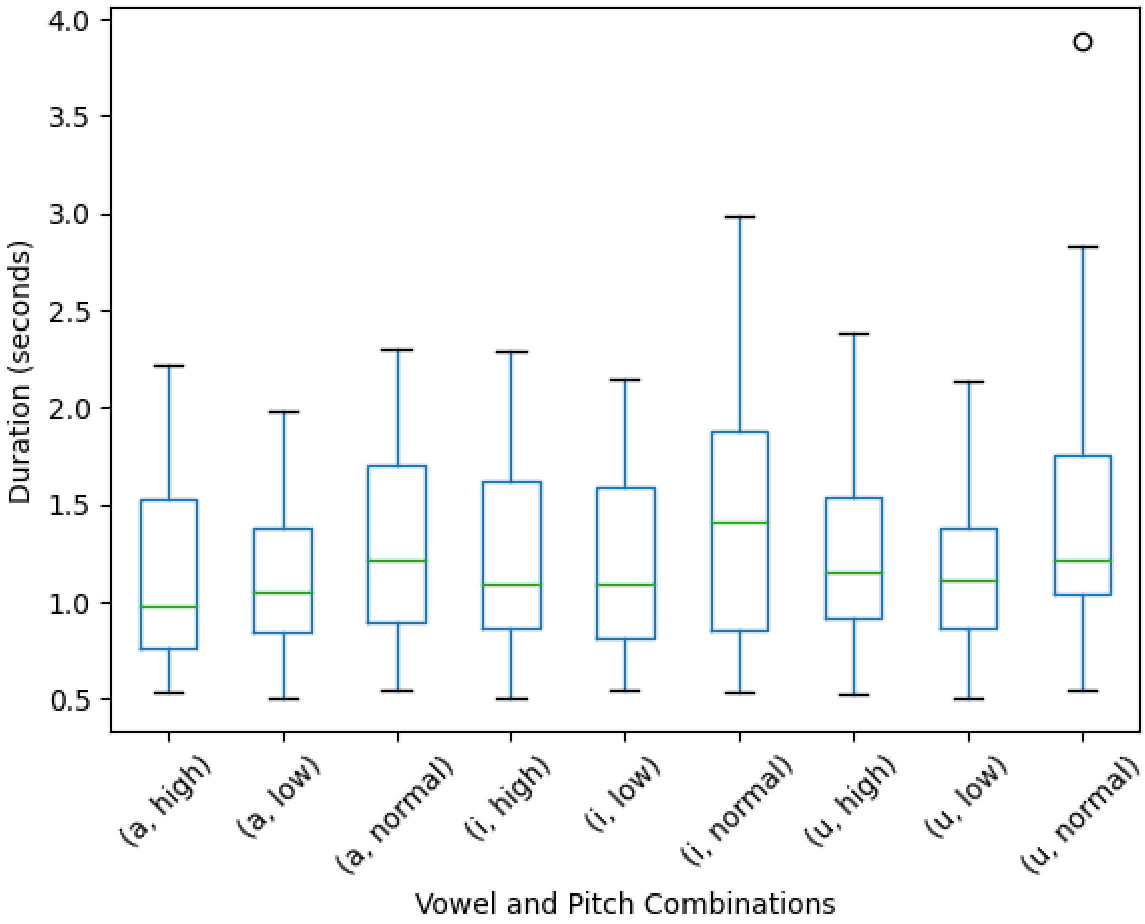


Figure 3.11: Boxplot of vowel and pitch combination distribution for normal people

Vowels	(a, high)	(i, high)	(u, high)
Maximum duration	2.3 seconds	2.3 seconds	2.3 seconds

Table 3.1: Vowel observation with high pitch for normal people

Vowels	(a, low)	(i, low)	(u, low)
Maximum duration	2.3 seconds	2.3 seconds	2.3 seconds

Table 3.2: Vowel observation for low pitch normal people

Vowels	(a, normal)	(i, normal)	(u, normal)
Maximum duration	2.4 seconds	3 seconds	2.7 seconds

