People’s Democratic Republic of Algeria

Ministry of High Education and Scientific Research

AMO Bouira University

Science and Applied Sciences Faculty

Electrical engineering department

Master’s Degree Thesis

Department: Electrical Engineering

Field: Science and Technology

Major: Electronics

Specialization: Embedded Systems Electronics

Theme

Applying Artificial Intelligence Methods to a Task Involving Recognition

Supervised by:

- REZKI Mohamed

Co-Supervised by:

- MEDJEDOUB Smail

Realized by :

- **AZERARAK Lydia**
- **CHIHATI Selma**

2023 / 2024

Acknowledgments

This work was carried out within the Department of Science and Applied Sciences at the University of Bouira.

First and foremost, we thank ALLAH Almighty for all the will, courage, and patience He has given us to complete this work.

We would like to thank, first and foremost, our supervisor, Mr. Rezki Mohamed, for his high-quality scientific guidance, his valuable advice, and his encouragement, which have contributed effectively to the advancement of this work.

We also thank the two juries who have honored us by carefully studying our work.

Finally, thanks to all those who have contributed, directly or indirectly, to the accomplishment of this modest work.

Thank you all!

2023 / 2024

Dedication 1

I dedicate this modest work to my beloved parents,

whose unwavering support and encouragement have been my greatest source of strength.

To my dear brother and sisters, your boundless love and constant support

have been invaluable throughout this journey.

CHIHATI Selma

2023 / 2024

Dedication 2

This work is dedicated to my parents,
whose enduring support and love have been a constant source of strength;
to my husband, whose patience, encouragement, and belief in me have been invaluable;
to my siblings, who have always stood by my side; and to all my friends, whose
encouragement and support have motivated me throughout this journey.
Their collective faith in me has made this accomplishment possible, and I am deeply
grateful for their presence in my life.

AZERARAK LYDIA

Abstract

Automatic identification of human speaker is one of the most challenging aspects of speech processing. To address this complex task, identifying the speaker's gender is essential, followed by age and other attributes. Gender recognition broadly relies on extracting meaningful features from speech signals and classifying them as either male or female.

This thesis explores gender recognition from speech using machine learning techniques, employing a large dataset (8422 voice samples) from the VoxForge database. The project is conducted in three phases. In the first phase, preprocessing, the focus was on extracting and preprocessing a set of relevant features - pitch and MFCCs - that effectively differentiate gender. In the second phase, modeling, we evaluate the performance of six different machine learning algorithms, including SVM with four kernels, k-NN, Random Forest, XGBoost, K-Means, and GMM. Among these, XGBoost demonstrated the highest accuracy, followed by Random Forest and GMM. The final phase involves the practical application of the findings through the development of a real-time gender classification system, featuring a graphical user interface (GUI) for user interaction.

Key words: Artificial Intelligence, Machine Learning, Speech processing, Gender recognition, MFCCs, Pitch, K-Nearest Neighbors, SVM, Random Forest, XGBoost, K-Means, GMM.

Résumé

L'identification automatique des locuteurs humains est l'un des aspects les plus complexes dans le traitement de la parole. Pour aborder cette tâche, l'identification du genre du locuteur est essentielle, suivie de l'âge et d'autres attributs. La reconnaissance du genre repose principalement sur l'extraction de caractéristiques discriminantes des signaux vocaux et leur classification en voix masculine ou féminine.

Cette thèse explore la reconnaissance du genre à partir de la parole en utilisant des techniques d'apprentissage automatique et en s'appuyant sur un large ensemble de données comprenant 8422 échantillons vocaux issus de la base de données VoxForge. Le projet s’est déroulé en trois phases. Dans la première phase, nous nous focalisons sur l'extraction et le prétraitement de caractéristiques pertinentes - le pitch et les MFCCs - qui permettent de différencier efficacement le genre. Dans la deuxième phase, la modélisation, nous évaluons et comparons la performance de six algorithmes d'apprentissage automatique en termes d’accuracy. Ces algorithmes sont : le SVM avec ses quatre noyaux, le k-NN, la Forêt aléatoire, XGBoost, les K-Moyennes et GMM. Les resultats de l’évaluation indiquent que XGBoost est le plus performant, suivi par la Forêt Aléatoire et GMM. La phase finale consiste à mettre en pratique les résultats à travers le développement d'un système de classification du genre en temps réel, avec une interface graphique utilisateur (GUI).

Mots clés : Intelligence Artificielle, Apprentissage Automatique, Traitement de la Parole, Reconnaissance du Genre, MFCCs, Pitch, k plus proches voisins, SVM, Forêt aléatoire, XGBoost, K-Moyennes, GMM.

ملخص:

تهدف هذه الأطروحة إلى التعرف على الجنس من خلال الصوت باستخدام تقنيات التعلم الآلي، بالاعتماد على مجموعة بيانات كبيرة تضم 8422 عينة صوتية مستمدة من قاعدة بيانات VoxForge. تم تنفيذ المشروع على ثلاث مراحل. في المرحلة الأولى، ركزنا على استخراج ومعالجة الخصائص المميزة، مثل طبقة الصوت (Pitch) ومعاملات الترددات الميلافية (MFCCs)، التي تتيح التفريق الفعال بين الجنسين. في المرحلة الثانية، قمنا بتقييم ومقارنة أداء ستة خوارزميات للتعلم الآلي من حيث الدقة. هذه الخوارزميات هي: آلة المتجهات الداعمة (SVM) بأربعة أنوية، k-NN، الغابة العشوائية ((Random Forest، XGBoost، التجميع بالوسائل (K-Means) ونماذج المخاليط الغاوسية (GMM). أشارت نتائج التقييم إلى أن XGBoost هو الأفضل أداءً، يليه الغابة العشوائية و GMM. تتضمن المرحلة النهائية تطبيق النتائج من خلال تطوير نظام تصنيف الجنس في الزمن الحقيقي، مزود بواجهة مستخدم رسومية (GUI).

الكلمات المفتاحية

، طبقة الصوت(MFCCs) الذكاء الاصطناعي، التعلم الآلي، معالجة الصوت، التعرف على الجنس، معاملات الترددات الميلافية ، التجميع XGBoost، (Random Forest) ، الغابة العشوائية(SVM) ، آلة المتجهات الداعمة(k-NN) ، الجيران الأقرب(pitch) (GMM) ، نماذج المخاليط الغاوسية(K-Means) بالوسيط .

List of contents :

List of figures:

List of tables

List of Symbols and Acronyms

AI	Artificial Intelligence
ANN	Artificial Neural Networks
ASR	Automatic Speech Recognition
CSV	Comma-Separated Values
DT	Decision Tree
FFT	Fast Fourier Transform
FN	False Negative
FP	False Positive
GMM	Gaussian Mixture Model
GUI	Graphical User Interface
KNN	K-Nearest Neighbors
MFCC	Mel Frequency Cepstral Coefficient
ML	Machine Learning
NLP	Natural Language Processing
PCA	Principal Component Analysis
RBF	Radial Basis Function
RF	Random Forest
RNN	Recurrent Neural Network
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
XGBoost	Extreme Gradient Boosting

General introduction

The ability to recognize and categorize human speech is a fundamental aspect of human communication, enabling us to convey emotions, thoughts, and intentions. In the realm of modern technology, replicating this natural human skill through artificial intelligence has opened new avenues in various fields, from personal assistants like Siri and Alexa to sophisticated security systems and medical diagnostics [1]. This thesis delves into one such application of speech recognition: gender recognition from voice.

Human speech is a rich source of information, characterized by various acoustic features such as pitch, frequency and intensity. Gender recognition, a crucial aspect of speech recognition, leverages these acoustic attributes to distinguish between male and female voices [2]. This process is like the natural ability of the human ear to identify gender differences based on voice [3]. With the advent of machine learning, automated systems can now emulate this capability with high accuracy, providing significant advantages in numerous practical applications.

One of the key challenges in developing an effective automatic gender recognition system is selecting the appropriate features and algorithms that can effectively capture and differentiate the subtle nuances in male and female voices. Machine learning, especially supervised learning algorithms, has significantly advanced the field of gender recognition by enabling the analysis of large datasets to train models that accurately classify gender.

The main objective of this research is to develop a robust machine learning classification model for gender identification using recorded voice samples. This involves experimenting with various machine learning algorithms, both supervised and unsupervised, conducting rigorous evaluations and comparisons to identify the best-fitting model. The chosen model will then be integrated into a practical application to effectively demonstrate the findings and real-world applicability of the research outcomes.

Our thesis is organized into three chapters as follows:

- The first chapter explains how speech is produced and perceived, and the key acoustic features that differentiate voices, focusing on Mel-frequency Cepstral Coefficients

(MFCC) and pitch. Additionally, it introduces the fundamental concepts and applications of speaker recognition systems, setting the stage for exploring gender recognition frameworks critical in speech processing.

- The second chapter presents the basics of artificial intelligence (AI) and machine learning (ML). It covers key algorithms, their applications, and objectives.
- The third chapter details the methodology for developing a gender recognition system using voice features, including data preprocessing, feature extraction, and performance evaluation of various machine learning algorithms. Additionally, it introduces a real-time gender classification application with a graphical user interface.

Chapter 1 :

Fundamentals of Voice and its Characteristics

Chapter 1:

Fundamentals of Voice and its Characteristics.

1. Introduction:

Speech is one of the primary means of communication between human beings, and its simplicity makes it the most popular communication method in human society (it is easier to speak to someone than to write or draw for them) [3], this communication is facilitated by a natural biological process wherein air expelled from the lungs is converted into speech as it passes through various organs including the vocal cords, tongue, teeth, and lips .However, this simplicity, (for humans), involvesa highly complex process carried out by our brain, from the production of speech to its perception and understanding. This complexity makes speech difficult to automate for a machine [4] [5].

Human speech contains a wealth of information marked by diverse acoustic features such as loudness, frequency, intensity, and pitch. Gender recognition, a subset of speech recognition systems [6], depends on analyzing these acoustic features, which are crucial for distinguishing between male and female voices. By examining these characteristics, gender recognition systems can accurately identify and differentiate gender, thereby improving their overall functionality and accuracy. This capability enhances the efficiency of gender recognition in speech processing technologies [7].

This chapter delves into the fundamentals of vocal signals, exploring their characteristics and properties, including human speech production and perception. It provides an overview of various speech features, emphasizing the importance of MFCC and Pitch for applications such as speech recognition and gender speaker recognition.

2. The human speech:

In this section, we present briefly human speech production and perception.

2.1.Definition of speech :

Literally, "voice" refers to the sounds produced when air passes through the mouth. Metaphorically, "voice" can also represent the unique way individuals express themselves. More technically, the voice is the sound created by the vibration of vocal cords, which is then shaped by the resonance of the vocal tract [8].

2.2.Speech production:

The production of speech begins in the brain, where the message and its lexical and grammatical structure are formulated. Following this mental preparation, a representation of the sound sequence is created along with a series of commands. These commands are executed by the speech organs to produce the spoken sound [9].

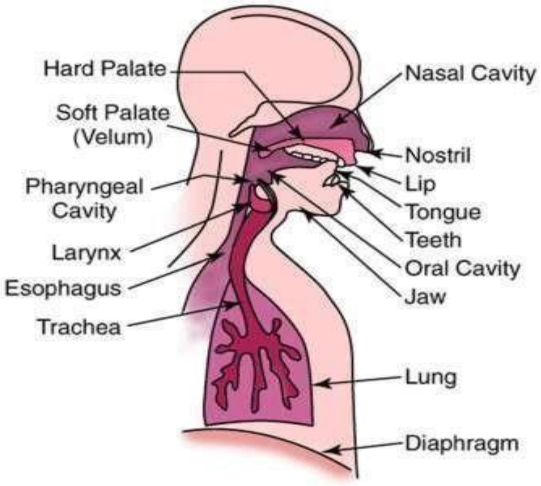


Figure 1.1: Production of speech sound [10].

To understand this phenomenon, we will explore the articulatory process of producing speech sounds to lay a foundation for understanding vowels and consonants in subsequent sections [11]. The journey begins in the mind, where a message is crafted along with its lexical and grammatical structure. This is followed by the creation of a sound sequence representation and the generation of commands that guide the speech organs to deliver the utterance. This requires a phonetic plan and a motor plan [12]. Following these mental operations, the physical production of speech sounds occurs. Speech is generated by an airstream that originates in the lungs, travels through the trachea, and passes through the oral and nasal cavities. This involves four primary processes: initiation, phonation, oro-nasal processing, and articulation. The initiation process starts with the expulsion of air from the lungs. Otherwise, speech sounds primarily involve a "pulmonic regressive air stream,"[13] although not all languages use this method, as some incorporate ingressive sounds. Phonation occurs at the larynx, where air passes through two horizontal tissue folds known as the vocal folds. The space between these folds is referred to as the glottis.

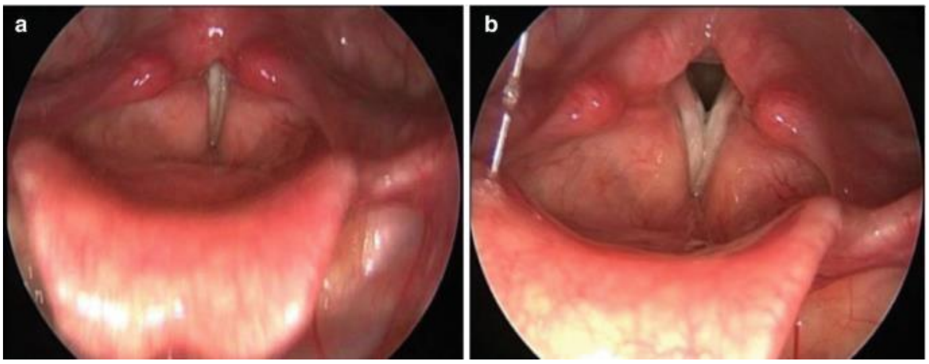


Figure 1.2: Vocal folds, (a) Closed glottis, (b) Open glottis [14].

The glottis may be in three states: fully closed, as shown in Figure 1.2 (a), preventing any air passage; slightly open, allowing the vocal folds to vibrate and produce "voiced sounds"; or fully open as shown in Figure 1.2 (b), as during regular breathing, reducing vocal fold vibrations and creating "voiceless sounds."

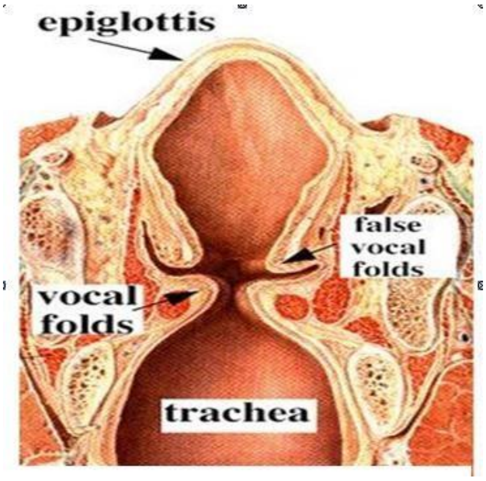


Figure 1.3: The Larynx [15]

After passing through the larynx and pharynx, air enters either the nasal or oral cavity, directed by the velum as depicted in Figure 1.4. This oro-nasal process enables the distinction between nasal consonants (/m/, /n/, /ŋ/) and other sounds, as referenced in [13].

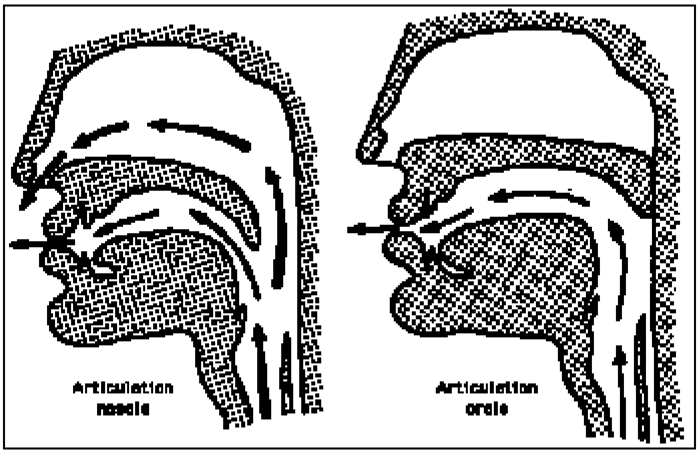


Figure 1.4 : The oro-nasal process [15]

The articulation process, most evident in the mouth, is crucial for differentiating most speech sounds. Within the oral cavity, which serves as a resonator, various articulators which can be active or passive including the upper and lower lips, teeth, and the tongue (tip, blade, front, and back), as well as the roof of the mouth (alveolar ridge, palate, and velum).

Speech sounds are thus characterized by their articulation points and the manner of their production [13].

Speech can be categorized based on this articulatory process:

- **Voiced Speech:** This occurs when air, expelled from the lungs, travels through the windpipe to the vocal cords. If the vocal cords vibrate, segmenting the airflow into periodic signals, the resulting speech is termed voiced.
- **Unvoiced Speech:** Here, despite airflow from the lungs, the vocal cords do not vibrate, leading to aperiodic signals and resulting in unvoiced speech.
- **Silenced:** This category applies when the vocal cords remain still with no vibration after air is expelled from the lungs [5].

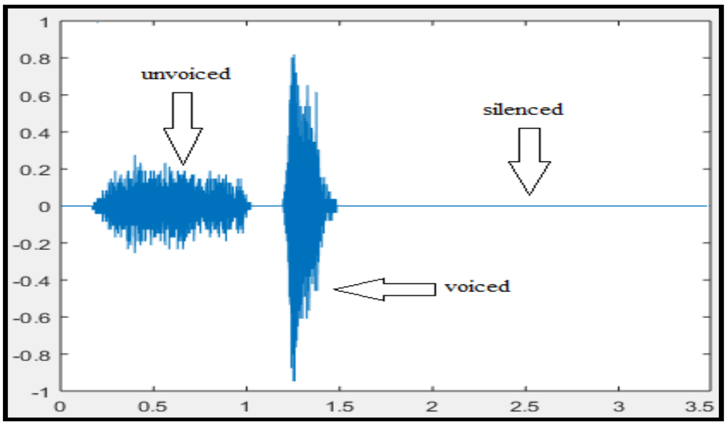


Figure 1.5: Types of Speech Produced [5].

The figure 1.5 represents the three types of speech produced.

2.3.Speech perception:

This section introduces speech perception, focusing on its physical aspects essential for speech recognition.

2.3.1. The auditory system:

The integration of knowledge about the human auditory system could enhance the effectiveness of automatic speech recognition systems. This section briefly outlines the human auditory system as shown in Figure 1.6, which details the structure of the human ear.

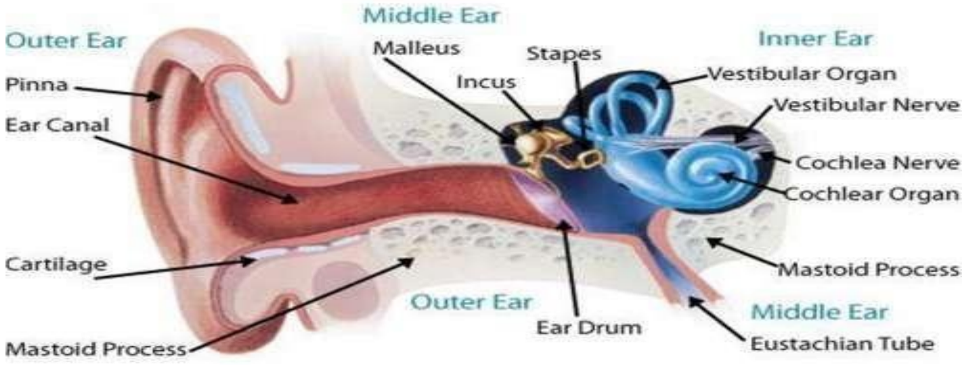


Figure 1.6: Structure of the human ear [16].

The auditory system, also known as the hearing system, comprises various components essential for effective auditory perception. These include the outer ear, middle ear, inner ear, and the auditory nervous system. The outer ear comprises the pinna and the ear canal, with the pinna being the external part that channels sounds into the ear canal. Within the middle ear, sound vibrations are transmitted from the tympanic membrane, commonly known as the eardrum, to the inner ear through a set of small bones named the malleus, incus, and stapes, collectively referred to as the ossicles. The inner ear contains the cochlea, a spiral-shaped structure lined with tiny hair cells that convert sound vibrations into neural signals. Finally, the auditory nerve transports these signals from the cochlea to the brain stem's nucleus and then to the temporal lobe, where sounds are processed and interpreted [17].

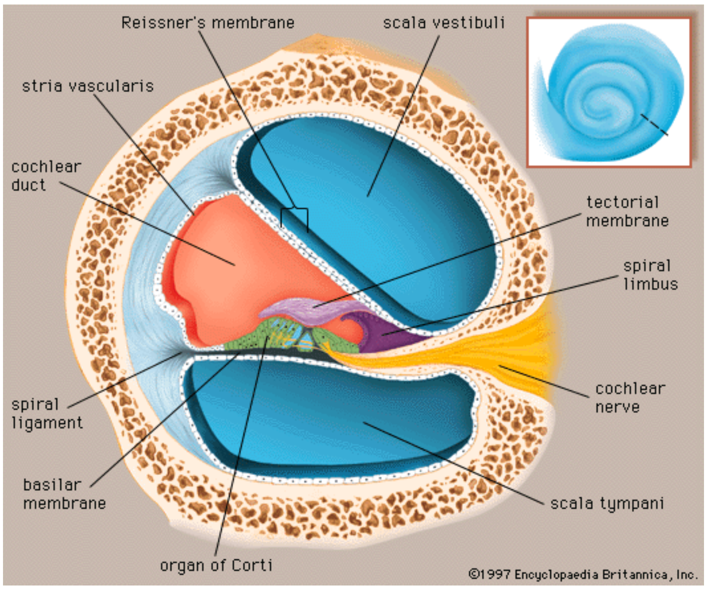


Figure 1.7: structures of the cochlea; human ear [18].

Figure 1.7 represents the structure of the cochlea.

2.4.Characteristics of the Speech Signal:

To delve back into the fundamental of the vocal signal, it is essential to understand speech as a vibratory output produced by the human vocal tract's articulatory mechanisms. The principal attributes of the vocal signal are outlined as follows [19][20]:

- **Redundancy:** The speech signal is characterized redundancy, incorporating multiple layers of information such as phonetic sounds and grammatical structures. This redundancy contributes to its resilience against interference, but it also requires the selective extraction of relevant information from the signal to avoid quality degradation.
- **Continuity:** Speech is a continuous flow of sound over time, which requires prior discretization of the signal for digital processing and analysis.
- **Variability:** There is significant variability in the speech signal, arising from the multitude of factors that influence spoken language, including dialectal differences, individual speaker characteristics, and contextual variations.
- **Non-stationarity:** The notion of a stationary process serves as an oversimplified model, as real-world phenomena are rarely strictly stationary due to continual evolution within their associated physical systems. Although such models are practical and useful for analyzing brief, consistent segments of speech, they fall short in capturing the full dynamics of speech signals. The non-stationary nature of the speech signal contributes to its complexity; therefore, its analysis often necessitates consideration under the premise of local stationarity (quasi-stationarity).

2.5.Acoustic properties of speech signal:

2.5.1. Frequency:

Frequency refers to the number of pressure oscillations per second, expressed in Hertz (Hz). Higher frequencies produce sounds with higher pitches, whereas lower frequencies generate lower-pitched sounds. The range of frequencies that can be heard by humans extends from 20 to 20,000 Hz [28].

