الجمهورية الجزائرية الديمقراطية الشعبية
The People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Rese
جامعة محمد بوضياف بالمسيلة
University Mohamed Boudiaf of M'sila

كلية الرياضيات والإعلام الآلي
Faculty of Mathematics and Informatics

قسم الإعلام الآلي
Department of Computer Science

Domain: Mathematics and Computer Science

Thesis Presented to Fulfill the Partial Requirement
for **Master's Degree** in Computer Science

Specialty: Information System and Software Engineering ISSE

Prepared By: Rania Agoune

Supervised By:

Said Gadri

ENTITLED

Speech Emotion Recognition Using Deep Learning Models

Jury Members

Mustapha Bourahla	President
Said Gadri	Supervisor
Said Hamani	Examiner

Academic Year 2024/2025

اهداء

﴿ يَرْفَعِ اللَّهُ الَّذِينَ آمَنُوا مِنْكُمْ وَالَّذِينَ أُوتُوا الْعِلْمَ دَرَجَاتٍ ﴾

الحمد لله،

الحمد لله حُبًا وشُكرًا وامتنانًا على البدء والختام**...**

ها **أنا** اليوم أُتوّج لحظاتي الأخيرة في ذلك الطريق، الذي كان يحمل في باطنه الكثير من العثرات، ورغم كل ما واجهته، ظلّت قدمي تخطو بثبات، بكل صبر، وطموح، وعزيمة، وتفاؤل، وحُسن ظنٍ بالله.

أُهدي بكل حب **تخرجي** إلى **نفسي العظيمة** القوية التي تحملت المشاق وواجهت التحديات بكل ثبات**...**

إلى ذلك الذي لا ينفصل اسمي عن اسمه،

إلى الرجل العظيم الذي كان لي عمودي الفقري، الذي ساندني بكل حب، في ضعفي، وفي لحظاتي المتعثرة، أخرج ما في داخلي من قوة، وآمن بي، وشجعني للوصول إلى طموحاتي، إلى من انتظر هذه اللحظة ليفتخر بي، إلى قوتي، مسندي، وضوئي الذي أنار حياتي... **والدي العزيز**، طيّب الله ثراه**...**

وإلى التي تعجز الكلمات عن وصفها، إلى من كانت النور في عتمتي، إلى من كان دعاؤها سرّ نجاحي، **أمي الغالية**، متّعكِ الله بالصحة والعافية**...**

إلى ضِلعي الثابت، وأمان أيامي، إلى من شددتُ عضدي بهم، فكان لهم بالغ الأثر في تجاوز العقبات والصعاب، **إلى إخوتي وأخواتي الأعزاء**،

السند الدائم، والدعاء الصادق، والدعم الذي لا ينقطع، أدامكم الله لي نورًا وذخرًا في حياتي...

إلى كل من كان لي عونًا وسندًا في هذا الطريق، إلى **صديقتيّ العزيزتين**، اللتين كانتا لي رفقةً وسندًا في الدرب**...**

والى صديقتي ورفيقة دربي **خليدة** حماك الله ورعاك**...**

إلى **أساتذتي الكرام**، ممن لم يترددوا في دعمي ومدّ يد العون**...**

أُهديكم جميعًا هذا الإنجاز، وثمرة جهدي، التي لطالما حلمت بها، وها أنا اليوم أُكمِلها وأتمّها، بفضل الله سبحانه وتعالى.

شكر و عرفان

بسم الله نحمد الله ونشكره أن لولا توفيقه لما نحن هنا ونساله أن يكون خاصا لوجهه

الكريم

وان ينفعنا وينفع الناس جميعا

ومصداقا لقوله تعالى: ﴿ نِّعْمَةً مِّنْ عِندِنَا ۚ كَذَٰلِكَ نَجْزِي مَن شَكَرَ ﴾ القمر٣٥

أتقدم بخالص الشكر والتقدير الى الله تعالى الذي وفقني لإتمام هذا العمل ،ثم إلى أستاذي الفاضل**"سعيد قادري"** ، على توجيهاته القيمة ودعمه المستمر طوال فترة إعداد المشروع ، نسأل الله ان يجزيه عن فضله احسن الجزاء.