Table 3.3: Vowel observation with normal pitch for normal people

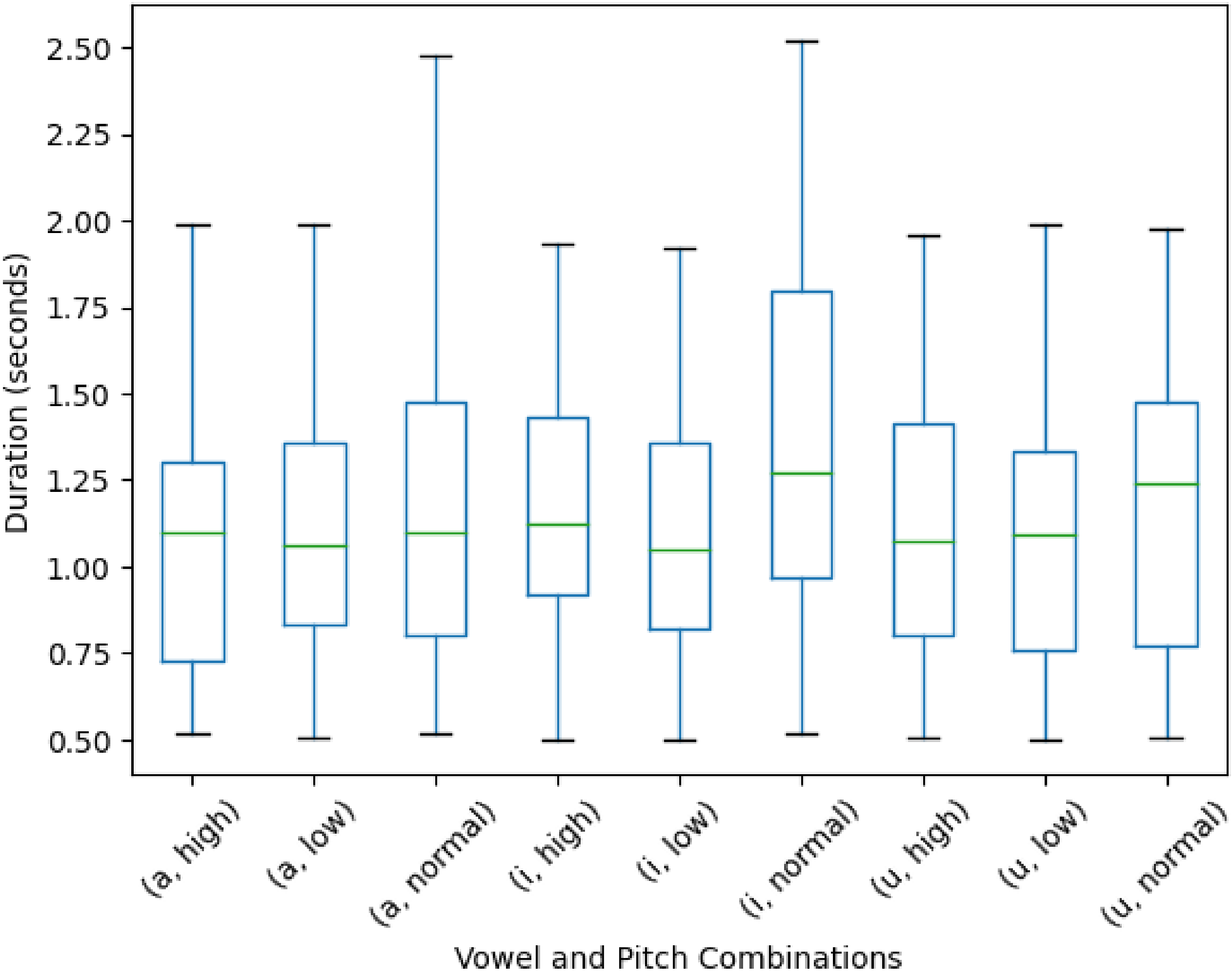


Figure 3.12: Boxplot of vowel and pitch combination distribution for dysphonia people

Vowels	(a, high)	(i, high)	(u, high)
Maximum duration	2.3 seconds	2.3 seconds	2.3 seconds