2.5.2. Duration:

Corresponds to the time a sound lasts. It is generally measured in seconds [28].

2.5.3. Amplitude:

It is the variation of maximum pressure reached compared to the reference pressure. It is calculated by measuring the acoustic pressure level (P) and the reference acoustic pressure level (Po), where the reference pressure level approximately corresponds to the threshold of human ear perception [29].

2.5.4. Spectrum:

The frequency representation that defines the intensity of speech, typically obtained through Fourier analysis. The quasi-stationarity of the speech signal requires the implementation of effective analysis and modeling methods for short-term processing of vocal signals over a window duration commonly ranging between 20 milliseconds and 30 milliseconds, referred to as frames, with overlap between these windows to ensure the analysis of temporal continuity of features [30].

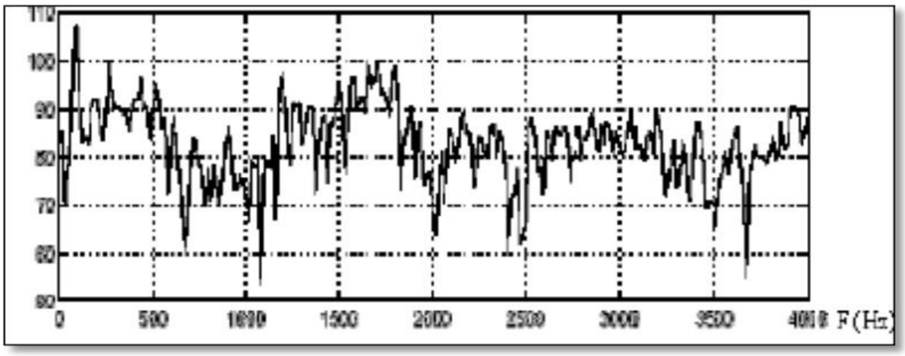


Figure 1.9: Example of the spectrum of an unvoiced signal [30].

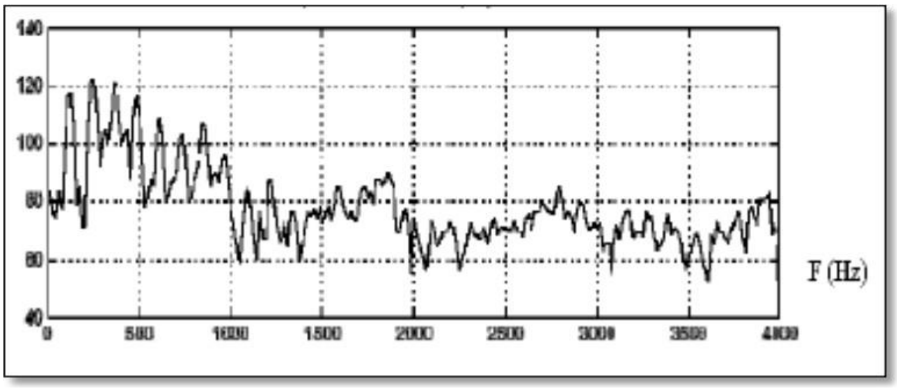


Figure 1.10: Example of the spectrum of a voiced signal [31].

2.6.Speech Features:

Due to the significant variability in speech signals, it is beneficial to conduct feature extraction to minimize this variability. The objective is to isolate pertinent information from the speech signal. This process basically transforms the speech signal in a numerical representation [5]. In this section we briefly present different speech features.

2.6.1. Spectral Features:

2.6.1.1.Mel Frequency Cepstral Coefficient Measurement:

Mel Frequency Cepstral Coefficients (MFCC) is among the most prevalent speech features, it closely mirrors the auditory capabilities of the human ear by capturing similar parameters that the human ear extracts. Speech production begins when air expelled from the lungs travels through the esophagus, causing the vocal cords to vibrate and convert the air into quasi-periodic signals. Subsequently, the signal undergoes modulations in frequency that is largely influenced by the shape of the vocal tract. The vocal tract comprises the oral cavity (including the mouth, tongue, and lips), the pharyngeal cavity (encompassing the throat), and the nasal cavity [21]. This shape forms an envelope that, if discerned accurately by the human ear, allows for the precise identification of the phoneme being produced. MFCC effectively determines this shape and represents its envelope, enabling accurate phoneme recognition [22].

This section describes how each step is carried out during the MFCC feature extraction. The following blockdiagram in Figure 1.8 represents the MFCC feature extraction process.

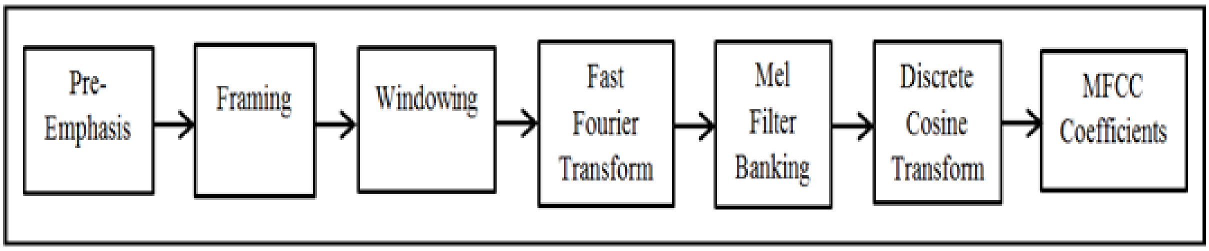


Figure 1.8 : MFCC feature extraction process [5].

The implementation steps involved in the MFCC feature extraction process and their significance are outlined as follows:

- **Pre-emphasis** :

At this step the signal passes through a filter, which emphasizes higher frequencies in the signal. This means that it increases the energy of the signal where it's low while also adjusting for the high frequency components of the speech signal [5].

In this initial phase, the signal undergoes filtering to accentuate its higher frequency components. This enhancement increases the energy levels in portions of the signal that exhibit lower amplitude, concurrently adjusting for the dominant high-frequency elements of the speech signal [5].

$$Z'(n) = Z(n) - \alpha * Z(n-1) \qquad (1)$$

- Z'(n) is the pre-emphasized signal at time n.
- Z (n) is the original signal at time n.
- Z (n-1) is the original signal at the previous time step (n-1).
- The value α is known as the pre-emphasis coefficient and typically varies from 0.9 to 1.

- **Framing** :

In this phase, the speech signal is segmented into units called frames [5]. The process involves segmenting the continuous flow of speech into fixed-length samples. This segmentation is essential for the block-wise analysis of the signal [23].

- **Windowing :**

After segmenting the speech signal into frames, the next step entails applying a windowing technique to each frame. This technique is intended to minimize signal discontinuities at thebeginning and end of each frame. Specifically, each frame resulting from the segmentation process is multiplied by a Hamming window, represented as S (n) * h (n), where w (n) is the hamming windowfunction defined by [5]:

$$h(n) = 0.54 - 0.46 * \cos(2\pi n / N-1) \qquad (2)$$

Where:

- S(n) is the speech signal frame.
- h(n) is the window function applied to each sample in the frame.
- 0.54 and 0.46 are constants that define the shape of the Hamming window.
- $\cos(2\pi nN-1)$is the cosine function scaled to the length of the frame.
- n varies from 0 to −1, where N is the number of samples in the frame.

➢ Fast Fourier Transform:

In this phase, each frame, which contains N samples in the time domain, undergoes transformation into the frequency domain using the following equation [5]:

$$Z_i\,(k) = \sum^{N} Z_i\,(n)\;h\,(n)\;e^{-2\pi\,ke-2\,\pi\,Ke/N} \qquad (3)$$

Where:

- Zi (n) represents the signal in the time domain.
- Z i(k) represents the signal in the frequency domain, ranging from 1 to K.
- H (n) is a window function of N samples, and K is the length of the FFT.

After the signal is converted from the time domain to the frequency domain, the next step involves the estimation of different frequencies present in the signal. This is achieved by computing the power spectrum for each frame through a periodogram estimate.

This step is analogous to the function of the cochlea in the human ear, which is located in the middle ear and filled with a viscous liquid that is sensitive to vibrations. As different frequencies are present in the speech signal, the liquid in the cochlea vibrates at different locations, depending on which different auditory nerves to carry the signal to the brain, indicating the presence of specificfrequencies [22].

To represent the periodogram estimate, the absolute values of the FFT for each frame are takenand squared. The periodogram estimate for a speech frame is represented as:

$$P_i = \frac{1\,|\,Z_i\,(K)\,|}{N} \qquad (4)$$

➢ Mel Filter Bank

As previously stated, the Mel Frequency Cepstral Coefficients (MFCC) take into account the sensitivity of human auditory perception to different frequencies by converting the frequency domaindata into Mel scale representations [5].

$$m = 2595\log_{10}\left(1+\frac{f}{700}\right) \qquad (5)$$

Where :

- Frequency (f): The input frequency in Hertz (Hz) to be converted.
- Normalization: The frequency is divided by 700 to normalize it.
- Addition of 1: Ensures the argument of the logarithm is positive.
- Scaling by 2595: Adjusts the logarithmic value to match the Mel scale

$$f = 700(10^{m/2595} - 1) \qquad (6)$$

Where:

- Mel Scale Value (m): The input value on the Mel scale.
- Division by 2595: Reverses the scaling applied in the first equation.
- Exponentiation: Converts the logarithmic value back to a linear scale.
- Subtraction of 1: Reverses the earlier addition of 1.
- Multiplication by 700: Scales the result back to the original frequency range in Hertz.

➢ Discrete Cosine Transform

This step involves transforming the signal back into the time domain [5]. The formula used for this transformation is detailed below:

$$C_m = \overset{k=1}{\sum^{N}} \cos\left[m * (k - 0.5) * \frac{\pi}{N}\right] * E_k \qquad (7)$$

Where:

- C_m is the m-th Mel-Frequency Cepstral Coefficient (MFCC).
- E_k is the log-transformed energy output obtained from previous step (Mel filter bank).
- N is the total number of Mel filters.
- m is the index of the MFCC (The value of **m** is ranging from 1, 2, 3…….L (**L** is the number of cepstral coefficients).

2.6.1.2. **Formants:**

Formants are resonant frequency bands in the speech signal that correspond to the vocal tract’s natural resonances. They play a pivotal role in speech intelligibility and phonetic analysis

2.6.1.3.Spectral entropy:

Spectral entropy measures the distribution of power within a signal's spectrum and is an important feature in speech recognition. The normalized power distribution of a signal is treated as a probability distribution in the frequency domain, utilizing spectral entropy [31]. The probability distribution is calculated using the following equation:

$$P(k) = \sum_{K=1}^{N} \frac{S(K)}{S(K)'} \qquad \text{For } k = 1 \text{ to } N \qquad (8)$$

Where:

- p(k) is the probability distribution.
- S (K) is the k^{th} power spectrum where S (k) =| X (k)|2. X (k) is the discrete Fourier transform of the signal.
- S′ (K) is the sum of the power spectral densities.

The power spectrum and probability distribution are needed to compute the spectral entropy with the following equation:

$$SE = \sum_{K=1}^{N} P(k) \log_2 P(k), \qquad \text{For } k = 1 \text{ to } N \qquad (9)$$

Where:

- SE: is the spectral entropy.

2.6.1.4.Spectral flatness:

It is a feature of acoustic signals which is useful in digital signal processing [32]. Spectral flatness is the term used to describe the ratio of the geometric mean to the arithmetic mean of a powerspectrum. It is mainly used to quantify the noise- like or tone-like nature of a sound signal. By evaluating spectral flatness, we can identify whether a signal has a flat or non-flat spectrum. The power spectrum being perfectly flat, and the ratio of the arithmetic mean and geometric mean being 1 indicates that the arithmetic mean and geometric mean are equal. The geometric mean is always greater or equal to the arithmetic mean, so the ratio can’t be more than 1[33].

2.6.2. Prosodic features

2.6.2.1.The fundamental frequency (Pitch):

A crucial element of speech recognition is the fundamental frequency, which characterizes the essential physical property of the signal. This is often referred to as the approximate frequency of thequasi-periodic voiced speech signal [24].

The pitch of the voice corresponds to both low and high sounds, which depend on the number of air vibrations. These vibrations are produced by the vocal cords, which can vibrate between 16 and 20,000 times per second. The frequency of these vibrations per second, measured in Hertz, determinesthe pitch of the sounds and is referred to as the fundamental frequency.

The vocal cords vibrate at a certain pitch, also known as tone. This pitch, whether low or high,varies depending on the tension exerted on the vocal cords, which in turn depends on the physical aspects of the vocal cords and thus differs between men and women.

Control over pitch is possible through the tension and stretching of the vocal cords in addition to the modulation of air pressure. This air pressure is controlled by the respiratory muscles, including the diaphragm, abdominal muscles, and neck muscles. It primarily involves muscular control, thus mastering pitch is a matter of training.

The fundamental frequency typically ranges:

- Around 100 Hz in men.
- Around 200 Hz in women.
- Between 200 and 300 Hz in children. Children have much higher voices due to theimmaturity of their vocal cords [25].

The estimation of speech signal pitch has been widely studied, and the method can be divided into three types of methods: time, frequency, and spectro-temporal.

- Temporal approaches, based on the principle of autocorrelation, involve comparing the waveform of a frame with a shifted version of itself [26].
- Frequency-based approaches rely on detecting the harmonic structure of a signal containingan F0 in the frequency representation. They may consider spectral module, cepstral analysis,and implement a preliminary step of partial detection in harmonic relation [26].
- The spectro-temporal approach uses a filter bank to separate the signal and process all output signals. This separation using filter banks is intended to simulate the decomposition of the signal into time and frequency that occurs in the inner ear. Modeling can be more or less close to biology. These methods use temporal methods on each channel and utilize the separation of information into frequency bands to improve voiced/unvoiced decision and F0 estimation. The three sets of methods mentioned above are deterministic [27].

2.6.2.2.Intensity:

Intensity distinguishes between strong and weak tones; it corresponds to the amplitude of the wave. Amplitude is given by the maximum deviation of the amplitude, which characterizes the wave. For sound, this quantity is pressure, so the amplitude is given by the difference between the strongest and weakest pressure applied by sound waves. When the wave's amplitude is large, the intensity is high, and thus the sound is louder. Sound intensity is measured in decibels (dB), thus corresponding to the amplitude or the peak value of the sinusoidal wave, i.e., the pressure difference between compression areas and rarefaction zones [28].

2.6.2.3.Energy:

It corresponds to the sound intensity, which is related to the pressure in the area above the larynx. The amplitude of the speech signal varies over time depending on the type of sound [30], and its energy is given by:

$$E=\sum_{n=0}^{N-1} S^{\wedge\,(\mathrm{n})} \qquad (10)$$

In summary, Mel-frequency cepstral coefficients (MFCC) and pitch are widely recognized as essential features in speech recognition tasks such as speaker and gender recognition. The main advantage of MFCC is the robustness towards noise and spectral estimation errors under various conditions [34]. A. Reynolds conducted a study comparing various features and discovered that MFCC outperforms the others [35]. In the other hand, pitch provides crucialfundamental frequency information, which is instrumental in distinguishing between speakers and genders based on voice characteristics, when the number of speakers is small (on the order of 10), pitch can be reliably used to discriminate gender speakers. When the number of speakers is greater, pitch has to be combined with other parameters.

Therefore, the combination of MFCC and pitch features is fundamental in speech recognition tasks, providing a robust foundation for extracting essential information from speech signals and enabling accurate classification and identification of speakers[36].This is particularly important for applications where speech is often contaminated by noise or has varying levels of quality [37][38],what drove Shao and al [39] to propose an integrated extraction of pitch and Mel-frequency cepstral coefficients (MFCC) for speech recognition and reconstruction using an auditory model [40].

3. Speech processing:

Speech recognition, speaker recognition, and gender recognition are pivotal components of modern speech processing systems. Speech recognition technology enables computers to convert spoken language into text, facilitating applications such as voice assistants, dictation systems, and automated customer service. Within this framework, speaker recognition distinguishes individual speakers based on their unique voice characteristics, while gender recognition identifies the gender of the speaker. These capabilities collectively enhance the functionality and utility of speech processing technologies across various domains [41].

Speech processing techniques depend on capturing speech signals typically through a microphone and transferring them to a computer via digitalization. These techniques are employed to extract specific information from the speaker, including:

- Speech recognition: Determining the words spoken (speech-to-text transcription).
- Speaker recognition: Identifying the speaker's identity (e.g., John, Lisa).
- Sex identification: Determining the speaker's gender (male or female).
- Language recognition: Identifying the language spoken (e.g., English, Spanish).
- Speech detection: Detecting whether someone is speaking (speech activity detection) [41].

4. Overview of speech recognition :

In this section, we will talk about speech recognition, and speaker gender recognition as a subset of speech recognition.

4.1.Definition of Speech Recognition:

Speech recognition, also known as automatic speech recognition (ASR), computer speech recognition, or speech-to-text, enables software to transform human speech into written text. Although often confused with voice recognition, speech recognition involves translating spoken language into text, whereas voice recognition specifically identifies an individual user’s voice [42].

4.2.Architecture of Speech Recognition:

The operation of an automatic speech recognition system involves extracting various speech features from the acoustic signal for each word or sub-word unit. These features capture how the word or sub-word changes over time, forming a pattern that characterizes the word or sub-word. During the training phase, all words in the vocabulary are read aloud, and their patterns are stored. When recognizing a word later, its pattern is compared to the stored patterns, and the word that gives the best match is selected, this approach is commonly known as pattern recognition [43].

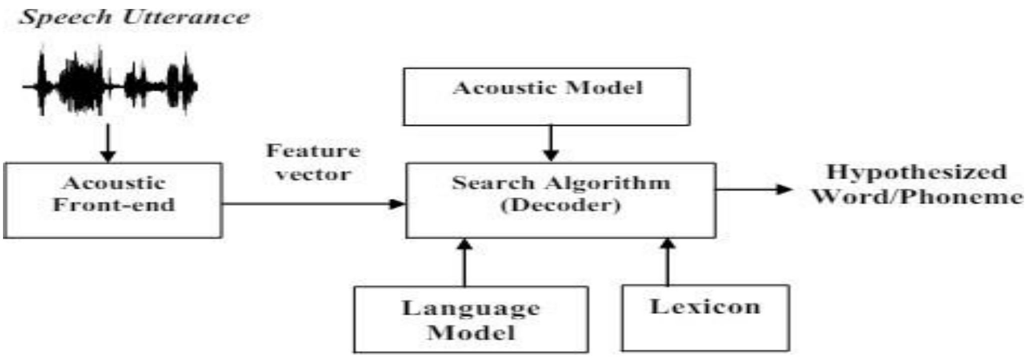


Figure 1.11: Speech recognition Architecture [43].

A typical speech recognition system is developed with major components that include acoustic front-end, acoustic model, lexicon, language model and decoder as shown in figure 1.11:

- **Acoustic Front-end**: The acoustic front-end involves processing the signal and extracting features. In speech recognition, the primary goal of feature extraction is to transform the speech signal into a sequence of compact acoustic feature vectors that contain sufficient information for recognition in subsequent stages
- **Acoustic Model**: The acoustic model is a critical component of automatic speech recognition systems, representing acoustic features for phonetic units. In building an acoustic model, choosing appropriate basic modeling units is essential in acoustic model development. Different types of sub-word units can significantly impact speech recognition performance depending on the specified target language.

- **Language Model**: The language model comprises an extensive list of words and their probabilities of occurrence within a given sequence.
- **Decoder:** This software program processes user-spoken sounds by searching the acoustic model for corresponding sounds. When a match is made, the decoder identifies the phoneme associated with the sound and continues matching phonemes until a pause in the user's speech. Then, it queries the language model for a matching sequence of phonemes. If a match is made, the decoder returns the corresponding text (word or phrase) to the calling program [43].

4.3.Speech Recognition Approach:

Basically, there exist three approaches of speech recognition [9].

Acoustic Phonetic Approach.

Pattern Recognition Approach.

Artificial Intelligence Approach.

- **The acoustic Phonetic Approach**: This approach is based on the idea that spoken language consists of a finite number of distinctive phonetic units (phonemes). These units are characterized by specific acoustic properties that are reflected in the speech signal over time.
- **The pattern-matching approach**: This method involves two main steps: pattern training and pattern comparison. It uses a well-defined mathematical framework to create consistent speech pattern representations from a set of labeled training samples. These representations are then used for reliable pattern comparison through a formal training algorithm.
- **Artificial Intelligence Approach**: This hybrid method combines elements of both the Acoustic Phonetic and Pattern Recognition approaches. It leverages knowledge from linguistic, phonetic, and spectrogram information to improve speech recognition.

4.4.Applications of Speech Recognition:

Speech recognition technology has many applications, from virtual assistants to translation and custom voice commands, details can be found on [43][44].

- **Navigation Systems:** Speech recognition software is used in navigation systems, allowing drivers to give voice commands to vehicle devices while keeping their eyes on the road and hands on the wheel.

- **Virtual Assistants**: Virtual assistants like Siri, Google Assistant, and Alexa use speech recognition to interpret user requests, answer questions, and perform tasks.
- **Healthcare:** Automatic speech recognition is used in the medical field to convert speech into text for medical reports, clinical notes, and updating electronic health records.
- **Call Centers:** Speech recognition systems are used in call centers to automate customer interactions, analyzing speech input and responding to customer requests.
- **Language Translation:** Machine translation software uses speech recognition to convert human speech from one language to another.
- **Voice Search:** Speech recognition systems are part of search engines, allowing users to surf the web using voice commands.

5. Gender Speaker Recognition:

Speaker and gender recognition form integral parts of speech recognition systems. They enable these systems not just to convert spoken words into text but also to identify the speaker (speaker recognition) and discern the speaker's gender (gender recognition). These functionalities significantly enhance the versatility and applicability of speech recognition across diverse industries and applications [6][45].

5.1.Introduction to Speaker Recognition:

Speaker recognition involves identifying individuals based on their voice. This task is typically divided into two main categories: speaker identification and speaker verification. Speaker identification determines which known voice best matches a speaker from a group, while speaker verification decides whether a speaker's claimed identity matches their provided voice sample. Speaker verification systems are less computationally demanding compared to speaker identification systems. Verification involves comparing a speaker's voice sample against one or two enrolled models, whereas identification requires matching against multiple models [46].

- **Speaker identification:** Speaker identification is the process of distinguishing a specific speaker from a group of different speakers. In this process, the system prompts the user to provide a speech utterance. The system then identifies the user by comparing the speech characteristics of the utterance with those stored in the database, and lists, which contain the most likely speakers could have given that speech utterance [47].

- **Speaker verification:** Speaker verification involves either accepting or rejecting a speaker's claim of identity. The system prompts the user, who asserts their identity as the speaker, to provide an ID. The system verifies the user by comparing the characteristics of the provided speech utterance with those provided by the user. If the comparison meets a predefined threshold, the user's identity claim is accepted; otherwise, it is rejected [47].

5.1.1. Architecture of a Speaker Recognition System

The goal of automatic speaker recognition is to extract distinctive features for speaker differentiation. This process involves three primary steps: feature extraction, modeling, and testing. The system functions in two modes: training mode, where a reference feature model is developed, and testing mode, where the input signal is compared with the stored models in its database to verify or identify the speaker [48].