والى الأساتذة الكرام **أعضاء اللجنة** المناقشة على موافقتهم لمناقشة هذا العمل

كل الشكر والعرفان **لطاقم الإدارة** وكل **الأساتذة الكرام**

كل الشكر لزميلاتي اللتين أعانتاني في هذا العمل **"هلتالي ندى الريحان "** و **"قاضي دعاء حنان"**

وبالتوفيق للجميع.

Table of Contents

List of Figures

List of Tables

General Introduction

Speech is the primary medium of communication for humankind, and articulating language is perhaps one of the most advanced and singular human traits. An occurrence of nature that always set an allure before the eyes of scientists and psychologists now presents its challenges and opportunities in the multimedia and Artificial Intelligence (AI) domain. One such fascinating strand of research emerging from it is Speech Emotion Recognition (SER), which tries to answer the question concerning whether or not machines can recognize and understand human emotions on the basis of voice signals alone.

Ekman and other psychologists went on to develop an elementary understanding of human emotions, setting out six: happiness, sadness, disgust, fear, surprise, and anger, which apply universally across cultures. These emotions, among others, are embedded in how people express and communicate. Parametric features such as pitch and tone, along with intensity, rhythm, and spectral content, convey emotional cues revealing the inner state of the speaker. To illustrate, anger is exhibited by a pitch increase and quick utterance, whereas anger sees slower utterance that has low energy.

One challenging task to run. Emotional expressions are strongly context-dependent and change with age, gender, culture, and individual speaking style. The manner of appearance and prosody of emotions in spontaneous speech may be substantially different from emotion embodiment under acted or scripted speech recordings. These make the development of a reliable SER system a very multilayered problem and worthwhile research challenge.

The central problem addressed in this thesis concerns the question of how one may reliably and accurately recognize human emotions from speech, in spite of the myriad variations in emotional vocal expression. The problem would thus require one to merge the theoretical study of human emotion with contemporary signal processing and deep learning.

To answer this research question, the present work is organized as follows:

- **Chapter 1: Basic Concepts and Techniques of Speech Emotion Recognition**
  This chapter presents the theoretical background of human emotion, its role in communication, and the different techniques used to detect emotions in speech, with a focus on acoustic features and psychological foundations.
- **Chapter 2: Foundations of Artificial Intelligence and Deep Learning**

  This chapter explores the principles of artificial intelligence, including machine learning and deep learning. It emphasizes the use of convolutional neural networks (CNNs) and introduces key datasets used in SER systems.
- **Chapter 3: System Implementation and Technologies Used**

  This part provides a detailed technical description of the developed system, including programming tools, data preprocessing, feature extraction, the architecture of the CNN model, and performance evaluation metrics.
- **Chapter 4: SER System Implementation & Some User Interfaces and Interface**
  This chapter explains the design and development of a web interface that integrates the trained SER model, allowing real-time emotion prediction from audio files or recorded speech.

The thesis concludes with a general summary of findings, discussions on the limitations of the current system, and suggestions for future improvements and research directions in the field of speech emotion recognition.

Chapter 1
Basic Concepts and Techniques of Speech Emotion Recognition

1 Introduction

Speech Emotion Recognition (SER) is an advanced field within Natural Language Processing (NLP) and Artificial Intelligence (AI). This domain intersects with Sentiment Analysis but differs in that it relies on analyzing vocal signals rather than written texts. This chapter aims to provide a comprehensive overview of speech emotion recognition, including its connection to sentiment analysis, the role of emotions in human interactions, techniques for detecting emotions in speech, as well as the applications of this field across various sectors such as customer service, mental health, entertainment, and education. The chapter also addresses the challenges facing speech emotion recognition systems and proposed solutions to overcome them, along with future trends that will contribute to the development of these systems.

2 Introduction to Sentiment Analysis

Sentiment Analysis is a natural language processing technique that interprets and classifies emotions expressed in text or speech. It employs various approaches, including lexicon-based, Machine Learning, and hybrid methods. we want to use This analysis in speech emotion recognition

Speech Emotion Recognition is related to Sentiment Analysis, which is a subtopic of Natural Language Processing (NLP) that focuses on interpreting opinions from a piece of text. Speech Emotion Recognition differs as it infers the speech signal from audio data; instead of written text like Sentiment Analysis. Together, Speech Emotion Recognition and Sentiment Analysis can capture semantic and vocal emotion [1]

3 The Role of Emotions in Human Interaction

A magical mixture of learned behaviors and genes, we are told, minds with thoughts and hearts bursting with emotions add spice and color to the human condition.