Table 3.4: Vowel observation with high pitch for dysphonia people

Vowels	(a, low)	(i, low)	(u, low)
Maximum duration	2.3 seconds	2.3 seconds	2.3 seconds

Table 3.5: Vowel observation with low pitch for dysphonia people

Vowels	(a, normal)	(i, normal)	(u, normal)
Maximum duration	2.4 seconds	3 seconds	2.7 seconds

Table 3.6: Vowel observation with normal pitch for dysphonia people

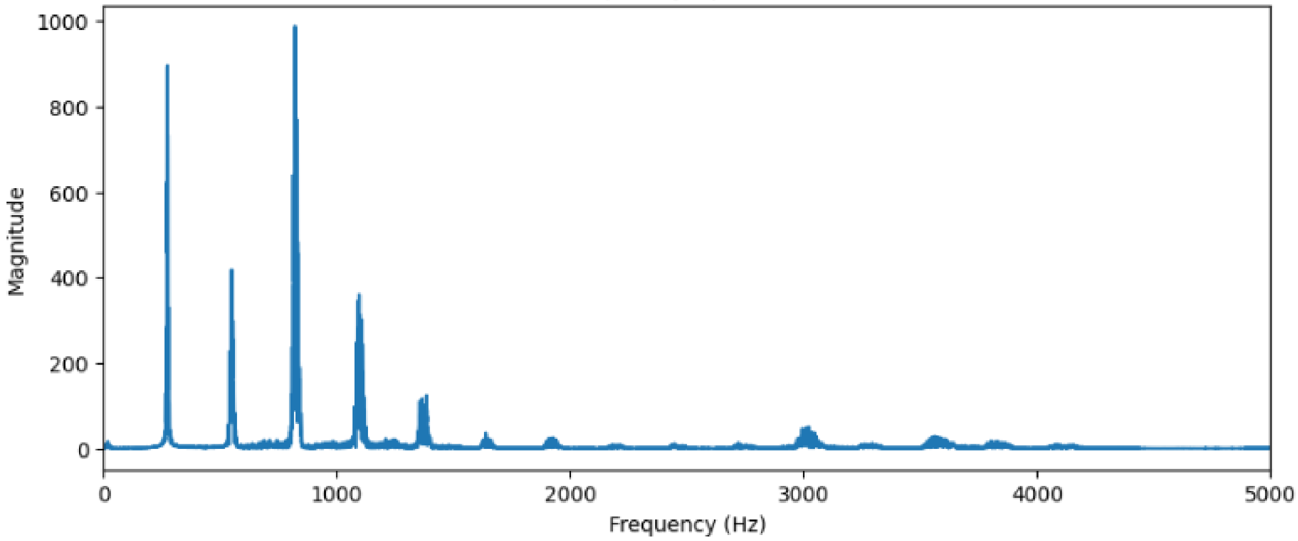


Figure 3.13: Normal male spectrum frequency for vowel ’a’ with normal pitch

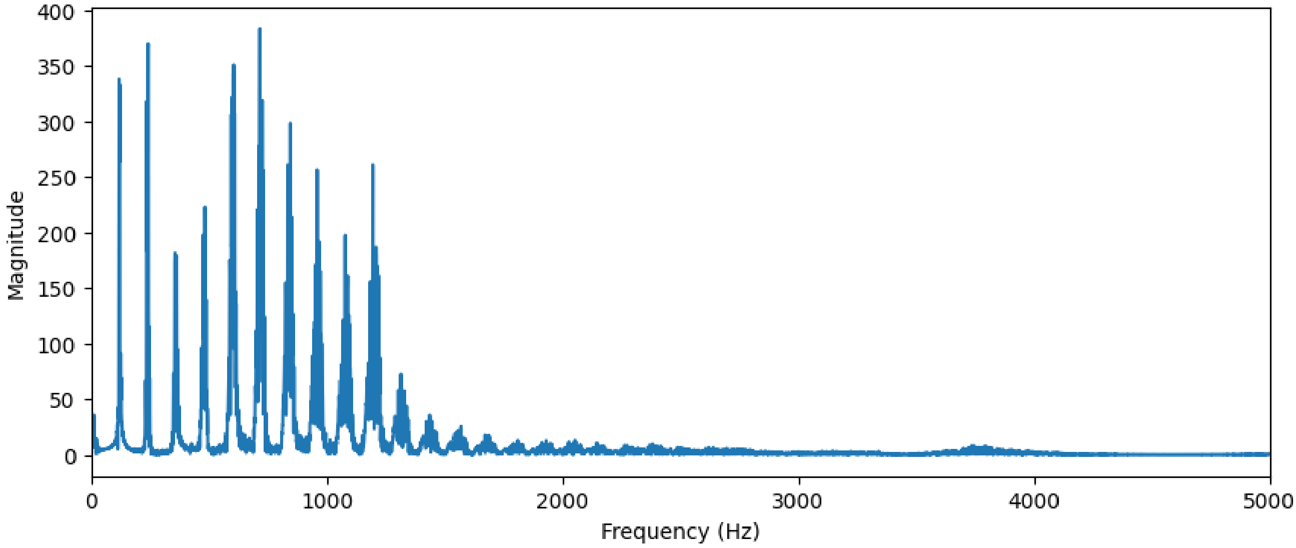


Figure 3.14: Dysphonia male spectrum frequency for vowel ’a’ with normal pitch

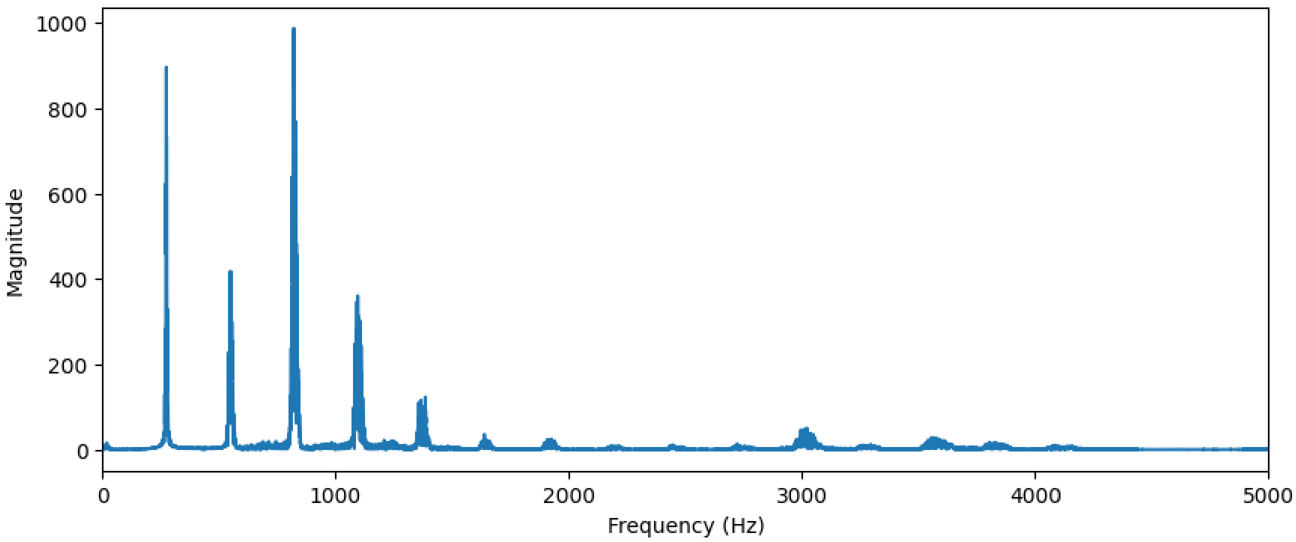


Figure 3.15: Normal female spectrum frequency for vowel ’a’ with normal pitch

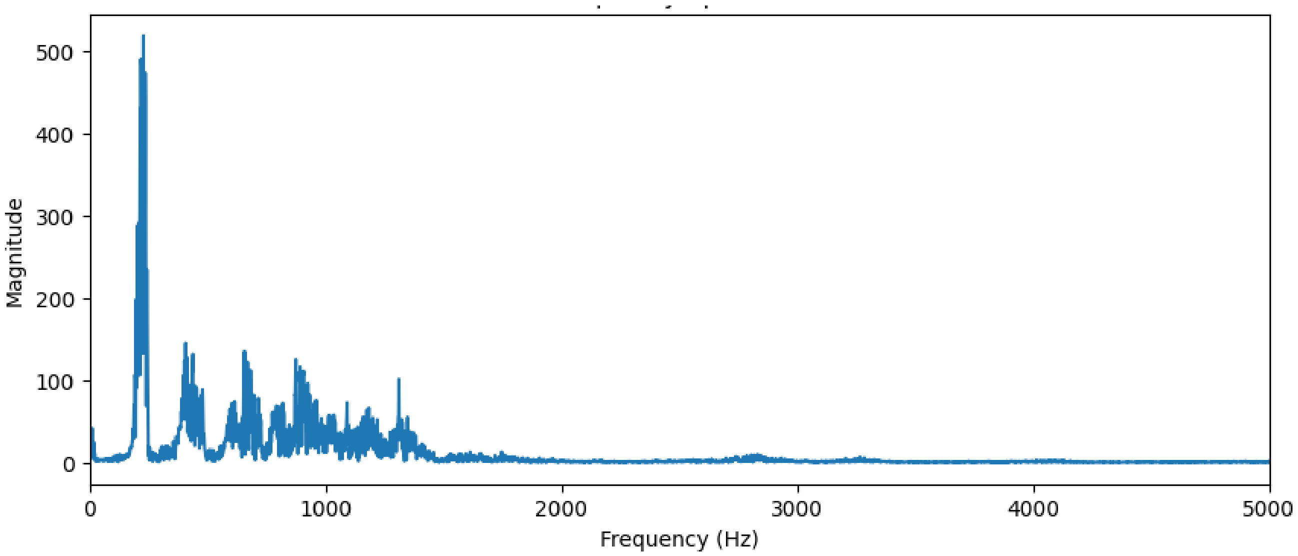


Figure 3.16: Dysphonia female spectrum frequency for vowel ’a’ with normal pitch

3.3.6 Feature extraction