The general structure of an automatic speaker recognition system is shownin Figure 1.12.(for both identification and verification speaker recognition)

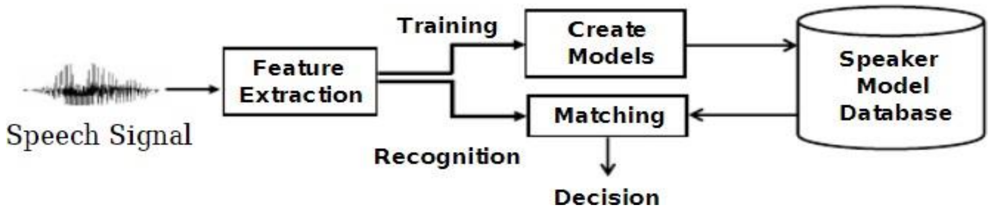


Figure 1.12: The general structure of an automatic speaker recognition system [49]

A speaker recognition system includes: Feature extraction, modeling and decision making.

- **Feature Extraction:** Feature extraction refers to the process of deriving characteristic values from a speech recording that specifically identify the speaker. This procedure involves creating a small collection of data obtained from an audio signal. Feature extraction plays a crucial role in enhancing the performance of both speech and speaker recognition systems [48].
- **Speaker modeling:** During the training phase, acoustic vectors extracted from each segment of the signal are utilized to construct a speaker model, which is subsequently stored in a database. This model serves the purpose of identifying and verifying the speaker's identity. [49].

- **Features matching:** also known as decision-making, is the final stage in speaker recognition systems where the extracted features are compared to the speaker models to determine whether the speaker is authentic or not. This stage involves several techniques to match the input features with the speaker models, including machine learning algorithms [48][49].

5.1.2. Applications of speaker recognition systems:

- **Personalization in Marketing:** Speaker Recognition enhances user experience, recommendations, and advertising by targeting ads based on individual voice features or clusters.
- **Voice Assistants:** In voice assistants, Speaker Recognition enables personalized settings like accessing favorite playlists per household member and implementing limited access and parental controls. Leading smart speakers such as Amazon Echo, Apple Homepod, and Google Assistant utilize Speaker Recognition.
- **Healthcare Applications:** In healthcare, Speaker Recognition authenticates telehealth services. For instance, Monument Health uses Voice Biometrics to verify healthcare providers accessing electronic health records (EHR) and for disease monitoring and diagnosis.
- **Speech Analytics:** Speaker Recognition in speech analytics provides insights into demographics such as gender, age, and language.
- **Law Enforcement and Forensics:** Speaker Recognition aids law enforcement by identifying criminals and facilitating efficient investigations for legal discovery applications. NSA's Voice in Real Time (Voice RT) exemplifies Speaker Identification for identifying criminals [50]

5.2.Introduction to Gender recognition:

Gender information is a distinctive and crucial property in speech. Determining this information from a speech signal is an important subject. Speaker verification systems also implicitly or explicitly use gender information. In general, identifying a speaker's gender is important for increasingly natural and personalized dialogue systems. [51]

5.2.1. Definition of Gender recognition:

Gender recognition through voice involves determining whether a speaker is male or female based on the acoustic features of their voice. This task uses a range of machine learning algorithms and techniques to achieve accurate classification [6].

Gender recognition encompasses two primary tasks: gender identification and gender verification. In gender identification, also known as 1:N matching, an unknown speaker's voice is compared against a database of N known speakers to find the best match. Gender verification, on the other hand, involves 1:1 matching to confirm whether a provided voice sample matches the claimed speaker's identity [52].

5.2.2. The architecture of a gender recognition system:

The architecture of a gender recognition system is shown in Figure 1.13. The system consists of the training and prediction phases.

- **Training phase:** During the training phase, the system receives the speech signal, undergoes pre-processing steps (such as noise removal and dimensionality reduction), and extracts acoustic features. Subsequently, a machine learning model is constructed and trained using these extracted features.
- **Recognition phase:** In the recognition phase, an unlabelled or unknown speech signal is inputted into the system. The model then predicts and outputs the gender associated with the input signal [53].

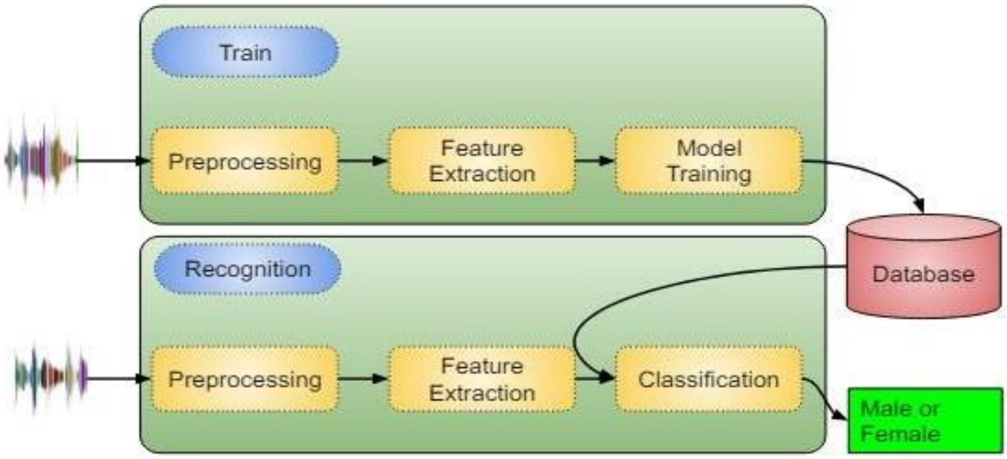


Figure 1.13: Architecture of a gender recognition system [53].

A gender recognition system comprises essential components as is mentioned on [53][48]: preprocessing, feature extraction, model training, and recognition as it is shown in Figure 1.13.

- **Preprocessing:** The voice (.wav files) must be converted to a format understandable by the system. Preprocessing is crucial to eliminate external noises.
  - **Silence Removal**: Eliminating silent or pause signals from the voice data.

  - **Pre-emphasis**: Normalizing the signal by adjusting its amplitude within each frame without affecting the duration.
  - **Framing**: Dividing the pre-emphasized signal into frames of varying lengths with overlapping percentages

- **Feature Extraction:** Once noise removal is completed, the process of feature extraction can be carried out. This involves deriving a concise dataset from the audio signal, focusing on acoustic features with high discriminative power for gender classification.
- **Model Training:** Training the machine using collected features from the dataset to enable it to classify voice genders. Various machine learning algorithms such as SVM, KNN, and DT are used for this purpose.
- **Recognition** :
  - **Classification:** The trained models are employed to categorize input voice signals into male or female based on the extracted features.

5.2.3. Applications of Gender recognition systems:

The goal of gender recognition using voice is to accurately identify the gender of a speaker based on their voice, which can have various applications [3][92] such as:

- **Improved Interaction:** Automatic gender recognition from speech signals can enhance interactive information systems by tailoring services to users' genders and boosting the efficiency of human-computer interaction through voice authentication.
- **Consumer Product Benefits:** The consumer product industry can use gender recognition to optimize advertising and marketing strategies, focusing on gender-specific consumption patterns and improving customer relationship management in telephone counseling.
- **Security and Authentication:** Gender identification can be used for targeted access control, such as restricting access to gender-specific areas on social networking platforms
- **Healthcare and Personalized Medicine**: Using gender recognition to tailor diagnoses, treatments, and care plans for certain conditions that differ between males and females.

In conclusion, this overview provides a concise look at speech recognition, speaker recognition, and gender recognition, highlighting their architectural components and diverse applications. These technologies play crucial roles in enhancing user interaction, personalization, and accessibility across various domains, driven by continuous advancements in machine learning and signal processing techniques.

6. Conclusion :

In conclusion, this chapter has provided a comprehensive overview of speech and its fundamental characteristics. We have explored the definition of speech as a complex communication process involving the production of vocal sounds and their reception and interpretation by listeners. Furthermore, our chapter has highlighted the crucial role of the auditory system in speech perception, from the reception of sound waves in the outer ear to the neural processing of auditory signals in the brain. Additionally, we have presented various speech features such as MFCC, pitch, timbre, and intensity to decode the underlying linguistic content of speech. We have also highlighted the importance of MFCC and Pitch for various applications, including speaker recognition and speech reconstruction. Overall, a deeper understanding of the fundamentals of voice lays the groundwork for the development of robust and effective speech processing and recognition systems.

In the following chapter, we will talk about the fundamentals of artificial intelligence (AI) and offer an overview of diverse machine learning algorithms including supervised, unsupervised, semi-supervised, and reinforcement learning.

Chapter 2 :

Artificial Intelligence and Fundamentals of Machine Learning

Chapter 2 :

Artificial Intelligence and Fundamentals of Machine Learning

1. Introduction

Artificial intelligence (AI), and specifically machine learning (ML), has experienced significant growth in recent times, particularly in computing, and data analysis [54].Machine learning is essential for addressing a wide range of challenges across industries such as medicine, banking, and finance. It has been widely applied in research related to gender voice recognition and classification, employing data mining techniques and advanced machine learning methods. [55]. ML enables systems to autonomously improve their performance through learning from experience, without the need for manual intervention in programming [56].

This chapter provides an overview of various machine learning algorithms, including supervised, unsupervised, semi-supervised, and reinforcement learning, along with deep learning as a subset of this field. We provide an extensive overview of these machine learning algorithms, presenting their capability to enhance the functionality and intelligence of applications. Additionally, we explore the foundational principles of different machine learning techniques and their practical applications in various domains including smart cities, cybersecurity systems, healthcare, agriculture, and e-commerce.

2. Artificial Intelligence:

In the following, we briefly define artificial intelligence and its objectives.

2.1.Definition of AI:

Artificial intelligence (AI) involves creating and developing computing systems that can perform tasks typically requiring the intelligence of humans. These tasks include voice recognition, decision-making, and identifying complex patterns. Fundamentally, AI enables machines to operate and behave like humans. AI has achieved this by developing machines and robots used across several domains, such as robotics, healthcare, marketing, business insights, and others. [57].

2.2.Objectives of AI:

Artificial intelligence can be developed by reverse engineering human abilities and traits. This approach enables computers to function autonomously and intelligently by drawing insights from human behavior. The main goal is to create intelligent machines with the following key objectives [58]:

- ➢ Enhance problem-solving skills.
- ➢ Implement methods for representing knowledge.
- ➢ Supporting strategic thinking.
- ➢ Enable ongoing education.
- ➢ Foster people skills.
- ➢ Encourage innovation.
- ➢ Attain overall cognitive capability.
- ➢ Encourage collaboration between people and artificial intelligence.

2.3.Branches of Artificial Intelligence

Artificial intelligence is a technique that penetrates nearly all domains. To more effectively address diverse societal challenges, it is segmented into several parts, each addressing a distinct problem. The primary domains of artificial intelligence include: Robotics, Expert Systems, Machine Learning, Fuzzy Logic, Neural Networks, and Natural Language Processing (NLP).

These branches enable specialized solutions and advancements across various domains [59].

Out of these branches, our focus will be on Machine Learning.

3. Machine Learning:

In the following sections, we define machine learning and present various learning techniques, highlighting their applicability in solving real-world problems.

3.1.Definition of Machine Learning:

Machine learning is a component of artificial intelligence (AI) enables machines to learn automatically from data, enhance their performance based on past experiences, and make predictions. It involves a set of algorithms that process large amounts of data. These algorithms are trained using the data and based on this training; they build models to perform specific tasks [59].

3.2.Applications of machine Learning:

Machine learning has gained popularity across diverse applications because of its capacity to learn from historical data and make informed decisions. Below, we summarize popular application areas of machine learning technology:

- **Speech Recognition:** Speech recognition involves converting spoken commands into written text, is also referred to as "Speech to Text" or "Computer Speech Recognition." Currently, numerous speech recognition applications employ machine learning algorithms. Notable examples include: Google Assistant, Siri, Alexa, and Cortana, which use this technology to process voice commands [59].
- **Image Recognition:** Image recognition stands as one of the most prevalent applications of machine learning, utilized for identifying objects, individuals, locations, and digital images. A well-known use case is Facebook's automatic friend tagging suggestion, which employs machine learning's face detection and recognition algorithms to suggest tags when photos are uploaded [59].
- **Medical Diagnosis:** In the medical field, machine learning is instrumental in diagnosing diseases. This technology has advanced rapidly, enabling the creation of 3D models that predict the exact location of brain lesions, thereby facilitating the detection of brain tumors and other brain-related conditions [59].
- **Product Recommendations and E-commerce**: Product recommendation is a widely recognized application of machine learning and a key feature of almost all e-commerce websites. Machine learning aids businesses by analyzing consumer purchasing histories and making personalized product suggestions for future purchases derived from customer preferences and behavior [60].
- **Intelligent Decision-Making and and Predictive Analytics:** Machine learning plays a critical role in making informed decisions and predictive analytics. This involves using data-driven analytics to capture and exploit the connections between explanatory and predicted variables from historical data, to forecast unknown outcomes. Examples include Identifying suspects or criminals post-crime and detecting credit card fraud in real time. [60].

3.3.Types of Machine Learning:

Machine Learning algorithms can be classified into four primary types: Supervised learning, unsupervised learning, Semi-supervised learning, and Reinforcement learning [61], as illustrated in Figure2.1.

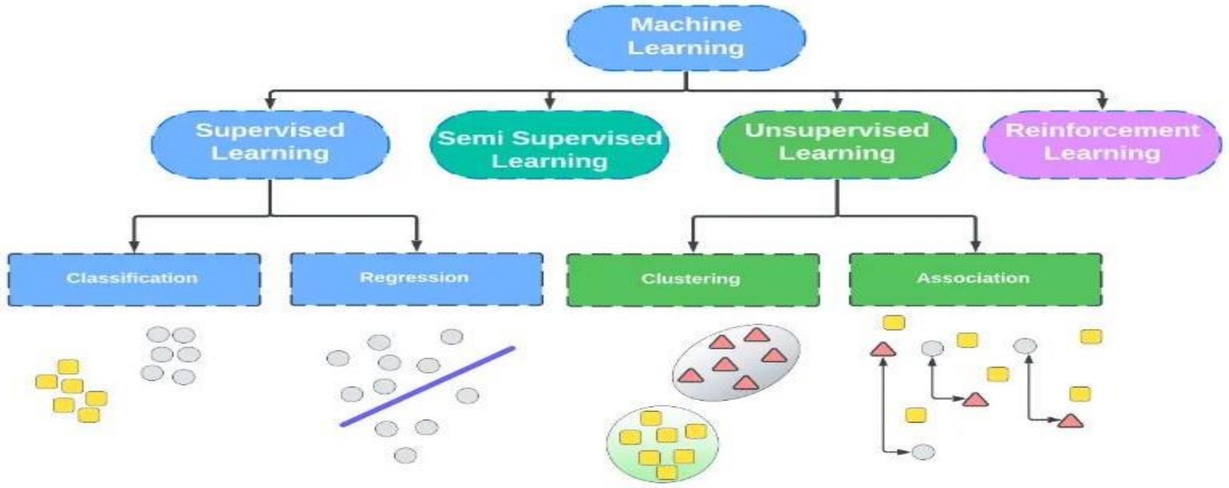


Figure 2.1: Various types of machines learning techniques [62].

In the subsequent section, we provide a brief overview of each learning technique and discuss their relevance in solving real-world problems.

3.3.1. Supervised Learning

Supervised learning involves training a model to map inputs to outputs based on example input-output pairs [63]. It employs labeled training data to infer a function, allowing the model to achieve specified goals from given inputs, making it a task-driven approach [64]. In supervisedlearning tasks include classification, which separates data into categories, and regression, which fits data to a model. An example is text classification, such as predicting the sentiment of a tweet or product review.

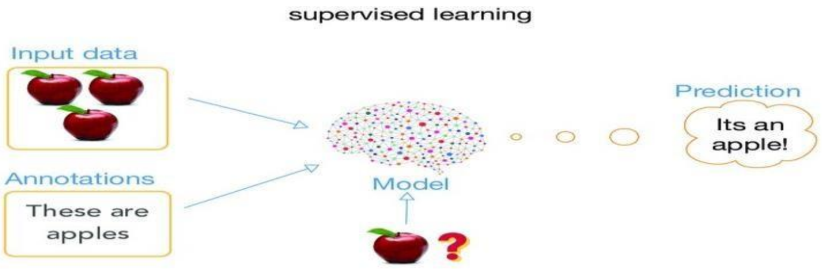


Figure 2.2: Example of supervised learning

Supervised learning addresses two main problems: Classification and regression.

In classification, the dotted line represents a linear boundary that separates the two classes, in regression; the dotted line models the linear relationship between the two variables.

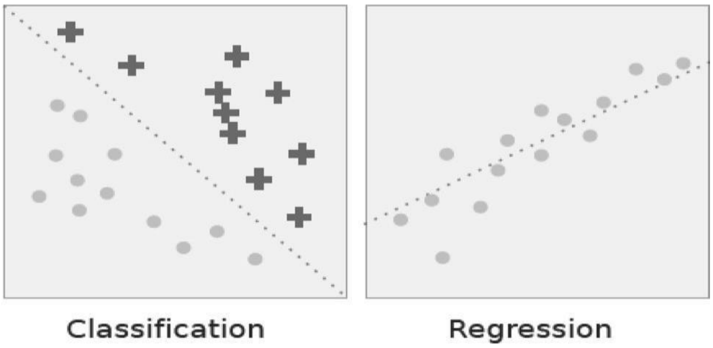


Figure 2.3 : Classification Vs Regression

3.3.1.1.Classification Analysis:

Classification is a supervised learning method in machine learning that deals with predictive modeling. It involves predicting a class label for a given example [63]. Mathematically, it maps a function (f) from input variables (X) to output variables (Y) as target labels or categories. This method is used to predict the class of data points and can be applied to structured or unstructured data. For instance, in email service providers, classification is used for tasks like spam detection ("spam" or "not spam").

3.3.1.1.1. Different types of classification:

In the following, we summarize the common classification tasks in Machine learning: binary, multi-class, and multi-label classification:

- **Binary classification:** This type of classification involves two distinct class labels, such as "true and false" or "yes and no" [63]. In binary classification tasks, one class typically represents the normal state, while the other represents an abnormal state. For example, in medical testing, "cancer not detected" might indicate the normal state, while "cancer detected" signifies the abnormal state..
- **Multiclass classification:** Traditionally, multiclass classification refers to tasks with more than two class labels [63]. Unlike binary classification, multiclass classification does not categorize examples into normal and abnormal outcomes. Instead, it assigns examples to one category from a predefined set of classes.

- **Multi-label classification:** In machine learning, multi-label classification is crucial when an example can be associated with multiple classes or labels. It extends the concept of multiclass classification by allowing examples to belong to multiple hierarchical levels of classes simultaneously. For instance, in multi-level text classification, an article on Google News might be classified under categories like "city name," "technology," or "latest news" all at once. Multi-label classification involves advanced algorithms capable of predicting multiple non-exclusive classes or labels, unlike traditional classification tasks where class labels are mutually exclusive [65].

3.3.1.1.2. Evaluation Metrics for Classification

When creating new machine learning models, comparing their performance with existing models is essential. This evaluation has two main objectives: eliminating underperforming approaches and further optimizing those that show promise. Particularly in the medical field, it is important to determine if an ML model performs better than a knowledgeable professional.

In supervised ML, the process begins by separating the data into main phases: training and testing phases. The training data is utilized to train and validate the model. Once trained, the model makes predictions on the test data, which are then compared to the true values of the test phase. This comparison enables us to determine if the predictions of a new ML model outperform those made by humans or models already exist in our test phase. [66].

Evaluation metrics for classification models offer a numerical evaluation of model performance. The selection of evaluation metric relies on the particular problem and the significance of false positives and false negatives. It is crucial to choose the right metric(s) to accurately evaluate the effectiveness of a classification model [67].

- **Definition of confusion Matrix:** The confusion matrix, also referred to as the error matrix, plays a crucial role for evaluating the performance of classification algorithms. It presents the classification outcomes of a binary classifier in a table. The confusion matrix provides a detailed analysis by displaying the quantity of true positives(TP), true negatives (TN), false positives (FP), and false negatives (FN) [67].

In the following we briefly present the confusion matrix, its components, and its applications evaluating the effectiveness of a binary classification model.

- **Components of a Confusion Matrix:** A confusion matrix is a 2x2 table utilized to display the outcomes of a binary classification model, comprising four essential components [67]:
  - True Positive (TP): The quantity of positive cases correctly classified as positive by the model.
  - False Positive (FP): The quantity of negative cases incorrectly classifiedas positive by the model.
  - True Negative (TN): The quantity of negative cases correctly classifiedas negative by the model.
  - False Negative (FN): The quantity of positive cases incorrectly classifiedas negative by the model.

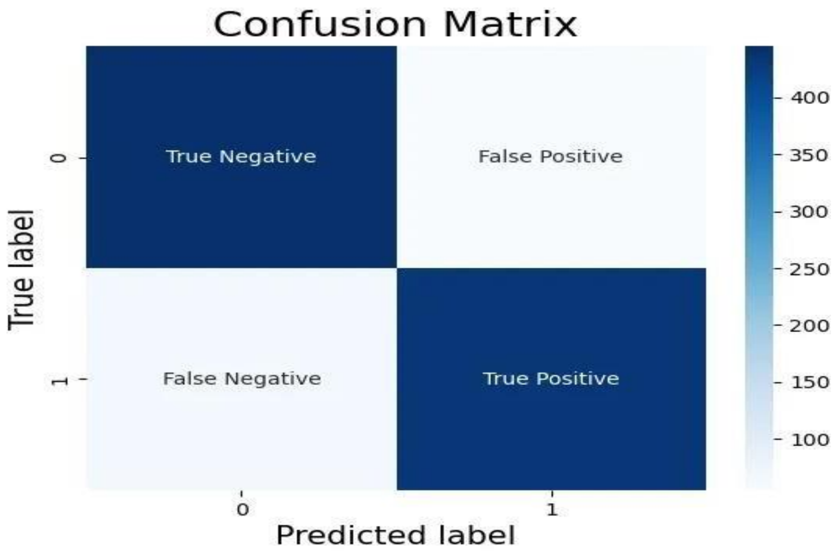


Figure 2.4: Representation of Confusion Matrix

We can analyze this matrix to compute various performance measures. These measures include accuracy, precision, recall, and F1 score. Each measure offers different information about the model's strengths and weaknesses [68].

- **Accuracy:** Accuracy stands as the most frequently employed performance metric for assessing a binary classification model [80]. It evaluates the performance of the model by calculating the ratio of correctly classified instances out of the total number of instances. In other words, accuracy indicates the proportion of all instances that the model has classified correctly [69]. The formula used to calculate Accuracy is as follows:

  Accuracy = (TP + TN) / (TP + TN + FP + FN)

- **Precision:** Precision is a metric that quantifies the ratio of true positive (TP) instances among those predicted as positive by the model. In other words, precision evaluates how accurately the model identifies positive predictions. A high precision score indicates accurate identification of positive predictions; while a low precision score suggests the model makes numerous false positive (FP) predicted values [67].

  The formula used to calculate precision is as follows:
  Precision = TP / (TP + FP)
- **Recall:** Recall measures the effectiveness of a classification model in identifying all relevant predictions from a dataset [69]. Specifically, recall is an evaluation metric that calculates the rate of positive predictions accurately detected by a binary classification model out of all actual positive predictions [67].