Emotions are at the core of human interactions, dictating how we communicate, why we build relationships, and how we make decisions. These are critical signals that assist individuals in accurately interpreting and reacting to social settings. Be it via facial expressions, tone of voice, or body language, emotions help individuals connect and engage with one another.

3.1 Human Interaction Theories of Emotion

There are several theories that explain how emotions affect human interaction. Charles Darwin said that emotions evolved as survival mechanisms to help humans react quickly to good and bad things in their environment. Building on this idea, Richard Lazarus' Cognitive Appraisal Theory highlights that people's emotions are oftentimes determined by how they perceive the events occurring rather than the events themselves. A second important perspective is the Social Constructionist view, which provides that emotions are not purely biologically determined but also occur in cultural contexts. From this perspective, emotional experience is profoundly affected by the norms, values, and expectations that prevail in a given community [2].

3.2 Emotional Communication & Expression

Communicating emotions through both verbals and nonverbals. This is a matter of tone, word choice, and emotional intelligence in verbal communication; how the message is received in the bigger picture. Emotions are not only expressed in words and non-verbal cues like body language, facial expressions, gestures, posture, and eye contact. Research by Paul Ekman has shown a set of universal facial expressions that correspond to basic emotions such as happiness, anger, sadness and fear. Also, non-verbal expression, which was previously limited to in-person communication, has a digital counterpart today. Most of the time our emotions are inadvertently perpetrated through emojis, punctuation and the type which we write. Such nuances can carry emotional weight — a gesture to maintain a connection, a nuance to tell the camera how you feel, color, tone, body language — and even the absence of — all of these things give flow to emotion, even without the person being in the same physical environment as you [3].

3.3 Emotions and Decision-Making

Decision-making is greatly affected by emotions and how people use them to make decisions aligned with their values and past experiences. Strong emotions can guide good decisions that are well-founded, but depending on the context and the person's ability to regulate their emotions, they can also lead to rash or reckless decisions [4].

In cases of conflict resolution and negotiation, emotions do matter even more, as they determine the way an individual reacts and interacts with others. Being able to control emotions and develop empathy is vital to achieving positive results. Moreover, emotional biases, including the popular affect heuristic, reveal that feelings can implicitly affect evaluations and choices in personal and professional contexts, emphasizing the multifaceted nature of emotion and reasoning.

3.4 The Impact of Suppressing vs. Expressing Emotions

The importance of emotion regulation in mental health and social relationships That is how much the repression of emotion - not to express anger, not to express sadness - only leads to higher stress levels, anxiety and lasting damage to interpersonal relationships. Indirectness indicates a failure of communication, while healthy, open, constructive emotions can actually improve communication and emotional health. Further, culture has a large impact on how we express and interpret emotion; in some cultures, emotional openness is encouraged while restraint and control is affirmed in others. In professional settings and workplaces especially, emotional intelligence in leaders leads to a connection with team members, which then leads to motivation and strong teams. Moreover, the emotions like anger, hope and empathy are substantial motivateors of societal movements, propelling mass action and social transformation. Now, this is a common symptom of depression, but I digress because the big picture here is that the ability to effectively manage and express emotions is critical not only to the health of the individual, but also to forming cohesive, empathetic communities [5].

4 Techniques for Emotion Recognition

Emotion recognition involves identifying and interpreting human emotions through various methods. Some key techniques include [6]:

4.1 Facial Expression Analysis

Uses computer vision and machine learning to detect emotions based on facial features. One common method is Ekman’s Facial Action Coding System (FACS), which identifies subtle muscle movements associated with emotions like happiness, anger, sadness, and fear. This technique is widely used in security, gaming, and human-computer interaction.

4.2 Speech Emotion Recognition

Analyzes vocal characteristics such as:

- ***Pitch:*** Higher pitch may indicate stress or excitement.
- ***Tone:*** Helps differentiate emotions like anger or happiness.
- ***Speech Patterns:*** Speed and pauses can suggest sadness or nervousness. It is applied in AI assistants, customer service, and emotion-aware communication systems.