In order to distinguish between various audio classes, such as distinct speakers, emotions, or speech disorders like dysphonia, feature extraction for voice classification entails finding and extracting pertinent characteristics from voice data **Figure 3.17**.

The following are some typical characteristics of voice classification:

the mel-frequency cepstrum (MFC) s designed to model features of audio signal and is widely used in various fields. This paper aims to review the applications that the MFCC is used for in addition to some issues that facing the MFCC computation and its impact on the model performance [3].
These coefficients capture important information about the spectral shape of the voice signal, including information about its timbre and overall spectral envelope.

Spectral Centroid is an estimate of the 'centre of gravity' of the spectrum within each subband. Originally proposed as a feature for speech recognition systems (Paliwal, 1998), it has been reported that SCF is a formant-like feature, as it provides the approximate location of the formant frequencies in the subbands (Paliwal, 1998) [7].
This feature can provide insights into the overall spectral distribution of energy in the signal, with higher values indicating a higher concentration of energy towards higher frequencies and vice versa

Zero Crossing Rate is a representation of the short-term power spectrum of a soundis the sign change rate of a signal.
This feature can provide information about the temporal characteristics of the signal, such as its rhythmic patterns or the presence of voiced or unvoiced segments.

F0 mean This feature can provide information about the intonation and pitch variability of the voice, which are important cues for tasks such as speaker identification, emotion recognition, and speech analysis.