  The formula used to calculate Recall is as follows:
  Recall = TP / (TP + FN)
- **F1-Score:** The F1-score is an evaluation metric that merges precision and recall, offering a comprehensive assessment of a binary classification model's effectiveness. It calculates the harmonic mean of recall and precision, giving both metrics equal significance. This balanced approach ensures that the F1-score reflects both the accuracy of positive instances (precision) and the capability to determine all positive predictions (recall) [67].

  The formula used to calculate F1-score is as follows:
  F1-score = 2 * (precision * recall) / (precision + recall)

In conclusion, evaluation metrics are crucial for evaluating the performance of classification models. Various metrics are available and selecting the appropriate one(s) relies on the specific problem, the cost of false positives and false negatives, and the degree of dataset imbalance.

Precision is useful when the cost of false positives is high, and recall is important when the cost of false negatives is significant. The F1-score offers an equitable measure of both recall and precision. While accuracy is the most frequently used evaluation metric, a high accuracy score signifies that the model makes a large proportion of correct predictions, whereas a low accuracy score indicates that the model makes too many incorrect predictions. [67].

3.3.1.2.Regression Analysis:

Regression analysis involves various machine learning techniques that predict a continuous (y) outcome variable depends on the values of one or more (x) predictor variables. The primary distinction between regression and classification lies in their predictions: classification forecasts discrete class labels, while regression predicts a continuous value. Figure 2.3 illustrates the distinctions between classification and regression models. Regression models are widely applied in several domains, including: cost estimation, financial forecasting, trend analysis, marketing, time series prediction, and drug response modeling. Popular regression algorithms such as linear regression, polynomial regression, lasso regression, and ridge regression [63].

3.3.2. Unsupervised Learning:

Unsupervised learning analyzes unlabeled datasets without human interaction, following a data-driven process [63]. It is utilized to identify meaningful patterns, structures, and trends within the data, and is commonly applied in tasks such as clustering, density estimation, feature learning, and dimensionality reduction.

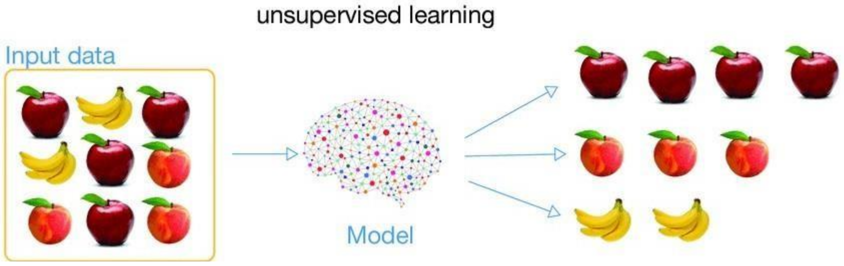


Figure 2.5: Example of unsupervised learning

Unsupervised learning problems are categorized into clustering and association tasks.

3.3.2.1.Clustering

Cluster analysis, or clustering is a technique in unsupervised machine learning that identifies and groups similar data points within large datasets without focusing on particular outcomes. This method groups objects such that those within the same cluster exhibit greater similarity to each other than to those in different clusters. It is often utilized in data analysis to uncover interesting trends or patterns, suchas identifying consumers groups depending on their actions [63].

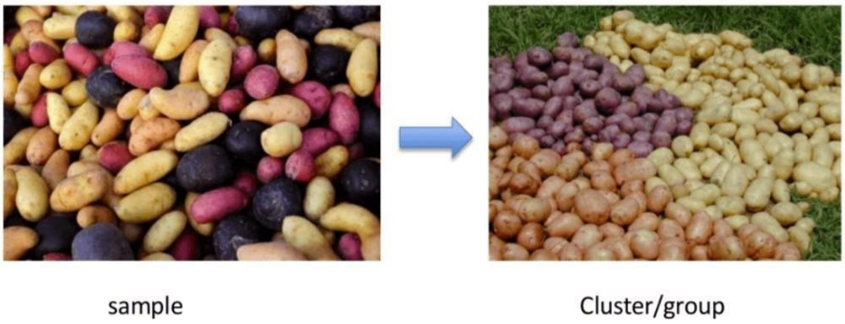


Figure 2.6: Clustering technique

Clustering is applicable in various fields, such as e-commerce, cybersecurity, mobile data processing, user modeling, health and behavioral analytics.

3.3.2.1.1. Different types of clustering methods:

Below, we briefly discuss different types of clustering methods.

- **Partitioning Methods:** This approach categorizes data into several clusters or groups according to their features and correspondences. Data scientists or analysts decide the number of clusters based on the specific requirements of the application, either dynamically or statically. K-means is the most widely used clustering algorithm based on partitioning methods [70].
- **Density-based Methods**: These methods distinguish clusters by recognizing contiguous regions of high point concentration in the data space, which are isolated by regions of low point concentration. Points that do not belong to any cluster are regarded as noise. Typical density-based algorithms include DBSCAN and OPTICS. These methods frequently encounter challenges when dealing with clusters that have similar density and high-dimensional data [71].
- **Hierarchical-based Methods:** Hierarchical clustering builds a cluster hierarchy in a tree-like structure. Strategies are generally of two types: (i) Agglomerative: a "bottom-up" approach where individual observations begin in separate clusters, which are then merged as the hierarchy progresses upwards, and (ii) Divisive: a "top-down" approach where all observations begin in a single cluster, and splits are recursively performed as the hierarchy descends [72].

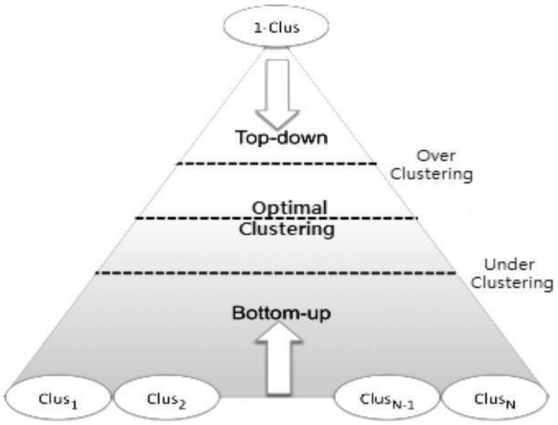


Figure 2.7: A graphical interpretation of the widely-used hierarchical clustering (Bottom-up and top down) technique.

- **Grid-based Methods:** Grid-based clustering is highly effective for managing extensive datasets. This method entails representing the dataset with a grid structure and subsequently merging grid cells to create clusters [60].
- **Model-based Methods**: There are primarily two types of model-based clustering algorithms: those employing statistical learning and those utilizing neural network learning. An instance of a statistical learning method is the Gaussian Mixture Model (GMM). [73].

3.3.2.2.Association:

Association rules enable the relationships between data elements in extensive databases. This unsupervised method focuses on uncovering meaningful relationships between objects within extensive datasets. For example, people buying new homes are most likely to buy new furniture [74].

Other Examples:

- Subgroups of cancer patients categorized by their gene expression profiles.
- Consumer groups segmented by their browsing and purchasing behaviors.
- Movie categories based on viewer ratings.

3.3.3. Semi-Supervised Learning:

Semi-supervised learning integrates objects of both supervised and unsupervised learning, working with both unlabeled and labeled data [63][64]. Thus, it lies between learning “without supervision” and learning “with supervision” [61]. The primary aim of semi-supervised learning models is to enhance predictive performance above what can be achieved using just labeled data. Applications encompass machine translation, data labeling, fraud detection, and text classification.

3.3.4. Reinforcement Learning:

Reinforcement learning enables software agents and machines to determine optimal behaviors within a specific context or environment to enhance efficiency [75]. This approach is environment-driven and based on a system of rewards and penalties [61]. It is particularly effective for tasks that require high levels of automation or operational optimization, such as robotics, autonomous driving, and supply chain logistics. However, it is not ideal for straightforward problems.

Therefore, to construct robust models across diverse application domains, various machine learning techniques assume critical roles based on their learning capacities, data characteristics, and the desired outcomes. In Table 2.1, we present various machine learning techniques, including examples. [60]. Subsequent sections offer a comprehensive exploration of machine learning algorithms designed to enhance the intelligence and capabilities of data-driven applications.

Learning type	Model building	Examples
Supervised	Algorithms or models learn from labeled data (task-driven approach).	Classification, Regression.
Unsupervised	Algorithms or models learn from unlabeled data (Data-Driven Approach)	Clustering, Associations, Dimensionality reduction.
Semi-supervised	Models are built using combined data (labeled +unlabeled)	Classification, Clustering.
Reinforcement	Models are based on reward or penalty (environment-driven approach)	Classification, Control.

Table 2.1: Various types of machine learning techniques with examples [60].

3.4.Machine Learning Algorithms:

In this part, we present different machine learning algorithms,
Figure 2.4 illustrates the general framework of a machine learning-based predictive model. n the first phase, the model is trained using historical data, and in the second phase, the model generates outcomes for new test data.

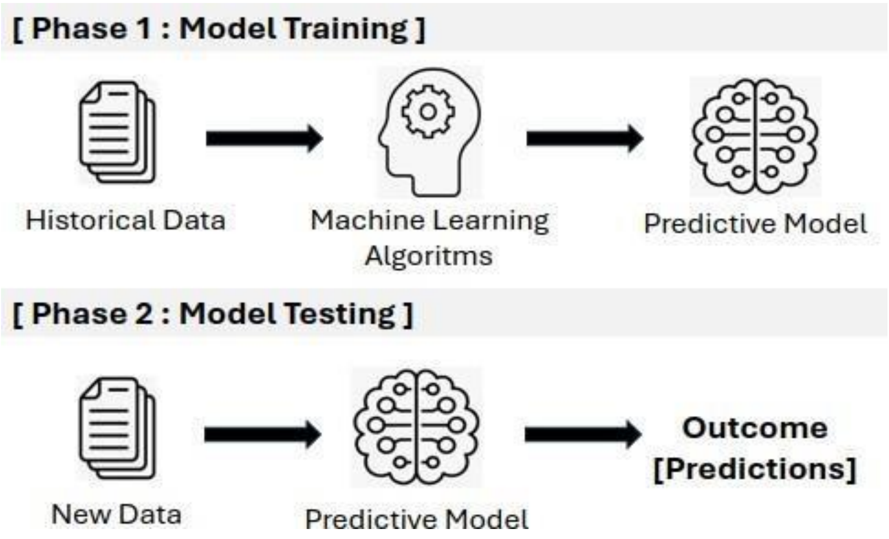

[Phase 1 : Model Training]
Historical Data
Machine Learning Algoritms
Predictive Model
[Phase 2 : Model Testing]
New Data
Predictive Model
Outcome [Predictions]

Figure 2.8 : A general structure of machine learning based predictive model considering both the training and testing phase [60].

3.4.1. Supervised Machine Learning Algorithms:

Numerous supervised algorithms have been introduced in the fields of machine learning and data science. Here, we provide an overview of the most commonly used methods across different application areas.

3.4.1.1.K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) [76] is an "instance-based learning" or non-generalizing learning algorithm, often referred to as a "lazy learning" algorithm. In place of constructing a typical internal framework, KNN stores all instances corresponding to the training data in n-dimensional space. It classifies new data points based on the collective decision of the labels assigned by the k nearest neighbors of each point. It uses measures such as the Euclidean distance function [65] to assess similarity. KNN is robust against noisy training data and its performance depends on data quality. One challenge with KNN is determining the ideal number of neighbors. It is suitable for both classification and regression tasks.

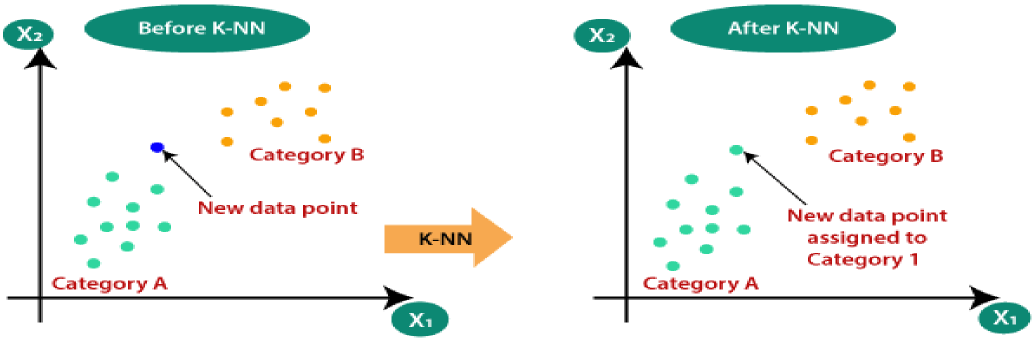


Figure 2.9 :K-nearest neighbors (KNN) algorithm [77].

3.4.1.2.Support Vector Machine (SVM):

The Support Vector Machine (SVM) is a supervised machine learning method utilized for classification, regression, and other tasks [78]. SVM constructs hyper-planes in high- or infinite-dimensional space to facilitate classification. The optimal hyper-plane maximizes the distance from the closest training data points of each class, which helps in minimizing the error in generalizing by the classifier.

SVMs are particularly efficient in high-dimensional spaces and their performance varies depending on the kernel functions employed. Commonly used kernel functions include linear, polynomial, radial basis function (RBF), and sigmoid [65]. However, SVMs may struggle with performance whenthe dataset contains a significant amount of noise, like overlapping target classes.

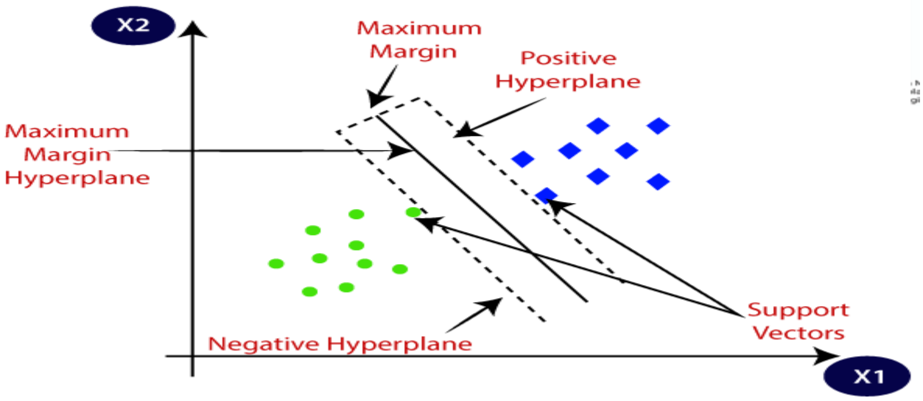


Figure 2.10: Support vector machine [79].

3.4.1.3.Decision Tree (DT)

Decision Tree (DT) [80] is a widely recognized non-parametric supervised learning technique utilized for both classification and regression tasks. DT categorizes instances by traversing down the tree from the root to leaf nodes. At each node, it evaluates the attribute defined for that node and directs instances along the branch of the tree that matches the attribute's value [65].A decision tree poses a question and then divides into subtrees based on the answer (Yes/No) [79].

The diagram below illustrates the typical structure of a decision tree.

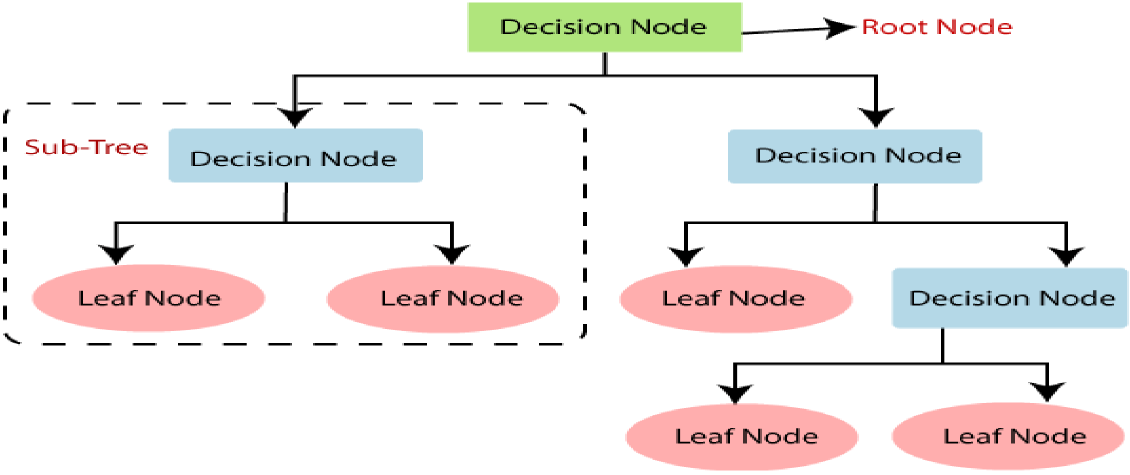


Figure 2.11: Decision tree structure [79].

3.4.1.4.Random Forest (RF):

The Random Forest classifier is a widely recognized ensemble classification method extensively utilized in machine learning and data science across diverse application domains.

This approach applies, "parallel ensembling," which involves training multiple decision tree classifiers simultaneously on different portions of the data, as illustrated in Figure 2.9 The outcomes are determined through either majority voting or averaging, aimed at reducing overfitting while enhancing prediction accuracy and control [65]. As a result, the Random Forest (RF) learning model, integrating multiple decision trees, generally outperforms single decision tree-based models [81]. RF achieves controlled variation in its decision trees by utilizing bootstrap aggregation (bagging) combined with random feature selection [82]. This approach is versatile, applicable to both classification and regression tasks, and performs effectively with categorical and continuous data.

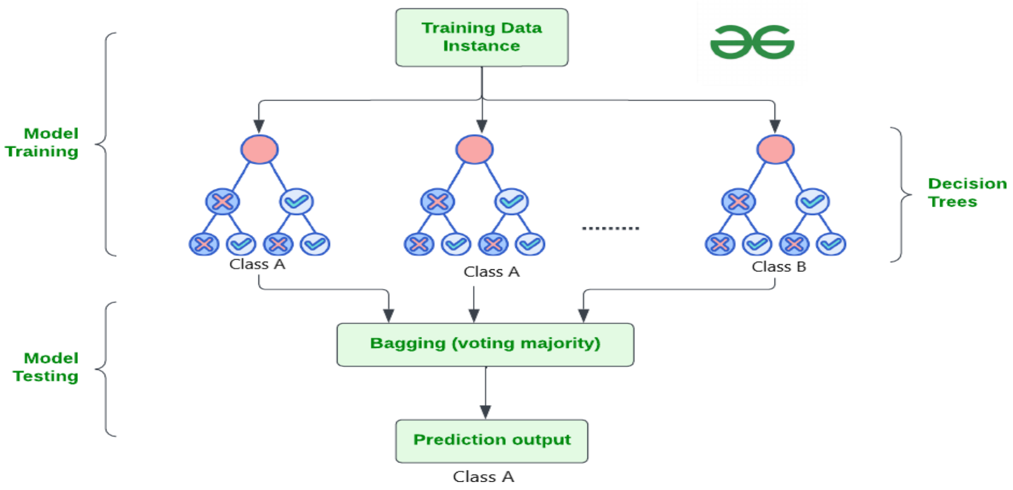


Figure 2.12: a random forest structure considering multiple decision trees.

3.4.1.5.Extreme Gradient Boosting (XGBoost):

Gradient Boosting, and Random Forests are similar [83], is an ensemble learning technique that creates a final model from a Sequence of separate models, typically decision trees. It utilizes the gradient to reduce the loss function, similar to the way neural networks [63] Employ gradient descent for optimizing weights. Extreme Gradient Boosting (XGBoost) is a variant of gradient boosting that makes finer approximations to find the optimal model. It calculates second-order gradients of the loss function to reduce loss and employs advanced methods for regularization (L1 and L2) [65], which help to reduce overfitting and improve model generalization and effectiveness. XGBoost is efficient, easy to understand, and capable of handling extensive datasets effectively. Figure 2.10 explains the generalstructure of Exterme Gradient Bosting algorithm.

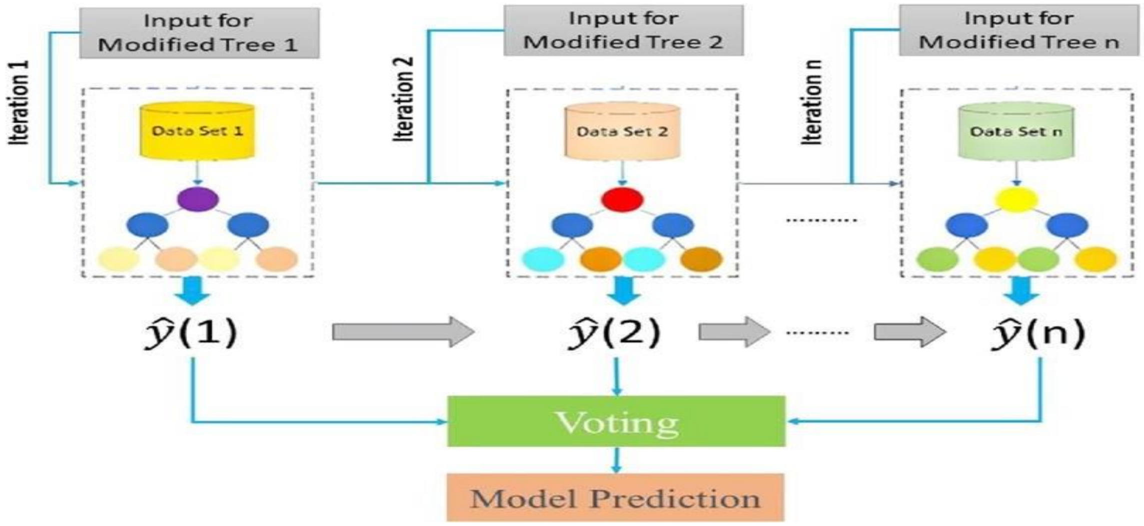


Figure 2.13: Exterme Gradient Boosting algorithm [84].

3.4.2. Unsupervised Machine Learning Algorithms:

Numerous unsupervised clustering algorithms have been introduced to effectively group data in machine learning and data science literature Below, we provide an overview of some widely used techniques across diverse fields of application.

3.4.2.1.K-means Clustering:

K-Means Clustering is an unsupervised learning algorithm that organizes an unlabeled dataset into distinct clusters. The value of K specifies the number of clusters to be formed; for instance, if K=2, the data will be grouped into two clusters, and if K=3, it will be divided into three clusters, and so on [54].

This algorithm assigns data points to clusters by minimizing the squared distance between each data point and its respective cluster centroid. It identifies k centroids and assigns every data point to the nearest cluster, aiming to keep the centroids as compact as possible. Because it initially selects cluster centers randomly, the outcomes can vary. K-means can be influenced by outliers, as extreme values can significantly affect the mean [85].

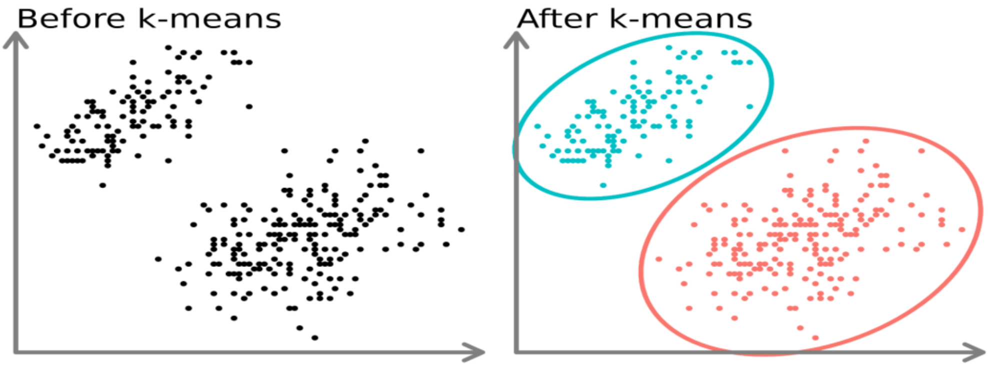


Figure 2.14 : Example of K-means algorithm

3.4.2.2.GMM Clustering:

Gaussian Mixture Models (GMMs) are commonly used for data clustering and represent a distribution-based clustering algorithm. A GMM is a probabilistic model where data points are generated by a mixture of a finite number of Gaussian distributions with unknown parameters. To determine the Gaussian parameters for each cluster, the Expectation-Maximization (EM) algorithm can be employed [65]. EM is an iterative method that uses a statistical model to estimate the parameters. Unlike K-means, GMMs account for uncertainty and provide the likelihood that a data point belongs to one of the k clusters. GMM clustering is more robust than K-means and performs well with non-linear data distributions.

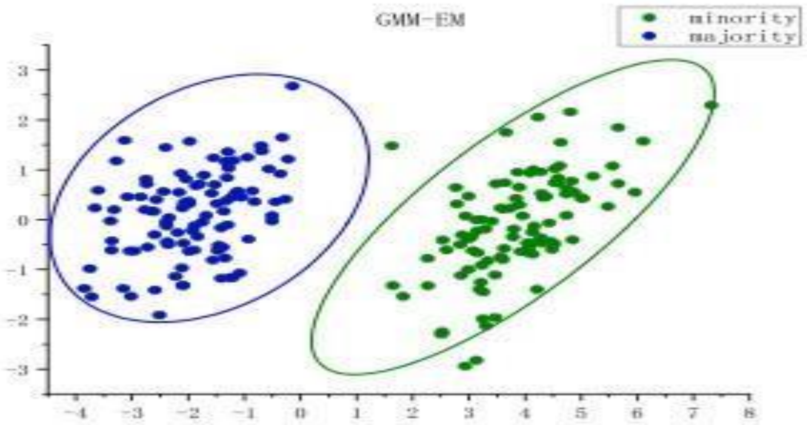


Figure 2.15 : GMM-Expectation-maximization (EM) [86].

3.5. Dimensionality Reduction and Feature Learning:

High-dimensional data processing is a significant challenge in machine learning and data science. Dimensionality reduction, an unsupervised learning technique, is crucial as it enhances human interpretation, reduces computational costs, and prevents overfitting and redundancy by simplifying models. This can be achieved through feature selection and feature extraction. The main difference between these processes is that “feature selection” [87] retains a subset of the original features, while “feature extraction” creates new ones [88].

3.5.1. Feature Selection:

Feature selection, also known as variable or attribute selection, involves choosing a subset of unique features (variables, predictors) for building machine learning models. It reduces model complexity by eliminating irrelevant or less important features, facilitating faster training. An optimal subset of selected features can minimize overfitting, simplify and generalize the model, and increase its accuracy [87]. Feature selection is a crucial concept in machine learning, significantly impacting the model's effectiveness and efficiency [89][90]. Techniques like the Chi-squared test, ANOVA test, Pearson’s correlation coefficient, and recursive feature elimination are popular methods for feature selection.

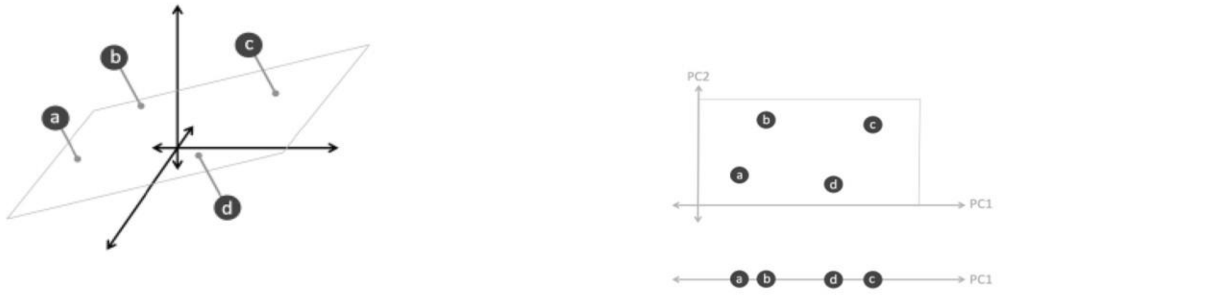
3.5.2. Feature Extraction:

Feature extraction techniques enhance data understanding, improve prediction accuracy, and reduce computational cost or training time in machine learning models. The goal is to reduce the number of features by generating new ones from the existing ones and then discarding the originals. This new reduced set of features captures most of the information from the original set [89][90].

For instance, Principal Component Analysis (PCA), is commonly used for dimensionality reduction, creating new components from existing features [88].

3.5.2.1.Principal Component Analysis (PCA):

Principal Component Analysis (PCA) is a well-known unsupervised learning technique in machine learning and data science. PCA transforms a set of correlated variables into a set of uncorrelated variables called principal components. For example, Figure 14 shows the effect of PCA, where Figure 14a depicts the original features in 3D space, and Figure 14b shows the principal components PC1 and PC2 on a 2D plane, and PC1 on a 1D line. PCA reduces the dimensionality of datasets, helping to build effective machine learning models [76]. Technically, PCA identifies the principal components with the highest eigenvalues from a covariance matrix and uses them to project data into a new subspace with equal or fewer dimensions [65].



(a) An example of the original features in a 3D space. (b) Principal components in 2D and 1D space.

Figure 2.16 : An example of a principal component analysis (PCA) and created principal components PCA1 and PCA2 in different dimension space [60].

4. Deep Learning:

4.1.Definition:

Deep learning is a subset of artificial neural networks (ANN) within the broader field of machine learning, characterized by its use of representation learning. It provides a computational architecture that combines multiple processing layers, including input, hidden, and output layers, to learn from data [63]. The primary advantage of deep learning over traditional machine learning methods is its superior performance, especially when working with large datasets [64].

Figure2.18 shows a general performance of deep learningover machine learning considering the increasing amount of data. However, it may vary depending on the data characteristics and experimental set up.

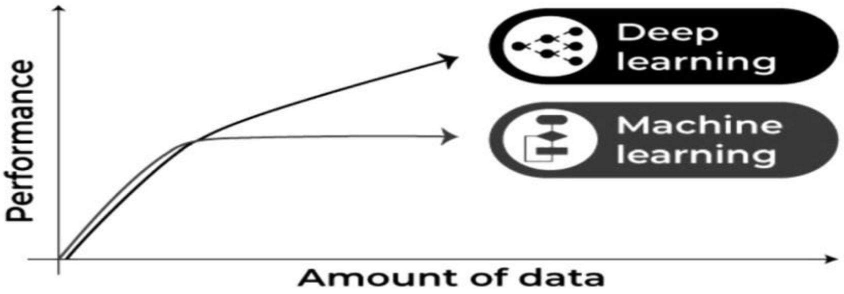


Figure 2.17 : Machine learning and deep learning performance in general with the amount of data [60].

4.2.Deep Learning Algorithms:

In the following, we present various types of deep learning methods that canbe used to build effective data-driven models for various purposes.

4.2.1. The Convolutional Neural Network (CNN):

The Convolutional Neural Network (CNN) enhances the design of the standard ANN by incorporating convolutional layers, pooling layers, and fully connected layers [], as illustrated in Figure 2.19. It was inspired by biological neural networks. By leveraging the two-dimensional (2D) structure of input data, CNNs are widely used in various fields, including image and video recognition, image processing and classification, medical image analysis, and natural language processing [60].

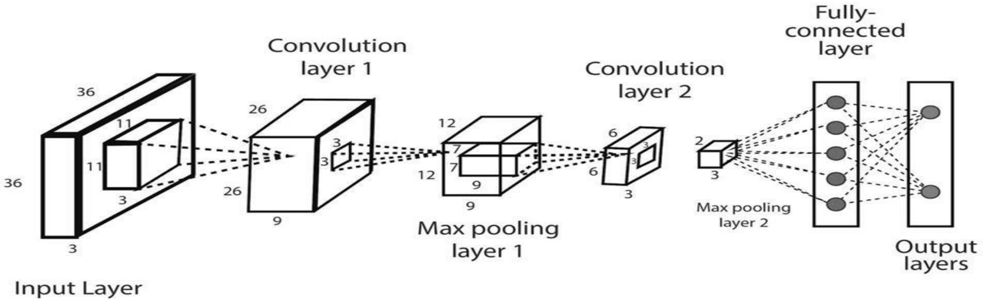


Figure 2.18: An example of convolutional neural network (CNN) including multiple convolutions and pooling layers.

4.2.2. Recurrent Neural Networks (RNN):

Recurrent Neural Networks (RNNs) are designed to process sequential data. Unlike traditional feedforward neural network models, such as CNNs, where data flows from the input layer to the hidden layer and then to the output layer without connections between neurons in the same layer, RNNs address this limitation. Feedforward architectures cannot handle problems where input data points are related to each other.

Points are related to each other. For instance, predicting the next word in a sentence requires knowledge of the previous words, as the words in a sentence are not independent of each other [83].

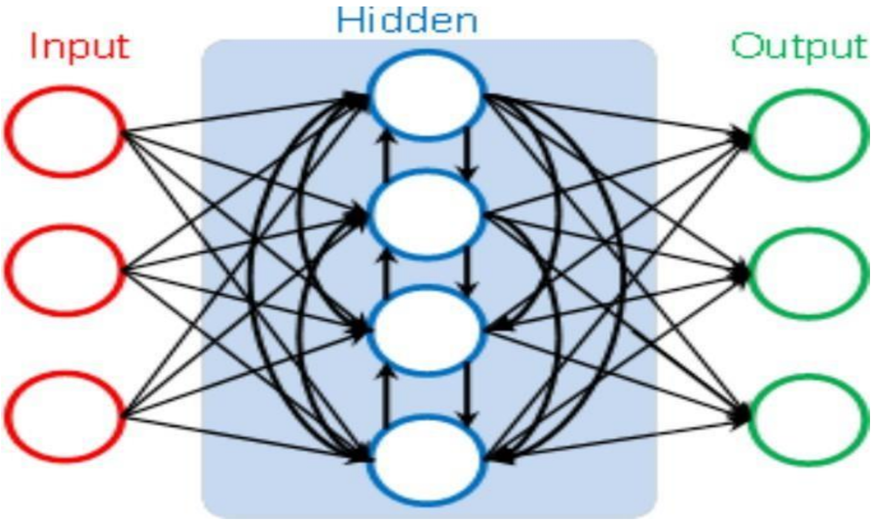


Figure 2.19 : A structure of an artificial neural network modeling with multiple processing layers.[83]

5. Conclusion :

In summary, Artificial intelligence, machine learning, and deep learning are fundamentally about machine perception the ability to interpret sensory data. Training algorithms use supervised learning to label data and unsupervised learning to group data. The key difference between supervised and unsupervised learning is whether a labeled training set is used.

In this chapter, we have detailed the basic concepts necessary for our work, starting with definitions of machine learning and an overview of some algorithms. We then outlined deep learning and its algorithms. In the next chapter, we will apply the architecture of our proposed approach and treat areal problematic which concern a gender speaker recognition.

Chapter 3 :

Applying Machine Learning Algorithms for Gender Recognition from Voice Samples

Chapter 3:

Applying Machine Learning Algorithms forGender Recognition from Voice Samples

1. Introduction:

In this chapter, we provide a comprehensive methodology for gender recognition from voice features using Python. The chapter is divided into three main parts: preprocessing, modeling and real-time gender classification.

The first part details the preprocessing steps taken to prepare the audio data, ensuring its quality and consistency.

The second part focuses on modeling. We evaluated various machine learning algorithms, bothsupervised and unsupervised, to determine the most accurate for gender classification based on voicefeatures. We provide a detailed analysis of the results, concluding with the identification of the mostsuitable algorithm for our task.

The third part introduces the real-time gender classification application we developed. This section focuses on the application's design, particularly its graphical user interface (GUI). The application uses the most accurate machine learning model for gender classification.

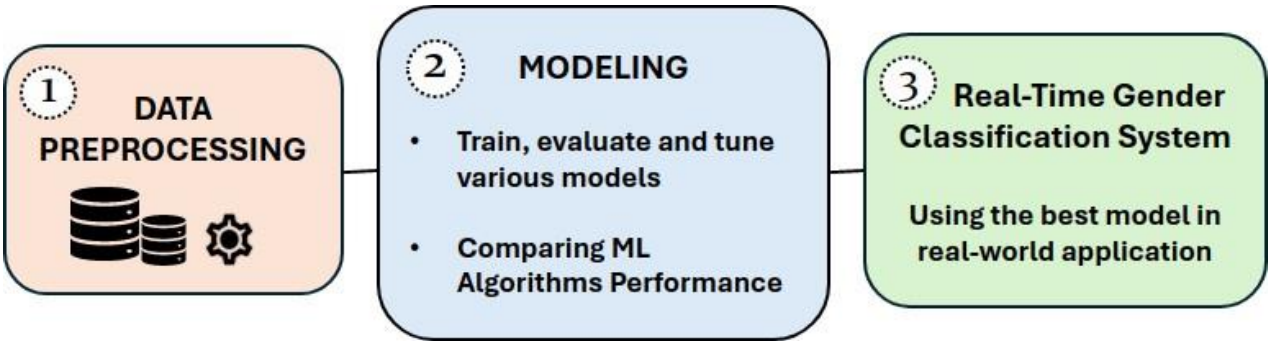

1
DATA PREPROCESSING
2
MODELING
• Train, evaluate and tune various models
• Comparing ML Algorithms Performance
3
Real-Time Gender Classification System
Using the best model in real-world application

Figure 3.1.1: The Main Steps of the Machine Learning Pipline

2. Preprocessing:

Before modeling, it is crucial to preprocess the audio data to ensure their quality and consistency, thereby improving the effectiveness of model training. This section outlines the preprocessing steps we have applied, as illustrated in the figure below.

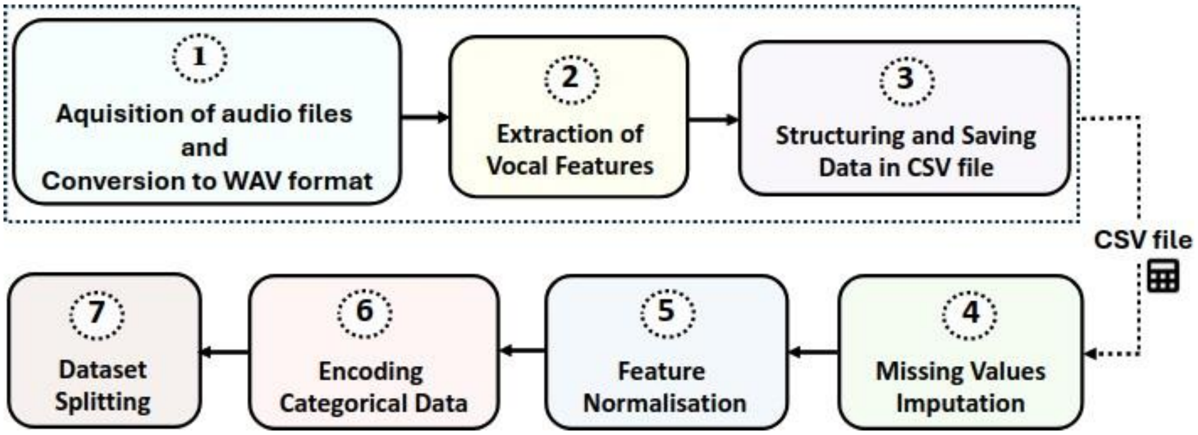

1
Aquisition of audio files and Conversion to WAV format
2
Extraction of Vocal Features
3
Structuring and Saving Data in CSV file
CSV file
4
Missing Values Imputation
5
Feature Normalisation
6
Encoding Categorical Data
7
Dataset Splitting

Figure 3.2: The Data Preprocessing Pipeline.

- **Dataset Description :**

The dataset consists of 8,422 raw audio waveforms sourced from the VoxForge audio database, an open-source speech recognition corpus. Each audio sample, submitted by users using their personal microphones, has a duration between 3 to 6 seconds and is approximately 200 kilobytes in size. The samples are evenly distributed between genders, with 4,211 male and 4,211 female samples, ensuring a balanced dataset (see figure 3.3).

Each sample is recorded as a ".WAV" file and pre-processed to extract 401 acoustic features necessary for gender classification. These processed samples are then saved into a "CSV" file comprising 8,422 rows and 402 columns: 401 feature columns, 1 gender label column (male or female).

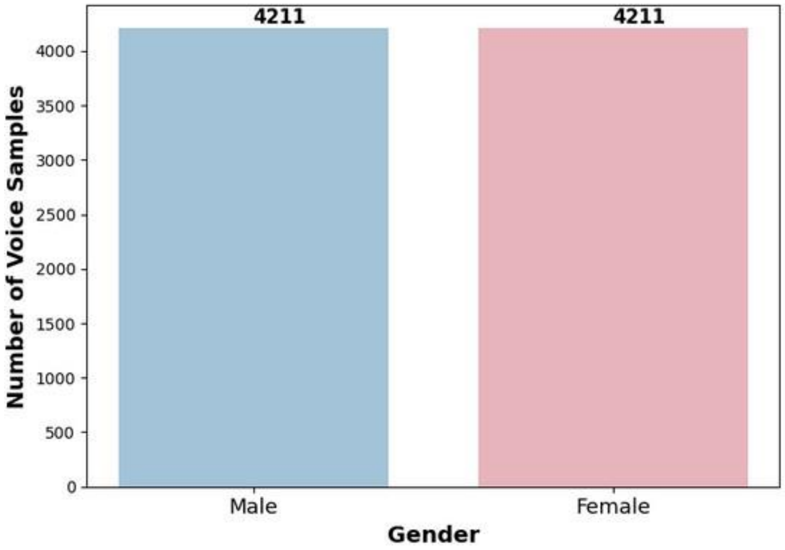
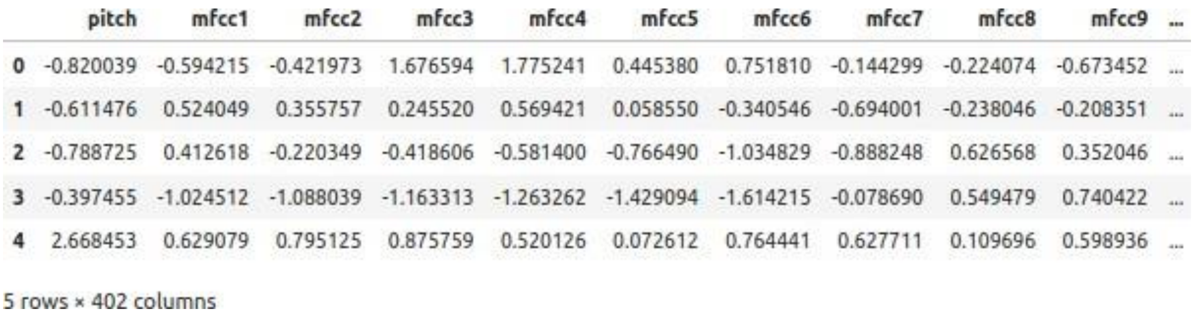


Figure 3.3 : Disribution of Voice Samples by Gender in the Dataset.



	pitch	mfcc1	mfcc2	mfcc3	mfcc4	mfcc5	mfcc6	mfcc7	mfcc8	mfcc9	...
0	-0.820039	-0.594215	-0.421973	1.676594	1.775241	0.445380	0.751810	-0.144299	-0.224074	-0.673452	...
1	-0.611476	0.524049	0.355757	0.245520	0.569421	0.058550	-0.340546	-0.694001	-0.238046	-0.208351	...
2	-0.788725	0.412618	-0.220349	-0.418606	-0.581400	-0.766490	-1.034829	-0.888248	0.626568	0.352046	...
3	-0.397455	-1.024512	-1.088039	-1.163313	-1.263262	-1.429094	-1.614215	-0.078690	0.549479	0.740422	...
4	2.668453	0.629079	0.795125	0.875759	0.520126	0.072612	0.764441	0.627711	0.109696	0.598936	...

5 rows × 402 columns

Figure 3.4: Preview of the First Five Rows of the Preprocessed Dataset.

2.1.Acquisition and Conversion of Audio File Format:

The foundation of our analysis begins with a data collection strategy focused on gatheringa diverse set of voice recordings. To ensure consistency and compatibility across the dataset, all collected audio files, regardless of their initial format, are converted to the ".WAV" format. This conversion process is vital as ".WAV" files offer uncompressed audio data, which is preferable for maintaining the integrity of the original recordings during processing.

We employ the pydub library for this purpose due to its versatility in seamlessly handling various audio formats. Each audio file undergoes format verification, and if not already in "WAV" format, it is converted using pydub. This standardization process not only preserved theoriginal quality of the audio recordings but also simplified subsequent analytical procedures.

2.2.Extraction of Vocal Features:

The next critical phase in our methodology is extracting meaningful acoustic features from each audio file, isolating characteristics that most indicate gender differences in speech. Thesevocal features are crucial for the performance of the machine learning learning models. As mentioned in Chapter 1, we focused on Mel-frequency Cepstral Coefficients (MFCCs) and pitch as the primary features.

2.2.1. MFCC Coefficients Extraction:

The process of MFCC extraction begins by resampling the audio to 8 kHz, focusing on frequencies essential for speech analysis.

Following resampling, we adjust the length of the audio samples to a uniform 40000 samples. This length adjustment is crucial for ensuring uniformity across the dataset, involvingboth truncating longer samples and padding shorter ones.

Using the *librosa.feature.mfcc()* function from the librosa library, we then compute a set of10 Mel-frequency Cepstral Coefficients (MFCCs).

Carefully specifying parameters for MFCC extraction is essential for achieving optimal results. We set a window size (n_fft) of 2048 in the Fourier transform to balance frequency andtemporal resolution. Additionally, a hop length of 1024 is set to control the spacing between successive frames.

The result of the MFCC extraction is a comprehensive feature vector comprising 400 values for each audio sample. Each value corresponds to a specific MFCC coefficient.

2.2.2. Pitch Extraction:

We employ the probabilistic YIN (pYIN) algorithm from the librosa library to extract pitch, specifically targeting the typical frequency range of human speech. This focus ensures we analyze the most relevant vocal frequencies.

The extraction process starts by detecting pitches within this frequency range and filteringthem to include only those from voiced segments, where actual vocal activity is present.

From these filtered pitches, we extract the highest pitch in each segment, representing the dominant frequency. This method allows us to isolate key pitch features from each audio sample effectively.

2.3.Structuring and Saving Data in CSV file:

After feature extraction, the data was meticulously organized into a structured format. Each sample’s feature set, consisting of one pitch and 400 MFCCs, was combined with its corresponding gender label into a pandas DataFrame.

The data organization and consolidation process was optimized through parallel processing, using a ProcessPoolExecutor to handle multiple files simultaneously. This improved the efficiency of the process and scaled effectively with the size of the dataset.

Finally, the consolidated DataFrame was saved in a CSV file, providing a well-organized and accessible format for further analysis and model training.

2.4.Missing Values Imputation:

After loading the CSV dataset, we used the *fit_transform* method from *sklearn.impute.SimpleImputer* class to handle potential missing values, employing the 'mean' strategy for replacement. This approach involves replacing any missing values with the mean of the available values in each column excluding the target column containing the labels. By doing so, we prevent information loss.