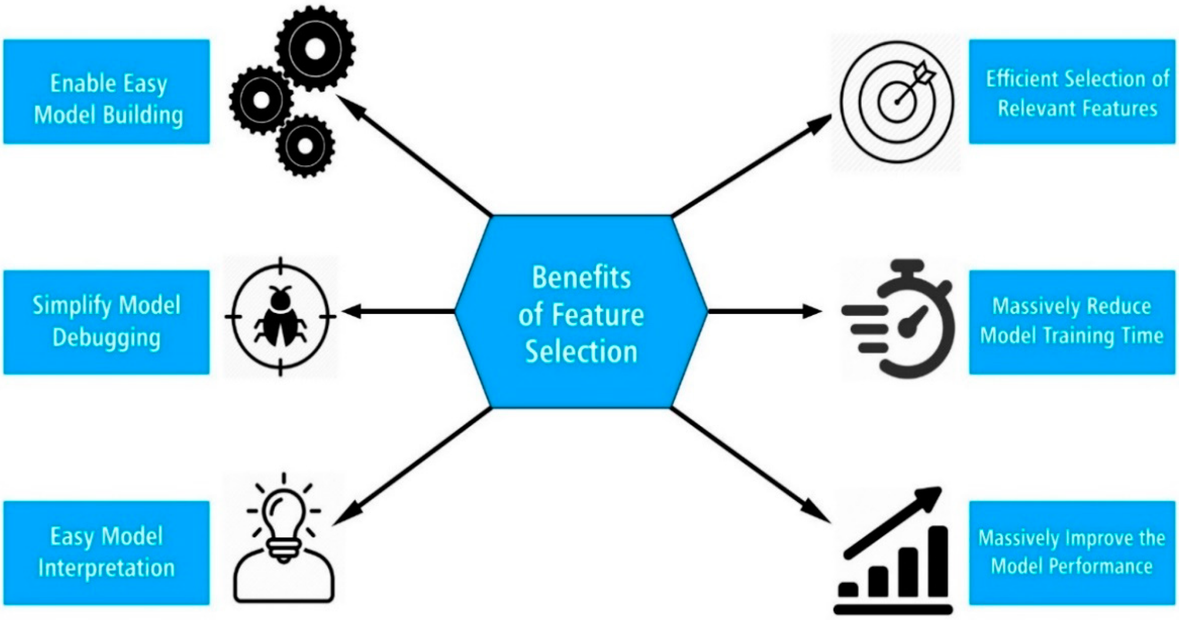

Enable Easy Model Building
Efficient Selection of Relevant Features
Simplify Model Debugging
Benefits of Feature Selection
Massively Reduce Model Training Time
Easy Model Interpretation
Massively Improve the Model Performance

Figure 3.17: Benefits of feature extraction

3.3.7 Modelling

For the modelling we test two types of models SVM model, Random Forest model and KNN model

- **Random forest:** Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees..

- **KNN (K-Nearest Neighbors) :** is a simple and intuitive machine learning algorithm used for classification and regression tasks. It classifies a new data point by comparing it to the labeled data points in its vicinity (nearest neighbors).

3.3.8 Model evaluation

We will now explore the evaluation measures used to evaluate our models

- **Accuracy:** Accuracy is a metric that measures the overall correctness of the predictions made by the model.
  The higher the accuracy, the better the model's ability to make correct predictions.
- **precision:** (Positive Predictive Value): The proportion of true positive results among the total predicted positives. It answers the question.
- **Recall:** Sensitivity or True Positive Rate): The proportion of true positive results among the total actual positives.

Experimental results

We have experienced two different machine learning models: FNN and LSTM
The metrics used to evaluate the models are accuracy precision and recall.

But, Here we compare the KNN and Random forest.

- **KNN:**
  The evaluation results of the scores:
  Accuracy : **Figure 1.18**
  Precision and recall: **Figure 1.18**
- **Random Forest:**
  The evaluation results of the scores:
  Accuracy : **Figure 1.18**
  Precision and recall: **Figure 1.18**

Results and discussion

Depend on previous results:
Accuracy Comparison:
Random Forest has the highest accuracy (66.08

NOTE
Class 0: referenced to dysphonia people.
Class 1:referenced to dysphonia people.