2.5.Feature Scaling:

We performed feature normalization using the *fit_transform* method from the *sklearn.preprocessing.StandardScaler* class. This method centers the features around a mean of 0 and scales them to have a standard deviation of 1. By ensuring all features are on a comparable scale and preserving the shape of their distribution, normalization prevents any single feature from dominating the learning process. This contributes to more efficient model training and faster convergence.

2.6.Structuring and Saving Data in CSV file:

After feature extraction, the data was meticulously organized into a structured format. Each sample’s feature set, consisting of one pitch and 400 MFCCs, was combined with its corresponding gender label into a pandas DataFrame.

The data organization and consolidation process was optimized through parallel processing, using a ProcessPoolExecutor to handle multiple files simultaneously. This improved the efficiency of the process and scaled effectively with the size of the dataset.

Finally, the consolidated DataFrame was saved in a CSV file, providing a well-organized and accessible format for further analysis and model training.

2.7.Missing Values Imputation:

After loading the CSV dataset, we used the *fit_transform* method from *sklearn.impute.SimpleImputer* class to handle potential missing values, employing the 'mean' strategy for replacement. This approach involves replacing any missing values with the mean of the available values in each column excluding the target column containing the labels. By doing so, we prevent information loss.

2.8.Feature Scaling :

We performed feature normalization using the *fit_transform* method from the *sklearn.preprocessing.StandardScaler* class. This method centers the features around a mean of 0 and scales them to have a standard deviation of 1. By ensuring all features are on a comparable scale and preserving the shape of their distribution, normalization prevents any single feature from dominating the learning process. This contributes to more efficient model training and faster convergence

2.9.Encoding Categorical Data:

In our CSV dataset, the categorical labels within the target column "gender" required numerical representation to align with machine learning algorithms. By employing the *fit_transform* method from the *sklearn.preprocessing.LabelEncoder* class, we converted this categorical variable into numerical equivalents, assigning '1' to 'male' and '0' to 'female'. This conversion ensures that machine learning algorithms can effectively interpret gender data.

2.10. Dataset splitting:

To evaluate the performance of our machine learning models, we split the dataset into separate training and testing sets. This approach allows us to assess the models' performance on unseen data, thereby minimizing the risk of overfitting. Our dataset, consisting of 8422 samples, was divided into 75% for the training set and 25% for the test set.

3. Modeling:

After meticulously preprocessing our dataset, we moved to the modeling phase, aiming to develop machine learning models capable of accurately identifying gender based on the extracted voice characteristics. This stage involved the selection, training, and evaluation of various models.

3.1.Selection of Machine Learning Algorithms:

To identify the most effective model for our binary classification task, we explored both supervised and unsupervised learning methods.

A. Supervised Algorithms :

1. Support Vector Machines (SVM): Explored with four different kernels:
   - Linear Kernel.
   - Polynomial Kernel.
   - Radial Basis Function (RBF) Kernel.
   - Sigmoid Kernel.

2. k-Nearest Neighbors (k-NN).

3. Random Forest.

4. XGBoost.

B. Unsupervised Algorithms :

5. K-Means Clustering.
6. Gaussian Mixture Models (GMM).

3.2.Performance Evaluation Metrics:

To assess model performance, we used the accuracy, precision, recall, and F1-score metrics, derived from the confusion matrix, with accuracy being the primary metric for comparing the relative effectiveness of the selected models.

In addition to the standard evaluation metrics, we also included inference time, which is the duration required for a trained model to make a prediction once it's deployed for real-worlduse. The inference time was calculated by measuring the total time taken to make predictions and then dividing this duration by the number of samples in the test set.

For training and evaluating our machine learning models, we used a high-performance setup with the following hardware specifications:

- **Processor :** AMD Ryzen 5 5600X, featuring 6 cores and 12 threads.
- **Memory :** 32 GB RAM.
- **Graphics :** NVIDIA Quadro RTX 6000/8000 GPU.

This setup significantly reduced the inference time.

To further enhance the performance of our models, we performed hyperparameter tuning, an iterative process that involves adjusting various model parameters to find the optimal configuration that maximizes model performance (see Figure 3.5).

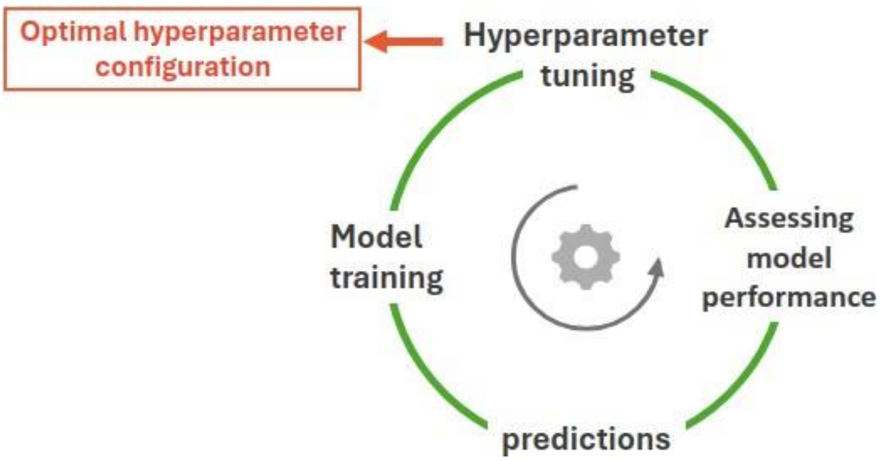

Optimal hyperparameter configuration
Hyperparameter tuning
Model training
Assessing model performance
predictions

Figure 3.5: Iterative Process of Hyperparameter Tuning.

3.3.Results and Discussion:

3.3.1. The Supervised Approach:

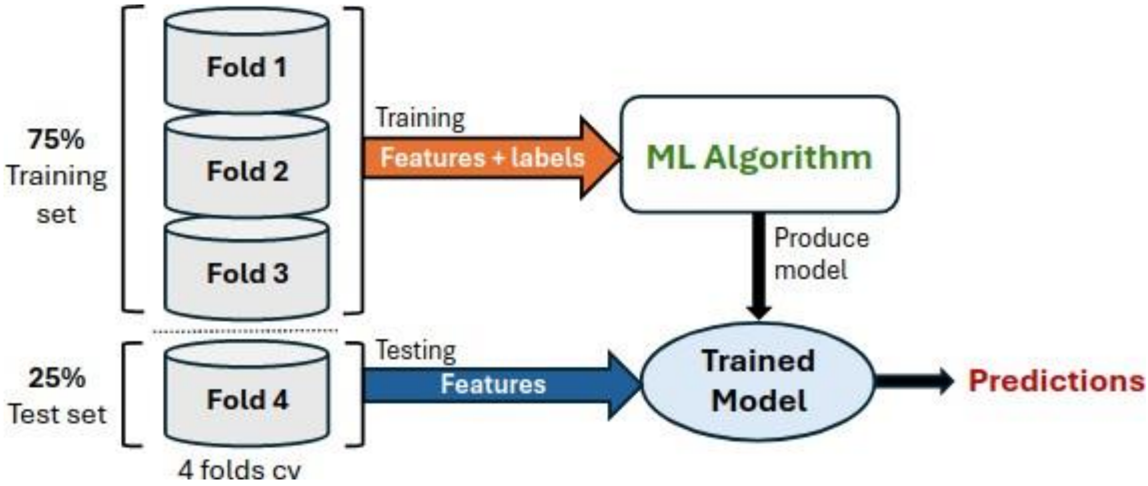


Figure 3.6: The process of training and testing supervised models.

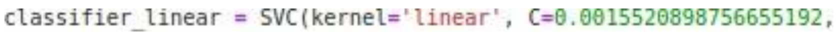
We used 4-fold cross-validation to evaluate the supervised models, ensuring robust performance assessment. Each model was trained on three folds and tested on the remaining fold, repeating the process four times to assess different sections of the data (see Figure 3.6).

A. Support Vector Machines (SVM):

- **Linear Kernel:**

In this analysis, we explore the performance of a Support Vector Machine (SVM) model using a linear kernel. The linear kernel is known for its simplicity and effectiveness in linearly separable data. However, our task's complexity, given the nuances in human voice data, demands a thorough examination and comparison with other kernels.

The best results were obtained with the following configuration (see Figure 3.7):



```
classifier_linear = SVC(kernel='linear', C=0.0015520898756655192,
```

Figure 3.7 : The best parameters of SVM classier with linear kernel.

The confusion matrix in figure 3.8 shows the following insights:

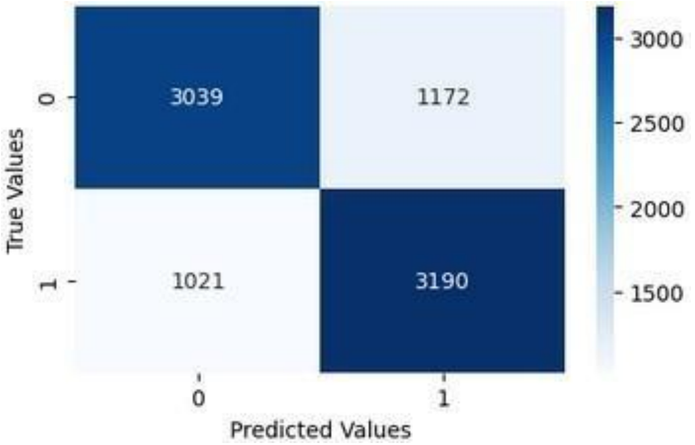


Figure 3.8: The confusion Matrix of the SVM model with the linear kernel.

- ✓ **True Positives (TP) for Class 1:** 3190 instances are correctly predicted as Class 1.
- ✓ **True Negatives (TN) for Class 0:** 3039 instances are correctly predicted as Class 0.
- ✓ **False Positives (FP):** 1172 instances are incorrectly classified as Class 1.
- ✓ **False Negatives (FN):** 1021 instances are incorrectly classified as Class 0.

The table below presents the performance metrics for the SVM model with a linear kernel.

Metric	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**Inference time** (milliseconds)
Value	**73.96**	73.99	73.96	73.95	0.88

Table 3.1 : Performance Evaluation Metrics of SVM with Linear Kernel

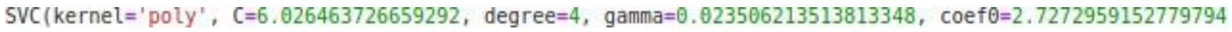
The Table 3.1 shows that the SVM model with a linear kernel shows balanced performance with accuracy, precision, recall, and F1-Score all around 74%, with an inference time of approximately 0.88 milliseconds, making it suitable for real-time applications.

The performance of the linear kernel serves as a benchmark for further comparisons with the other SVM kernels.

- **Polynomial Kernel** :

In continuing the exploration of SVM kernels for our classification task, we next evaluate the polynomial kernel.

The best results were obtained with the following configuration (see Figure 3.9):



```
SVC(kernel='poly', C=6.026463726659292, degree=4, gamma=0.023506213513813348, coef0=2.7272959152779794
```

Figure 3.9 : The Best Parameters of SVM classifier with polynomial kernel

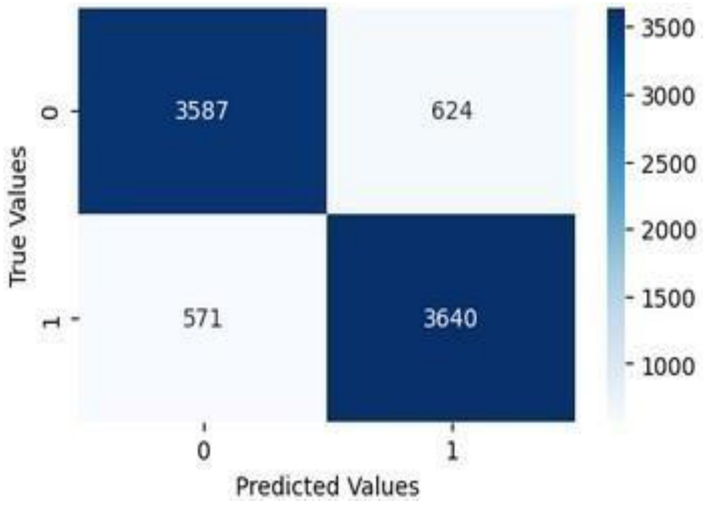


Figure 3.10 : The confusion Matrix of the SVM model with polynomial kernel.

The confusion matrix in figure 3.10 provides the following results:

3640 True Positives, 3587 True Negatives, 624 False Positives, and 571 False Negatives.

The following table presents the performance metrics for the SVM model employing a Polynomial kernel.

Metric	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**Inference time** (milliseconds)
Value	**85.81**	85.82	85.81	85.81	1.18

Table 3.2 : Performance Evaluation Metrics of SVM with Polynomial Kernel

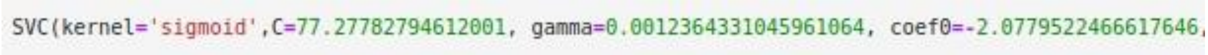
The SVM model with polynomial kernel demonstrates strong performance, with accuracy, precision, recall, and F1-Score all around 85.81% and an inference time of approximately 1.18 milliseconds, making it suitable for real-time applications.

These results, as presented in Table3.2 demonstrate an improvement over the linear kernel.

- **Sigmoid Kernel**:

In this section, we analyze the performance of a Support Vector Machine (SVM) model employing a sigmoid kernel.

The best results were obtained with the following configuration (see Figure 3.11):



```
SVC(kernel='sigmoid',C=77.27782794612001, gamma=0.0012364331045961064, coef0=-2.0779522466617646,
```

Figure 3.11 : The Best Parameters of SVM classifier with sigmoid kernel

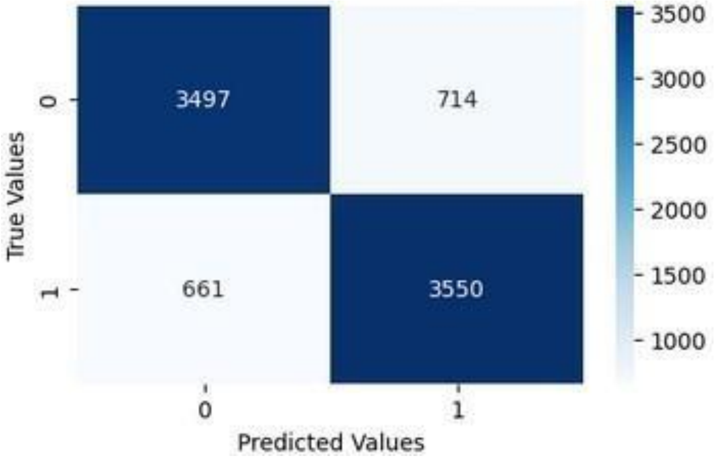


Figure 3.11.513: The confusion matrix of the SVM model with sigmoid kernel

The confusion matrix in figure 3.12 shows the following results:

3550 True Positives, 3497 True Negatives, 714 False Positives, and 661 False Negatives.

The following table presents the performance measures for the SVM model employing a Sigmoid kernel.

Metric	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**Inference time** (milliseconds)
Value	**83.67**	83.68	83.67	83.67	0.83

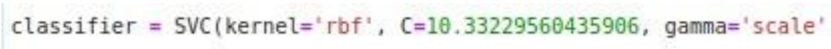
Table 3.3 : Performance Evaluation Metrics of SVM with sigmoid kernel

The table 3.3 shows that the SVM model with a sigmoid kernel achieves good performance,with accuracy, precision, recall, and F1-Score all around 83.67%, and an inference time of approximately 0.83 milliseconds. These results indicate that the sigmoid kernel is slightly lessperformant than the polynomial kernel.

- **Radial Basis Function (RBF) Kernel:**

The exploration of SVM kernels for gender classification from voice data extends to the Radial Basis Function (RBF) kernel.

The best results were obtained with the following configuration (see Figure 3.13):



```
classifier = SVC(kernel='rbf', C=10.33229560435906, gamma='scale'
```

Figure 3.13 : The best parameters of SVM classifier with RBF kernel

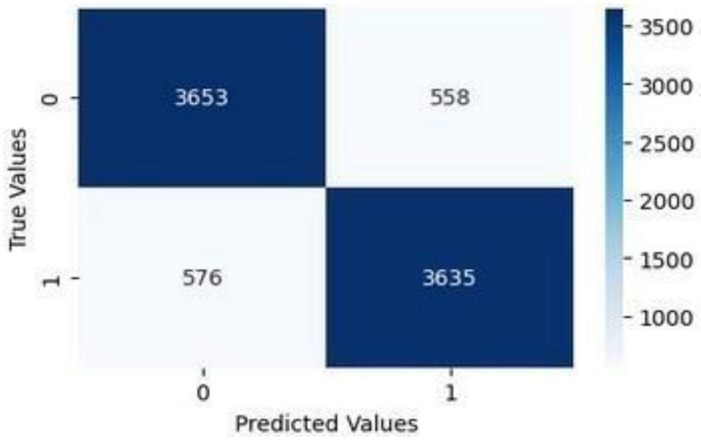


Figure 3.14: The Confusion Matrix of the SVM model with Gaussian Kernel

The confusion matrix in figure 3.14 shows the following results:

3635 True Positives, 3653 True Negatives, 558 False Positives, and 576 False Negatives.

The following table presents the performance measures for the SVM model employing a Gaussian kernel :

Metric	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**Inference time** (milliseconds)
Value	**86.54**	86.53	86.53	86.53	1.29

Table 3.4 : Performance Evaluation Metrics of SVM with RBF kernel

The table 3.4 shows that the SVM model with a RBF kernel achieves strong performance,with accuracy, precision, recall, and F1-Score all around 86.53%. The inference time is approximately 1.29 milliseconds. These results, as presented in Table 3.4, indicate that the Gaussian kernel is highly effective and efficient.

✓ **Selecting the Optimal SVM Kernel :**

SVM Kernel	Linear	Sigmoid	Polynomial	RBF
Accuracy	73.96 %	83.67 %	85.81%	**86.54 %**

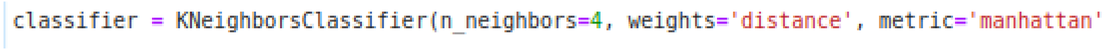
Table 3.5 : Accuracy Comparaison Across SVM kernels

As shown in Table 3.5, the Radial Basis Function (RBF) Kernel outperforms others with an accuracy of 86.54%, making it the optimal choice due to its ability to effectively handle non-linear relationships in data. It has therefore been selected for further comparative analysisagainst other machine learning models.

A. K-Nearest Neighbors (k-NN):

In this section, we evaluate the performance of the k-Nearest Neighbors (k-NN) model.

The best results were obtained with the following configuration (see Figure 3.15):



```
classifier = KNeighborsClassifier(n_neighbors=4, weights='distance', metric='manhattan'
```

Figure 3.15 : The best parameters with K-NN classifier.

The confusion matrix in figure 3.16 shows the following results:

3113 True Positives, 3706 True Negatives, 505 False Positives, and 1098 False Negatives.

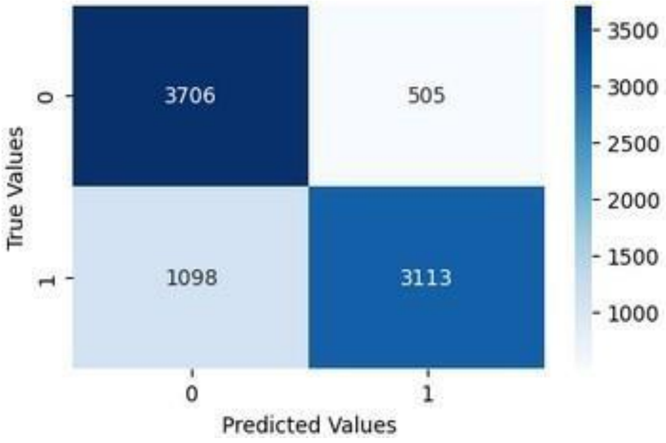


Figure 3.16 : The Confusion Matrix of the K-Nearest Neighbors model.

The following table presents the performance metrics for the k-Nearest Neighbors model:

Metric	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Inference time (milliseconds)
Value	**80.97**	81.59	80.97	80.87	0.70

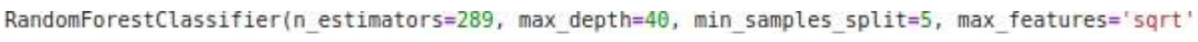
Table 3.6 : Performance Evaluation Metrics of k-Nearest Neighbors model

The table 3.6 shows that the k-NN model achieves good overall performance, with accuracy, precision, recall, and F1-Score all around 81%. The inference time is approximately 0.70 millisecondes. However, it performs slightly lower than the RBF kernel SVM.

B. Random Forest:

In this section, we evaluate the performance of the Random Forest model.

The best results were obtained with the following configuration (see Figure 3.17):



```
RandomForestClassifier(n_estimators=289, max_depth=40, min_samples_split=5, max_features='sqrt'
```

Figure 3.17 : The Best Parameters of the Random Forest Classifier

The confusion matrix in figure 3.18 shows the following results:

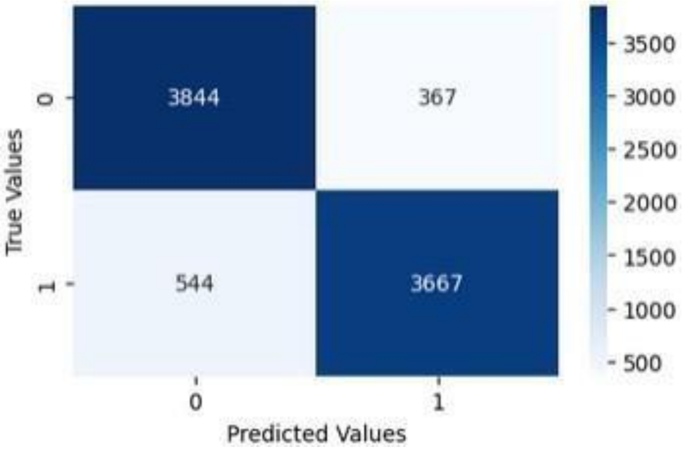


Figure 3.18 : The Confusion Matrix of the Random Forest model.

3667 True Positives, 3844 True Negatives, 367 False Positives, and 544 False Negatives.

The following table presents the performance metrics for the Random Forest model:

Metric	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**Inference time** (milliseconds)
Value	**89.18**	89.25	89.18	89.18	5.22

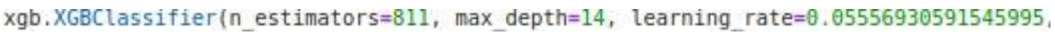
Table 3.7 : Performance evaluation metrics of the Random Forest model

The table 3.7 shows that the Random Forest model achieves strong performance, with accuracy, precision, recall, and F1-Score all around 89.18%. The inference time is approximately 5.22 milliseconds. These results indicate that the Random Forest model outperforms the other models in terms of accuracy. Although the inference time is longer compared to the k-NN and SVM models, it remains efficient for real-time applications.

C. XGBoost:

In this section, we explore the performance of the XGBoost model.

The best results were obtained with the following configuration (see Figure 3.19):



```
xgb.XGBClassifier(n_estimators=811, max_depth=14, learning_rate=0.05556930591545995,
```

Figure 3.19 : The best parameters of XGBoost classifier

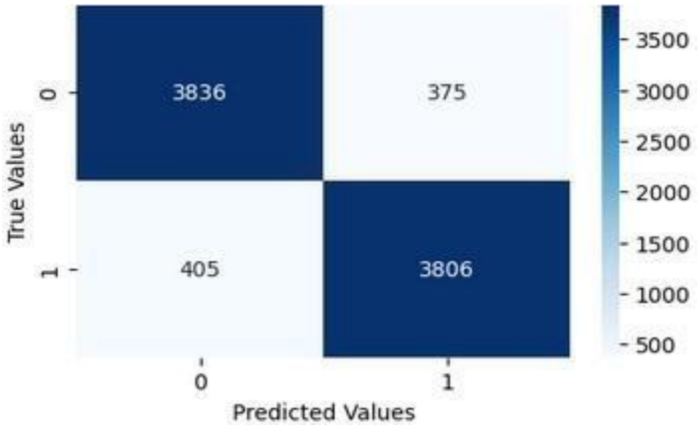


Figure 3.20 : The Confusion Matrix of the XGBoost model.

The confusion matrix in figure 3.20 shows the following results:

3806 True Positives, 3836 True Negatives, 375 False Positives, and 405 False Negatives.

The following table presents the performance metrics for the XGBoost model :

Metric	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**Inference time** (milliseconds)
Value	**90.74**	90.74	90.74	90.73	2.61

Table 3.8 : Performances Evaluation Metrics of XGBoost model

The table 3.9 shows that the XGBoost model achieves high performance, with accuracy, precision, recall, and F1-Score all around 90.74%. The inference time is approximately 2.61 milliseconds. These results indicate that the XGBoost model surpasses the other models in terms of accuracy. Despite a slightly longer inference time compared to some other models, it remains efficient and suitable for real-time applications.

3.3.2. The Unsupervised Approach:

A. Gaussian Mixture Model:

In this section, we evaluate the performance of a GMM-based system for our gender classification task.

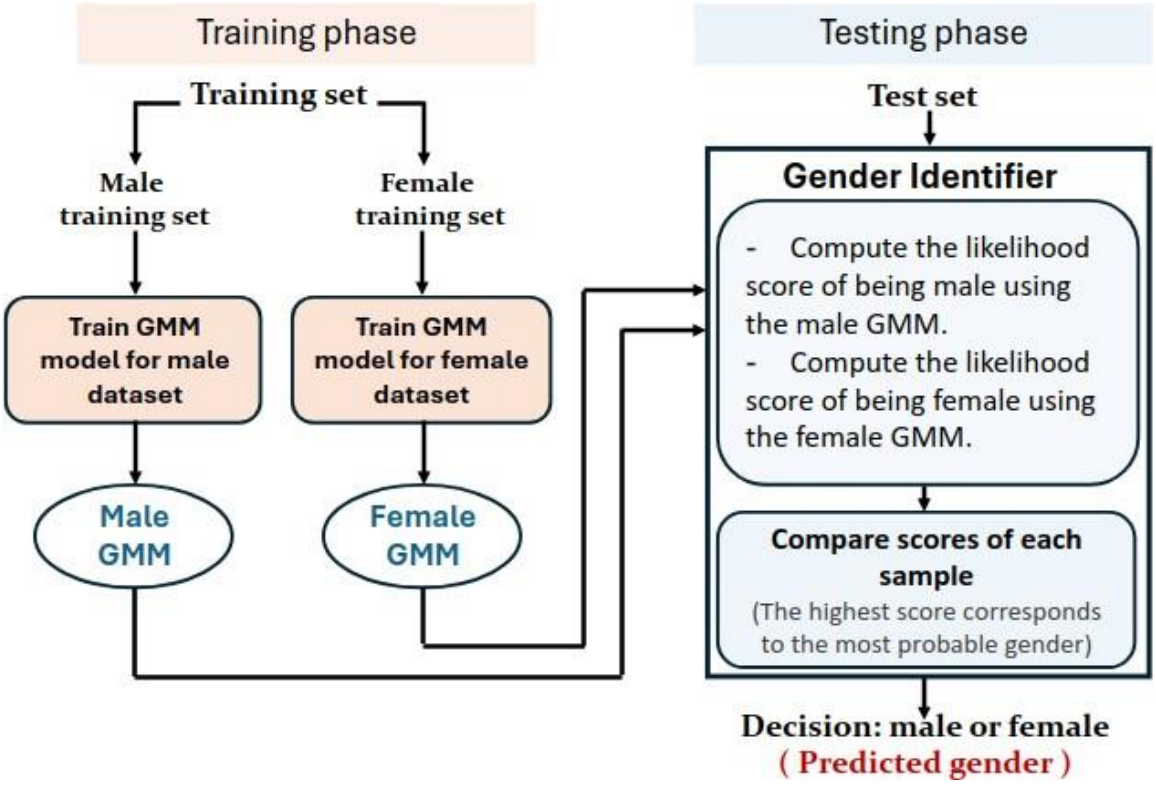
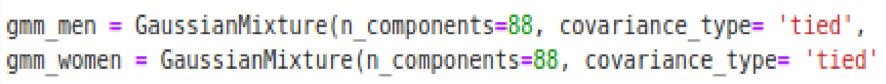

Training phase
Testing phase
Training set
Male training set
Female training set
Train GMM model for male dataset
Train GMM model for female dataset
Male GMM
Female GMM
Test set
Gender Identifier
- Compute the likelihood score of being male using the male GMM.
- Compute the likelihood score of being female using the female GMM.
Compare scores of each sample
(The highest score corresponds to the most probable gender)
Decision: male or female
(Predicted gender)

Figure 3.21 : The training and testing workflow for the GMM-based classification

The training process involved separately training GMMs for male and female datasets. The figure 3.21 illustrates the training and testing process in the GMM-based system.

The trained models were evaluated on a test set to compute the likelihood scores for each sample. The predicted gender was determined by comparing the likelihood scores from both the male and female GMMs, assigning the gender corresponding to the higher likelihood score.

The best results were obtained with the following configuration (see Figure 3.22):



```
gmm_men = GaussianMixture(n_components=88, covariance_type= 'tied',
gmm_women = GaussianMixture(n_components=88, covariance_type= 'tied'
```

Figure 3.22 : The Best Parameters with GMM.

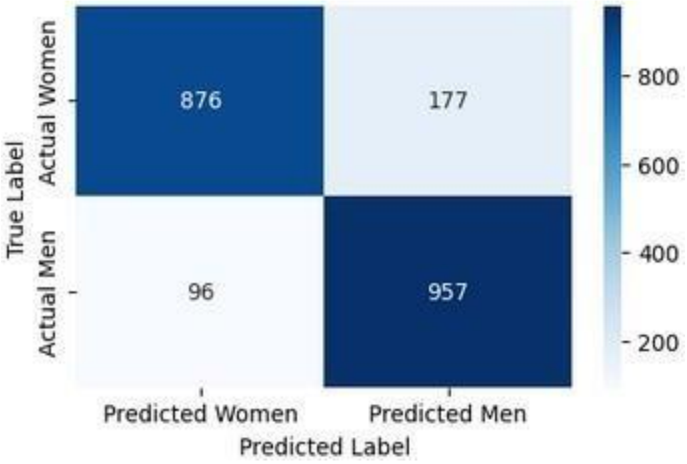


Figure 3.23: The Confusion Matrix of the GMM-based classifier

The confusion matrix in figure 3.23 shows:
957 True Positives, 876 True Negatives, 177 False Positives, and 96 False Negatives.

The table below presents the performance metrics of a The GMM-based classifier, separated by gender and the overall performance.

Metric	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**Inference time** (milliseconds)
Male	91.93	84.91	91.92	88.28	0.36
Female	83.67	91.20	83.66	87.27	
Overall	**87.80**	88.05	87.79	87.77	

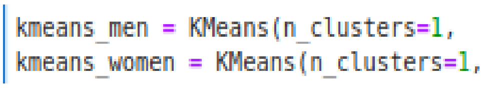
Table 3.9: Performance Evaluation Metrics of the GMM-based classifier

The GMM-based classifier shows strong performance, especially in classifying male voices, with high accuracy and recall. Although the accuracy for female voices is lower, the high precision for this class compensates somewhat for this disparity. The model is thus well- balanced and performs effectively for both genders, with an inference time of approximately 0.36 milliseconds, making it suitable for real-time applications.

B. K-Means :

In this section, we evaluate the performance of a KMeans-based classifier.

The best results were obtained with the following configuration (see Figure 3.24):



```
kmeans_men = KMeans(n_clusters=1,
kmeans_women = KMeans(n_clusters=1,
```

Figure 3.24 : The best parameters with K-Means

The training process involved creating two separate KMeans models, each with a singlecluster, and training them separately on male and female training sets.

The trained models were evaluated on a test set by comparing the distances to the cluster centers of the male and female models. The predicted gender was determined based on which cluster center was closer.

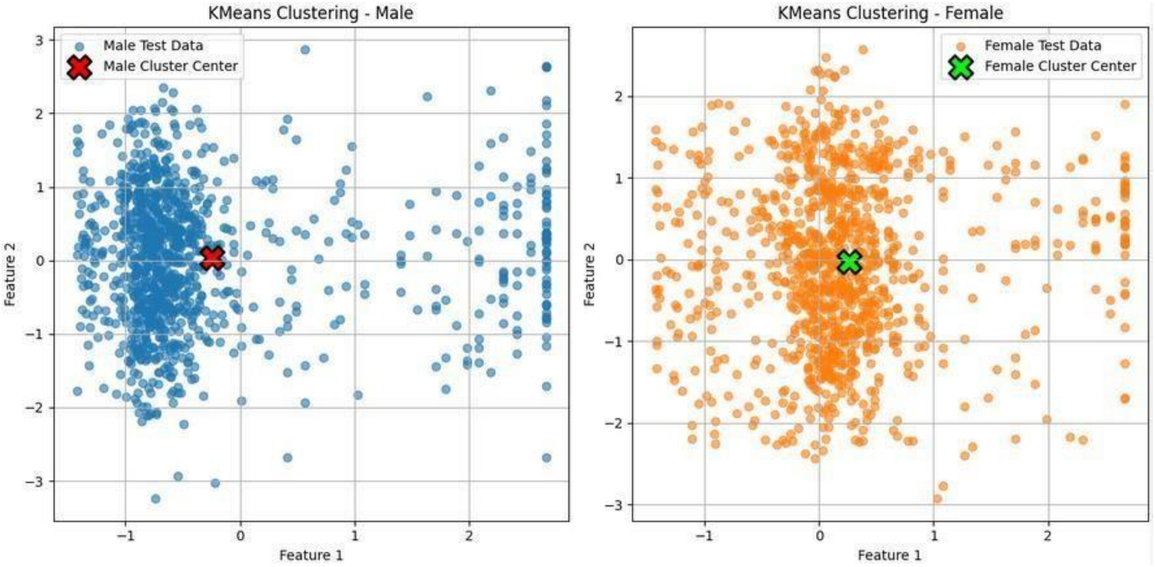


Figure 3.25 : K-Means Clustering of Gender Classification in PCA-Reduced Space

The figure above (Figure 3.25) illustrates a two-dimensional visualization of the KMeans clustering results for both genders. Each point represents a sample from the test set, and the position of the cluster centers indicates the centroid of the respective clusters.

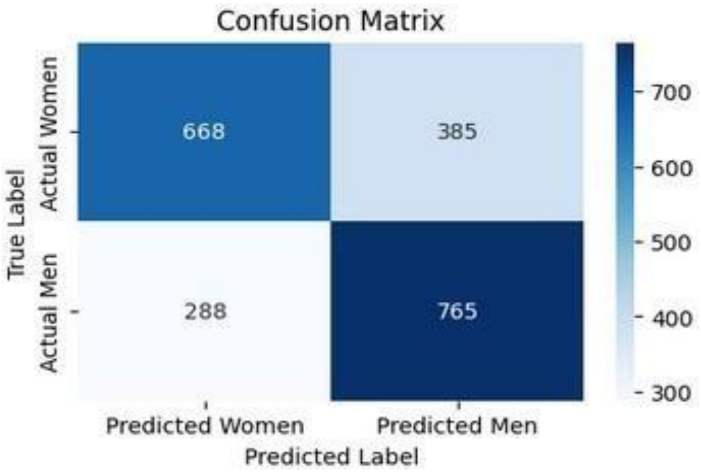


Figure 3.26: The Confusion Matrix of the K-Means based Classifier

The confusion matrix in figure 3.26 shows:
765 True Positives, 668 True Negatives, 385 False Positives, and 288 False Negatives.

The table below presents the performance metrics of a The Kmeans-based classifier, separated by gender and overall performance.

Metric	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**Inference time** (milliseconds)
Male	72.65	66.52	63.44	66.50	0.35
Female	63.44	69.87	72.64	69.45	
Overall	**68.04**	68.19	68.04	67.97	

Table 3.10 : Performance Evaluation Metrics of the K-Means-based classifier.

The k-Means-based classifier shows moderate performance. While the model performs better in accurately identifying male voices, it achieves higher precision and recall for female voices, leading to a better overall classification balance for females. The overall accuracy of 68.04% indicates that the model performs reasonably well but is less effective than other classifiers. Despite these performance limitations, the model's very low inference time of approximately 0.35 milliseconds makes it efficient for real-time applications despite its lower accuracy.

3.4.Accuracy-Based Model Comparison:

In this section, we provide a summary of the accuracies of six machine learning algorithms evaluated previously to identify the most effective model in distinguishing between male and female voices. The obtained results are specific to our dataset and may vary for other datasets.

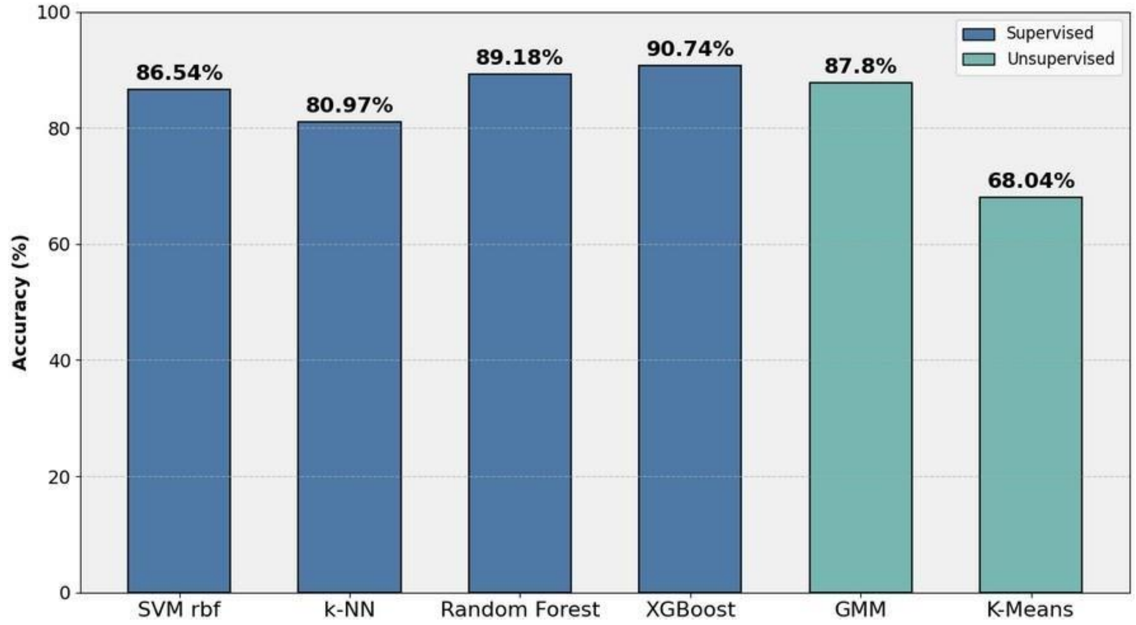


Figure 3.27 : Comparison of Machine Learning Algorithm Accuracies

The accuracies of the various supervised and unsupervised learning models are summarized in Figure 3.27.

The XGBoost model stands out as the optimal choice for classifying gender, achievingan accuracy of 90.74% with a reasonable inference time of 2.61 milliseconds. The Random Forest model also showed strong performance, with an accuracy of 89.18%, though it had a longer inference time of 5.22 milliseconds.

Among the unsupervised learning methods, the Gaussian Mixture Model (GMM) achieved the best results, with an accuracy of 87.80% and a very quick inference time of 0.36 milliseconds.

One of the challenges in classification, which potentially limits accuracy, is the diversity of the audio samples sourced from the VoxForge database. This dataset includes speakers of various languages and accents, along with recordings of varying qualities, while most studies use datasets acquired in controlled environments and acoustic conditions. These factors introduce significant variability in vocal properties and background noise, adding complexity to the classification task.

Despite these challenging conditions, the strong performance of the models, especially XGBoost, indicates their capability to generalize well. This robustness makes these models suitable for real-world applications where audio data is often collected in less-than-ideal environments.

4. Real-Time Gender Classification System

In this section, we present a real-time gender classification application that uses voice recordings to identify a person's gender. The application is designed to record audio, preprocessit to extract relevant features, and predict the gender using a pre-trained XGBoost machine learning model. The graphical user interfaces (GUI), implemented in Python, is built with the PyQt5 library, providing an intuitive and visually appealing experience for users.

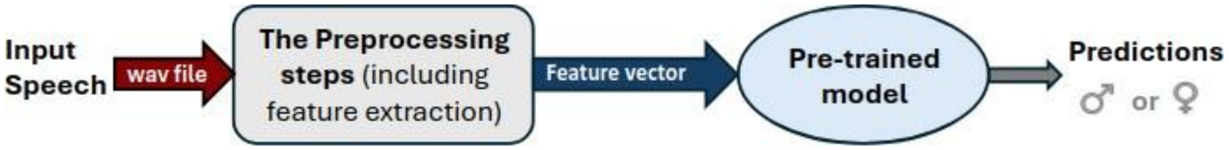


Figure 3.28 : Architecture of the Real-Time Gender Classification system

Figure below illustrates the flow of the application, from recording the input speech to generating the gender prediction.

4.1.Graphical User Interface Overview:

The application window, titled "Real-Time Gender Classification", prominently displays the title "Gender Identification" at the top, indicating the application purpose.

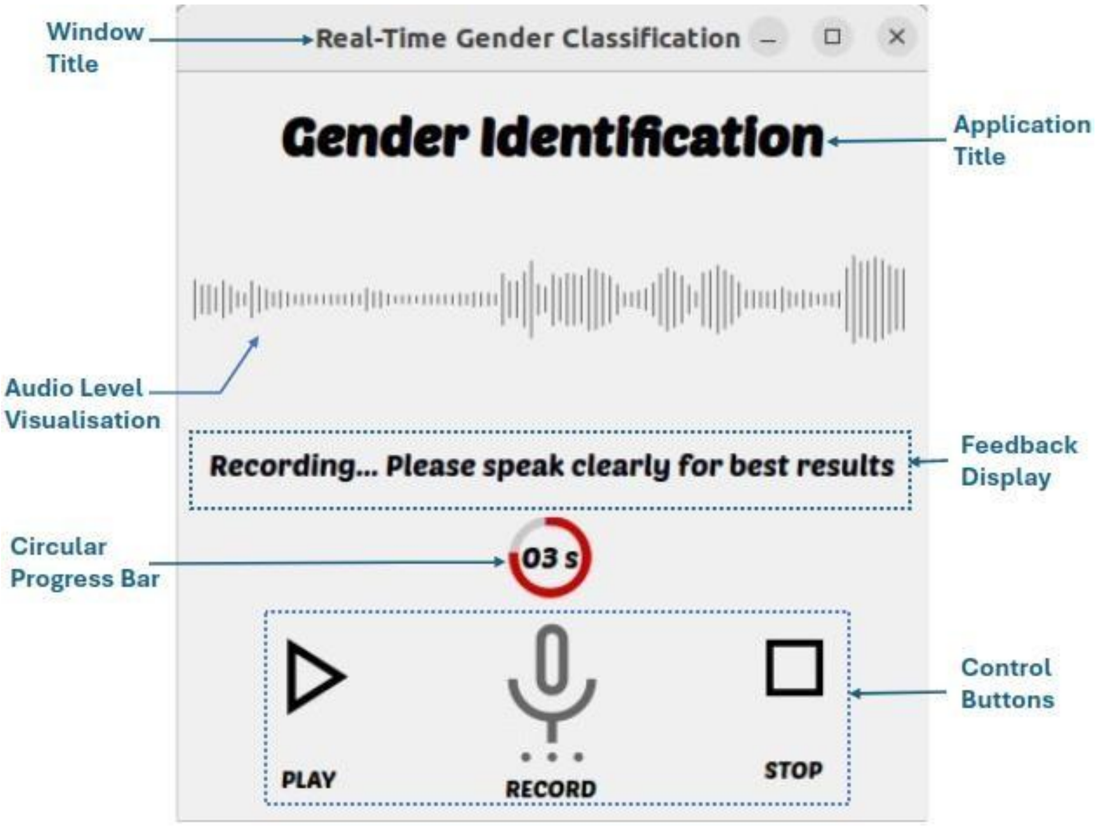


Figure 3.29 : Annotated Interface of the Real-Time Gender classification Application

The interface includes key components to guide users through the process, as shown in the figure. These components include:

- **Audio Level Visualisation**: The audio level bars consist of vertical bars that display the loudness of the recorded audio inreal-time. This helps users ensure their voice is being captured correctly.
- **Circular Progress Bar :** The Circular Progress Bar is a visual timer that runs during the 5-second recording session, indicating the elapsed time through a circular animation. It stops automatically when the recording is complete.
- **Playback Indicator :** A red vertical bar moves across the Audio Level Bar to indicate the current playback position.This feature allows users to visually track the playback progress in real time.
- **Control Buttons :** Includes buttons for recording, playing, and stopping audio, with clear icons and labels forintuitive operation.
- **Feedback Display :** The feedback display guides users through the process with clear messages and updates. Itserves multiple purposes, including:
  - → **Error messages:** Application provides clear error messages for various scenarios, suchas no recording available to play, a disconnected microphone or a recording that is tooquiet . Users are informed of any issues through clear and concise error messages.
  - → **User Guidance:** The feedback display provides specific messages to guide users through the process, ensuring they understand each step and can troubleshoot issues asthey arise.

- **Gender Prediction :**

Once a prediction is made, it shows the predicted gender in a large, easy-to-read font accompanied by a gender icon as shown in figure 3.30 and 3.31.

Predicted Gender: MALE ♂

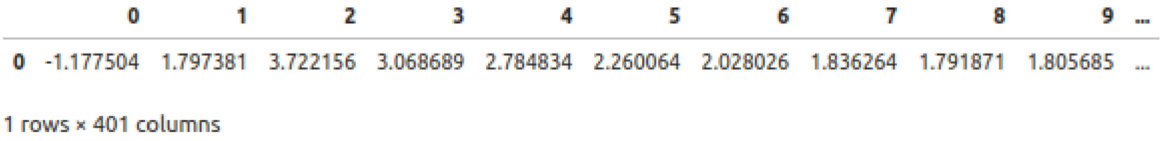
Figure 3.30: Display of predicted Gender Indicating Male.

Predicted Gender: FEMALE ♀

Figure 3.31: Display of predicted Gender Indicating Female

Here's how the application works in practice:

When the user clicks the 'Record' button, the application starts recording audio for a fixed duration of 5 seconds. The circular progress bar begins to fill, and the audio level bar updates in real-time to show the input volume. Once the recording is complete, the application preprocesses the audio, extracts features, and predicts the gender, displaying the result (see figure 3.28 and 3.32). The user can then play back the recording or make a new recording.