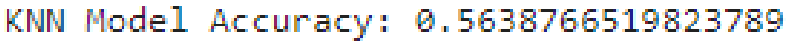
Detailed Classification Reports:

- **Class 0 Performance:** Precision: Random Forest (0.73) > KNN (0.66)
  Recall: Random Forest (0.69) > KNN (0.56)

- **Class 1 Performance:**

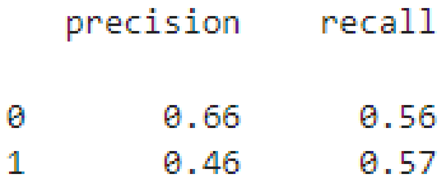
  Precision: Random Forest (0.57) > KNN (0.46)
  Recall: Random Forest (0.62) > KNN (0.57)

Conclusion
Random Forest is the best model overall for this dataset.



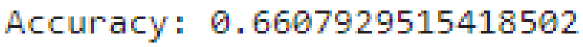
```
KNN Model Accuracy: 0.5638766519823789
```

Figure 3.18: Knn accuraccy



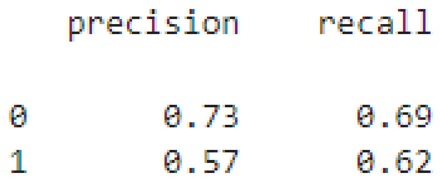
```
          precision    recall

0         0.66      0.56
1         0.46      0.57
```

Figure 3.19: Knn precision and recall report



```
Accuracy: 0.6607929515418502
```

Figure 3.20: Random forest accuraccy



```
        precision    recall

0            0.73      0.69
1            0.57      0.62
```

Figure 3.21: Random forest precision and recall report

3.4 conclusion

In conclusion, we have successfully achieved the objectives outlined in this project, and the results indicate that the random forest model is the most performant.

General conclusion

In this work, we proposed and implemented a system for pathological speech (dysphonia) detection. For training and testing data, we used recordings of the sustained vowels /a/, /i/, and /u/. In order to obtain voice quality information from these recordings, we implemented methods for extracting speech features. In order to design the most optimal classification model, we worked with three types of classifiers based on the following methods: random forests classifier (RFC), and K-nearest neighbors (KNN). The highest accuracy was achieved by the random forest classifier.

Despite encountering difficulties and challenges throughout the project, we successfully completed all priority tasks.
These obstacles, including technical issues and time constraints, taught us valuable lessons in time management, working methodologies, and adhering to specifications.
The experience helped us enhance our programming skills.

in conclusion, we have successfully achieved the objectives outlined in this project, and the results indicate that the random forest model is the most performant.

Looking ahead, future improvements can include incorporating new scenarios to enhance user assistance and implementing additional features and using other ML or DL algorithem.

Bibliography

[1] Barry, b. saarbruecken voice database. institute of phonetics; saarland university. available online. https://stimmdb.coli.uni-saarland.de/index.php4target.

[2] salaha zaiez institute. https://www.institutsalahazaiez.com/.

[3] Z. K. Abdul and A. K. Al-Talabani. Mel frequency cepstral coefficient and its applications: A review. *IEEE Access*, 10:122136–122158, 2022.

[4] O. Ceachir, R. Hainarosie, and V. Zainea. Total laryngectomy–past, present, future. *Maedica*, 9(2):210, 2014.

[5] L. Chahda, B. A. Mathisen, and L. B. Carey. The role of speech-language pathologists in adult palliative care. *International Journal of Speech-Language Pathology*, 19(1):58–68, 2017.

[6] S. M. Cohen, J. Kim, N. Roy, C. Asche, and M. Courey. Prevalence and causes of dysphonia in a large treatment-seeking population. *The Laryngoscope*, 122(2):343–348, 2012.

[7] P. N. Le, E. Ambikairajah, J. Epps, V. Sethu, and E. H. Choi. Investigation of spectral centroid features for cognitive load classification. *Speech Communication*, 53(4):540–551, 2011.

[8] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[9] S. Reghunathan and P. C. Bryson. Components of voice evaluation. *Otolaryngol Clin North Am*, 52(4):589–595, 2019.

[10] C. T. Sasaki and E. M. Weaver. Physiology of the larynx. *The American journal of medicine*, 103(5):9S–18S, 1997.

[11] R. Wirth and J. Hipp. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, pages 29–39. Manchester, 2000.