	0	1	2	3	4	5	6	7	8	9	...
0	-1.177504	1.797381	3.722156	3.068689	2.784834	2.260064	2.028026	1.836264	1.791871	1.805685	...

1 rows × 401 columns

Figure 3.32 : A segment of the Extracted Features Vector.

5. Conclusion:

This chapter presented a thorough methodology for gender recognition from voice features, including preprocessing, modeling, and real-time application development. During the preprocessing phase, we meticulously prepared the audio data for modeling by extracting crucial vocal features and ensuring data consistency through handling potential missing values, scaling features, and encoding categorical data.

In the modeling phase, we evaluated various machine learning algorithms. XGBoost proved to be the most accurate, with an accuracy of 90.74%, while the Gaussian Mixture Model (GMM) was the best unsupervised method, achieving an accuracy of 87.80%.

Additionally, we introduced a real-time gender classification application we developed using the XGBoost model. This application features a user-friendly GUI, which enables seamless audio recording, preprocessing, and gender prediction, making it practical for real-time use

General Conclusion and Perspectives

The journey from understanding the biological mechanisms of speech production to developing sophisticated machine learning models for gender recognition from voice has been both insightful and challenging. This thesis documents significant advancements in using artificial intelligence to replicate the intricate human ability to recognize gender from voice.

Initially, we studied the basics of human speech production and perception, identifying key acoustic features - MFCCs and pitch- that are crucial in differentiating male and female voices. We then delved into the realm of artificial intelligence, exploring various machine learning algorithms that effectively guided our model selection process.

The core of our research involved developing a structured methodology for gender recognition, which included meticulous preprocessing of audio data, extraction of essential features, and the evaluation of multiple machine learning models. The experiments were conducted using a dataset collected in unconstrained environments (incorporating factors such as background noise, different languages, accents, ages, and the emotional states of the speakers). Among the models tested, XGBoost demonstrated the highest accuracy at 90.74%.

Furthermore, we successfully developed a real-time gender classification application, showcasing the real-world applicability of our research findings.

To conclude, our experiments demonstrated the potential of advanced machine learning methods in effectively addressing the complexities involved in gender classification and the efficiency of the adopted approach, which remains open to further improvement. Moving forward, several promising directions for future research and development could be explored:

- Incorporating a broader range of acoustic features beyond MFCCs and pitch could further enhance the model's accuracy. This can be achieved by using advanced feature engineering techniques, including the extraction of new features from raw audio data, transforming existing characteristics or combining features to create new ones.

- Using deep learning models like CNNs and RNNs, for both feature extraction and classification. Deep learning can automate the extraction of intricate and subtle features directly from raw audio data and perform classification with high accuracy.
- Increasing the diversity and size of the dataset could improve the model's generalizability.
- Finding effective noise reduction techniques that can enhance audio clarity while preserving important features.
- Exploring the deployment of gender recognition models in various fields, such as healthcare, education, and security. The deployement envolves adapting the model to specific contexts and environments, ensuring accurate and efficient performance across diverse applications.

Bibliography:

[1] R. Dias and A. Torkamani, "Artificial intelligence in clinical and genomic diagnostics", Genome Medicine, vol. 11, no. 70, 2019. doi: 10.1186/s13073-019-0689-8.

[2] M. Nag, K. Singh, and K. Nirmal, "Survey Paper on Various Techniques of Voice Gender Recognition,"ResearchGate, May 2023. doi: 10.13140/RG.2.2.19493.06888/1.

[3] I. E. Livieris, E. Pintelas, and P. Pintelas, "Gender recognition by voice using an improved self-labeledalgorithm, "Machine Learning and Knowledge Extraction, vol. 1, no. 1, pp. 492-503, 2019. DOI: 10.3390/make1010030.

[4] A. Bendahmane, "Cours de Traitement Automatique de la parole" Polycopié de l'USTO, Oran, Algérie, 2014.

[5] A. Pahwa and G. Aggarwal, "Speech feature extraction for gender recognition,"Int. J. Image GraphicsSignal Process. vol. 8, pp. 17, 2016. [CrossRef]

[6] J. Ahmad, M. Fiaz, S.-i. Kwon, M. Sodanil, B. Vo, and S. W. Baik, "Gender Identification using MFCC for Telephone Applications – A Comparative Study, " International Journal of Computer Science and Electronics Engineering (IJCSEE), vol. 3, no. 5, pp. XXX-XXX, 2015. ISSN 2320–4028[Online].

[7] C. Ericsdotter and A. M. Ericsson, "Gender differences in vowel duration in read Swedish: Preliminaryresults, " Work. Pap. Lund Univ. Dep. Linguist. Phon., vol. 49, pp. 34–37, 2001.

[8] Vocabulary.com, "Voice," in Vocabulary.com Dictionary. Retrieved April 14, 2024, fromhttps://www.vocabulary.com/dictionary/voice

[9] N. Hammami, "Contribution to the automatic speech recognition of a language and its applications, "Ph.D. dissertation, University of Annaba, Annaba, Algeria, 2015. [Online]. Available: https://biblio.univ-annaba.dz/wp-content/uploads/2015/11/These-Hammami-Nacereddine-1.pdf

[10] J. L. Flanagan, "Speech analysis, synthesis, and perception," Springer, Berlin, 1965.

[11] M. Belinchón, J. M. Igoa, and Á. Rivière, "Psicología del Lenguaje: Investigación y teoría",Madrid: Trotta, 1994.

[12] A. Monaghan, "Phonetics: Processes of Speech Production, "1998. Available from http://www.compapp.dcu.ie/~alex/CA162/PHONETICS/processes.html.

[13] H. J. Gingrich, "English Phonology: An Introduction, " Cambridge: Cambridge University Press,1992

[14] M. J. Sandage, W. Zhang, and J. Ongkasuwan, "Paradoxical Vocal Fold Motion, " in Multidisciplinary Management of Pediatric Voice and Swallowing Disorders, J.McMurray, M.Hoffman, and M. Braden, Eds.available from https://doi.org/10.1007/978-3-030-26191-7_28 [Accessed: February, 16, 2024]

[15] A. Monaghan, "Phonetics: Processes of Speech Production, " 1998. Available from http://www.compapp.dcu.ie/~alex/CA162/PHONETICS/processes.html. [Accessed: March, 21, 2024]

[16] A. G. Adami, "Automatic Speech Recognition: From the Beginning to the Portuguese Language, "Universidade de Caxias do Sul, Centro de Computação e Tecnologia da Informação, Caxias do Sul, RS95070560, Brasil, [Online].Available from: https://www.researchgate.net/publication/266229294_Automatic_Speech_Recognition_From_the_Beginning_to_the_Portuguese_Langu age [Accessed: April, 21, 2024]

[17] Cleveland Clinic, "Hearing Loss: Tips for Making Communication Easier," Cleveland Clinic, 2024. [Online]. Available: https://my.clevelandclinic.org/health/articles/17054-hearing. [Accessed:Jun, 21, 2024]

[18] J. E. Hawkins, "Human ear," Encyclopedia Britannica, 17 Jun. 2024. [Online]. Available: https://www.britannica.com/science/ear. [Accessed: June. 10, 2024].

[19] Y. Aziza, "Modélisation AR et ARMA de la Parole pour une Vérification Robuste du Locuteur dans un Milieu Bruité en Mode Dépendant du Texte," Mémoire de magister, Université Ferhat Abbas– Sétif, Algérie, 2013, [Online], http://dspace.univ-setif.dz:8888/jspui/handle/123456789/1961, [Accessed:Jun, 21, 2024]

[20] A. Hadjer and B. Sabrina, "La Reconnaissance Automatique du Locuteur (RAL) par réseaux de neurones et GMM," Thèse d'ingénieur D'Etat, USTHB, Algérie, 2009.

[21] M. A. Anusuya, "Frontend Analysis of Speech Recognition - a Review," Int. J. Speech Technol.,Springer.

[22] Practical Cryptography PDF | PDF, Scribd. Accessed: Jun. 8, 2024. [Online]. Available from https://www.scribd.com/doc/303111229/Practical-Cryptography-pdf

[23] S. Bhupinder, R. Vanita, and M. Namisha, "Preprocessing in ASR for computer machine interaction with humans: A review," International Journal of Advanced Research in Computer Scienceand Software Engineering, vol. 2, pp. 396-399, 2012.

[24] E. Ramdinmawii and V. K. Mittal, "Gender identification from speech signal by examining the speech production characteristics," in 2016 International Conference on Signal Processing and Communication (ICSC), 2016, pp. 244–249. DOI: 10.1109/ICSPCom.2016.7980584.

[25] Le Mécanisme de Production du Son," Les Paramètres de la Voix, Available from : https://tpe-sur-la-voix.mozello.fr/ii-le-mecanisme-de-production-du-son/2-les-parametres-de-la-voix/ [Accessed: March, 29, 2024]."

[26] R. Boite, Traitement de la parole, Collection Electricité. Presses Polytechniques et UniversitairesRomandes, 2000.

[27] F. Signol, "Estimation de fréquences fondamentales multiples en vue de la séparation des signauxde parole mélangés dans un même canal," Thèse de doctorat, Université Paris-Sud 11, 2009.

[28] Sara Abdelouahed, "Analyse spectro-temporelle du signal vocal en vue du dépistage et du suivi des dysphonies chroniques," Mémoire de fin d'étude pour l'obtention du diplôme de doctorat en Signaux et Images en médecine, Université Abou BekrBelkaid Tlemcen, 2015.

[29] F. Racine, "Modélisation du son," Centre Galois, 2013.

[30] B. Salim and N. Ahmed, "Compression et codage de la parole par la transformation KLT."[Online]http://dspace.univbouira.dz:8080/jspui/bitstream/123456789/9651/1/M%C3%A9moire%20de%20fin%20d%E2%80%99%C3%A9tude%20master.pdf [Accessed: March, 29, 2024]."

[31] J. L. Shen, J. W. Hung, and L. S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," in Fifth International Conference on Spoken Language Processing,1998, vol. 98, pp. 232–235.

[32] N. Madhu, "Note on measures for spectral flatness," Electronics Letters, vol. 45, no. 23, pp. 1195–1196, 2009. DOI: 10.1049/el.2009.1977.

[33] M. A. Uddin, R. K. Pathan, M. S. Hossain, and M. Biswas, "Gender and region detection from human voice using the three-layer feature extraction method with 1D CNN," Journal of Information and Telecommunication, vol. 6, no. 1, pp. 27–42, 2021. [Online]. Available: https://doi.org/10.1080/24751839.2021.1983318

[34] S. Molau et al., "Computing Mel-frequency cepstral coefficients on the power spectrum," in Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE InternationalConference on, vol. 1, pp. 73-76, 2001.

[35] D.A. Reynolds, "Experimental evaluation of features for robust speaker identification," in "Speech and Audio Processing, IEEE Transactions on", vol. 2, no. 4, pp. 639-643, Oct. 1994.

[36] H. Ezzaidi, J. Rouat, and D. O'Shaughnessy "Combining pitch and MFCC for speaker recognition systems," Tech. Rep., ERMETIS, Université du Québec Chicoutimi, Canada.

[Online].Available:https://www.academia.edu/21189930/Combining_pitch_and_MFCC_for_speaker_recognition_systems?email_work_card=view-paper [Accessed: March, 17, 2024].

[37] S. Virkar, A. Kadam, S. Mallick, N. Raut, and S. Tilekar, "Proposed Model of Speech Recognition using MFCC and DNN," Int. J. Eng. Res. Technol., vol. 9, no. 05, pp. 12345-12350,May2020.[Online].Available:https://pdfs.semanticscholar.org/d90a/d9b8787da9da0c9351054e5056e1f372e3a2.pdf [Accessed: March, 20, 2024].

[38] MathWorks, "Speaker Identification Using Pitch and MFCC,"MathWorks AudioToolboxDocumentation.[Online].Available:https://www.mathworks.com/help/audio/ug/speaker- identification-using-pitch-and-mfcc.html [Accessed: March, 30, 2024]

[39] X Shao, B. Milner and S. Cox Integrated Pitch and MFCC Extraction for Speech Reconstructionand Speech Recognition Applications, In Eurospeech'03, Geneva, pp. 1725- 1728, 2003

[40] H. Rouat and H. Ezzaidi, "Pitch and MFCC dependent GMM models for speaker identification systems," in "Canadian Conference on Electrical and Computer Engineering, 2004 (IEEE Cat. No. 04CH37513), "2004.

[41] M. Faundez-Zanuy and E. Monte-Moreno, "State-of-the-art in speaker recognition," Escola UniversitariaPolitècnica de Mataró, TALP Research Center, Avda. Puig i Cadafalch 101-111, 08303 Mataro (Barcelona), Spain. https://arxiv.org/pdf/2202.12705 [Accessed: March, 17, 2024].

[42] IBM."SpeechRecognition."[Online].Available:https://www.ibm.com/topics/speechrecognition#:~:text=Speech%20recognition%2C%20also%20known%20as,speech%20into%20a%20written%20format. [Accessed: June, 17, 2024].

[43] S. Karpagavalli and E. Chandra, "A Review on Automatic Speech Recognition Architecture and Approaches," Int. J. Signal Process. Image Process. Pattern Recognit., vol. 9, no. 4, pp. 393-404, 2016.Available: http://dx.doi.org/10.14257/ijsip.2016.9.4.34. [Accessed: June 19, 2024]

[44] N. Sahota, "Speech Recognition: Applications, Features & Future," *LinkedIn*, Apr. 11, 2023. [Online]. Available: https://www.linkedin.com/pulse/speech-recognition-applications-features-future- neil-sahota-%E8%90%A8%E5%86%A0%E5%86%9B-/ [Accessed: June 19, 2024]

[45] M. Tanveer et al., "Ensemble deep learning in speech signal tasks: A review," Neurocomputing,2023.[Online].Available:https://www.sciencedirect.com/topics/computer- science/gender-recognition. [Accessed: May 19, 2024]

[46] T. Bäckström et al., "Introduction to Speech Processing," 2nd ed. Espoo, Finland: Aalto

University School of Science, 2022. [Online].Available: :https://speechprocessingbook.aalto.fi. [Accessed: March 30, 2024]

[47] .R. V. Pawar, P. P. Kajave, and S. N. Mali, "Speaker Identification using Neural Networks," in IEC (Prague), Aug. 2005, pp. 429-433 [Accessed: April, 1, 2024]

[48] Y. Mustafa, E. Tolga, and A. Nizamettin, "Performance Evaluation of Feature Extraction and Modeling Methods for Speaker Recognition," Annals of Reviews in Research, vol. 4, no. 3, pp. 555- 639, 2018. DOI: 10.19080/ARR.2018.04.555639.

[49] D. Istrate, M. Chenafa, M. Herbin, and V. Vrabie, "Biometric System Based on Voice Recognition Using Multiclassifiers," in Biometrics and Identity Management, First European Workshop, BIOID 2008, Roskilde, Denmark, May 7-9, 2008, revised selected papers.[Online].Available:https://www.researchgate.net/publication/221536350_Biometric_System_Based_on_Voice_Recognition_Using_Multiclassifiers [Accessed: May,22, 2024]

[50] Picovoice. (2023, May 8). Speaker Recognition: Applications and Use Cases. Picovoice. Available: https://picovoice.ai/blog/speaker-recognition-applications/ [Accessed: June 19, 2024]

[51] K. Amar, N. Hammou, and M. Moumene, "Reconnaissance vocale du genre basée sur l'apprentissage profond," Master'sthesis, supervised by M. E. Amine, MCA, Université de Mostaganem, 2022-2023.

[52] Md. S. Ali, Md. S. Islam, and Md. A. Hossain, "Gender recognition system using speech signal," Int. J. Comput. Sci. Eng. Inf. Technol. (IJCSEIT), vol. 2, no. 1, pp. 1-1, Feb. 2012.DOI:10.5121/ijcseit.2012.2101,Availablehttps://www.researchgate.net/publication/270274086_GENDER_RECOGNITION_SYSTEM_USIN G_SPEECH_SIGNAL

[53] I. A. Modupe, T. J. Sefara, "Yorub` a Gender Recognition from Speech using Attention-basedBiLSTM," Tshwane University of Technology, Department of Computer Science, 2019. [Online].Available: https://aclanthology.org/2019.nsurl-1.3.pdf

[54] I. H. Sarker, "AI-driven cybersecurity: an overview, security intelligence modeling and researchdirections,", "SN Comput. Sci"., 2021.

[55] E.H.Alkhammash, M.Hadjouni, and A.M. Elshewey. 2022. "A Hybrid Ensemble Stacking Modelfor Gender Voice Recognition Approach." Electronics 11, no. 11: 1750. https://doi.org/10.3390/electronics11111750.

[56] I. H. Sarker, M. M. Hoque, Md. K. Uddin, and A. Tawfeeq, "Mobile data science and intelligent apps: concepts, AI-based modeling and research directions," "Mob Netw Appl, "pp. 1–19, 2020.

[57] Coursera, "What is Artificial Intelligence?" [Online]. Available:https://www.coursera.org/articles/what-is-artificial-intelligence. [Accessed: April, 18, 2024].

[58] Turabit, "What Are the Objectives of AI?" [Online]. Available: https://www.linkedin.com/pulse/what-objectives-ai-turabit-ai-1f/.[Accessed:April,18, 2024].

[59] Z. Lateef, "Types of artificial intelligence you should know," [Online]. Available:https://www.edureka.co/blog/types-of-artificial-intelligence/. [Accessed: April, 18, 2024]

[60] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," Published online: 22 March 2021, Available from: https://link.springer.com/article/10.1007/s42979-021-00592-x

[61] M. Mohammed, M. Khan, and B. E. Bashier Mohammed, "Machine Learning: Algorithms and Applications", CRC Press, 2016.

[62] S. Mahendra, "What Are Machine Learning Models?" , Artificial Intelligence . [Online]. :https://www.aiplusinfo.com/blog/what-are-machine-learning-models/ Accessed on: may20, 2024.

[63] J. Han, J. Pei, and M. Kamber, "Data Mining: Concepts and Techniques. Amsterdam, Netherlands: Elsevier, 2011.

[64] I. H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity data science: an overview from machine learning perspective," "J. Big Data, "vol. 7, no. 1, pp. 1–29, 2020.

[65] F. Pedregosa et al., "Scikit-learn: machine learning in python," "J. Mach. Learn. Res". vol. 12, pp. 2825–2830, 2011.

[66] .ai, "Performance Metrics in Machine Learning: Complete Guide," [Online].Available: https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide. [Accessed: May, 30, 2024].

[67] Medium, "Evaluation Metrics for Classification,"[Online].Available: https://medium.com/@impythonprogrammer/evaluation-metrics-for-classification-fc770511052d [Accessed: May30, 2024].

[68] Encord, "Classification Metrics: Accuracy, Precision, Recall," [Online].Available from: https://encord.com/blog/classification-metrics-accuracy-precision-recall/. [Accessed: May, 30, 2024].

[69] Geeks for Geeks, "Confusion Matrix in Machine Learning," [Online]. Available: https://www.geeksforgeeks.org/confusion-matrix-machine-learning/. [Accessed:May, 30, 2024].

[70] J. MacQueen , "Some methods for classification and analysis of multivariate observations," in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, vol. 1, pp. 281-297, Oakland, CA, USA

[71] M. Ester et al., "A density-based algorithm for discovering clusters in large spatial databases withnoise," in *KDD*, 1996, pp. 226-231.

[72] I. H. Sarker, A. Colman, M. A. Kabir, and J. Han, "Individualized time-series segmentation for mining mobile phone user behavior," Comput. J. Oxf. Univ. UK, vol. 61, no. 3, pp. 349-368, 2018.

[73] C. Rasmussen, "The infinite Gaussian mixture model," in Advances in Neural Information Processing Systems, vol. 12, pp. 554-560, 1999.

[74] Guru99, "Supervised vs Unsupervised Learning: Key Differences," Guru99. [Online]. Available:https://www.guru99.com/supervised-vs-unsupervised-learning.html [Accessed : April,14,2024]

[75] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: a survey," J. Artif. Intell. Res, "vol. 4, pp. 237–285, 1996.

[76] D. W. Aha, D. Kibler, and M. Albert, "Instance-based learning algorithms," "Mach. Learn.," vol.6, no. 1, pp. 37–66, 1991.

[77] CourskNN,"qkzk.[Online].Available:https://qkzk.xyz/docs/nsi/cours_premiere/algorithmique/k nn/1_cours/. [Accessed: June 12, 2024].

[78] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and M. K. Radha Krishna, "Improvements to Platt's SMO algorithm for SVM classifier design," "Neural Comput"., vol. 13, no. 3, pp. 637–649, 2001.

[79] "Types of Machine Learning - Javatpoint" , www.javatpoint.com. [Online].Available from on :https://www.javatpoint.com/types-of-machine-learning Accessed:May22,2024]

[80] J. R. Quinlan, "C4.5: Programs for Machine Learning," " Mach. Learn.," 1993.

[81] I. H. Sarker, P. Watters, and A. S. M. Kayes, "Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage," " J. Big Data, " vol. 6, no. 1, pp. 1–28, 2019.

[82] Y. Amit and D. Geman, "Shape quantization and recognition with randomized trees," "Neural Comput"., vol. 9, no. 7, pp. 1545–1588, 1997.

[83] L. Breiman, "Random forests," " Mach. Learn., " vol. 45, no. 1, pp. 5–32, 2001.

[84] Z. Faska, L. Khrissi, K. Haddouch, et al., "A robust and consistent stack generalized ensemble- learning framework for image segmentation," J. Eng. Appl. Sci., vol. 70, p. 74, 2023. doi: 10.1186/s44147-023-00226-4.

[85] J. MacQueen et al., "Some methods for classification and analysis of multivariate

observations," in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, vol.1, pp. 281-297, Oakland, CA, USA

[86] G. Chen, B. Hou, and T. Lei, "A new Monte Carlo sampling method based on Gaussian Mixture Model for imbalanced data classification," Department of Mathematics, Dalian Maritime University, Dalian 116026, China, Published: 18 September 2023.

[87] I. H. Sarker, Y. B. Abushark, F. Alsolami, and A. Khan, "Intrudtree: a machine learning based cyber security intrusion detection model," "Symmetry," vol. 12, no. 5, p. 754, 2020.

[88] I. H. Sarker, Y. B. Abushark, and A. Khan, "Contextpca: predicting context-aware Smartphone apps usage based on machine learning techniques," "Symmetry", vol. 12, no. 4, p. 499, 2020.

[89] H. Liu and H. Motoda, " Feature Extraction, Construction and Selection: A Data Mining Perspective," vol. 453. Springer Science & Business Media, 1998

[90] I. H. Sarker, H. Alqahtani, F. Alsolami, A. Khan, Y. B. Abushark, and M. K. Siddiqui, "Context pre-modeling: an empirical analysis for classification based user-centric context-aware predictive modeling," "J. Big Data," vol. 7, no. 1, pp. 1–23, 2020

[91] X. Hao, G. Zhang, and S. Ma, " Deep learning, " International Journal of Semantic Computing, vol. 10, no. 03, pp. 417–439, 2016.

[92] Z. Hong, "Speaker Gender Recognition System," M.S. thesis, Dept. of Communications Engineering, Univ. of Oulu, Master’s Degree Programme in Wireless Communications Engineering,Oulu,Finland,2017.[Online].Available:https://oulurepo.oulu.fi/bitstream/handle/10024/9711/nbnfioulu-201706082645.pdf?sequence=1&isAllowed=y