4.3 Physiological Signals

Measures biological responses to detect emotions, including:

- ***Heart Rate:*** Increases with stress or fear, decreases with relaxation.
- ***Skin Conductance (GSR):*** Changes in response to excitement or anxiety.
- ***Brain Activity (EEG):*** Detects emotional states by analyzing neural patterns. These techniques are used in healthcare, wearable devices, and neuroscience research.

5 Application of SER

Speech Emotion Recognition is not just a theoretical concept. It has real-world applications that can enhance various sectors:

5.1 Customer Service

In customer service, SER can revolutionize how businesses interact with clients. By recognizing emotions, companies can:

- Respond more appropriately to customer needs.
- Identify which representatives are performing well.
- Detect potential issues before they escalate.

5.2 Mental Health

In the realm of mental health, SER has the potential to be a game-changer. It can assist therapists

by:

- Monitoring the emotional wellbeing of patients.
- Identifying changes in emotional states during sessions.
- Providing real-time feedback for better therapeutic outcomes.

5.3 Entertainment and Gaming

Have you ever played a video game that adapted to your mood? SER can make gaming experiences more immersive by:

- Adapting storyline based on player emotions.
- Enhancing character interactions in response to real-time emotional feedback.

5.4 Education

In educational settings, SER can facilitate better learning experiences by:

- Identifying students who may be struggling emotionally.
- Customizing learning materials based on a student's mood.
- Moreover, SER technology can help educators understand their students better. A class full of bored students can be transformed into an engaging environment if the instructor can read the emotional climate of the room.

6 Challenges and Solutions

Several challenges are inherent in speech emotion recognition systems [7], such as:

- ***Variability in Speech Data:*** The diversity in emotional expression across speakers, genders, and accents can lead to challenges in model generalization. To mitigate this, we will experiment with data augmentation techniques such as adding noise, pitch shifting, and time-stretching.
- ***Noise Robustness:*** To handle real-world scenarios with background noise, we will incorporate noise robust feature extraction techniques and test the models in noisy

environments.

- ***Emotion Imbalance:*** Some emotions may be underrepresented in the dataset. To address this, we will consider class balancing techniques such as oversampling underrepresented classes or using loss functions that penalize misclassifications of minority classes more heavily.

7 Emerging Trends Shaping the Future of SER

The field of speech emotion recognition (SER) is evolving rapidly with advancements in deep learning, artificial intelligence, and multimodal emotion recognition. Some of the recent trends include [8]:

- ***Deep Learning Models:*** The use of transformer-based architectures and convolutional neural networks (CNNs) for more accurate emotion detection.
- ***Multimodal Integration:*** Combining facial expressions, voice, and physiological signals to enhance recognition accuracy.
- ***Real-Time Applications:*** Implementation of SER in customer service, healthcare, and virtual assistants to improve user interaction.

8 The Future of SER

The really advanced technologies in AI and machine learning will eventually lead to a better-paced development of SER technologies. Future SER systems are likely to be even more accurate and adaptable in context, thereby recognizing human emotions from a greater variety of speech expressions and linguistic subtleties. Such advancements in SER technologies bode well for richer and more empathetic interactions between humankind and its creations. However, the related ethics of privacy and emotional ambience will still have to be scrutinized. In fact, the development and deployment of SER must follow paths that best safeguard human values and emotional well-being to stimulate positive outcomes [9].

9 Key Processing Stages in SER System

Speech Emotion Recognition (SER) systems involve a structured series of stages aimed at converting raw speech signals into meaningful emotion classifications. This process typically begins with the collection of emotional speech data, followed by rigorous pre-processing to

enhance data quality. Subsequently, relevant acoustic features associated with emotional expression are extracted. Finally, machine learning or deep learning algorithms are employed to classify these features into predefined emotional categories. The following sections detail each of these essential stages.

9.1 Data Collection

Data collection for SER consists of talking about and recording vocal expressions so that they reflect certain emotional states. This would warrant the collection of samples either in a controlled environment or in spontaneous emotional speech under naturalistic conditions. There should be great diversity and quality of collected audio data in order to construct a robust, generalizable emotion recognition model.

9.2 Pre-processing

Cleaning of the raw audio signal in any SER is maintained by pre-processing. This entails the removal of noise, normalization of volume and pitch variations, and segmentation of relevant speech frames. The input to the model is clean and emotion-rich and is subjected to silence removal and voice activity detection.

9.3 Feature Extraction

Feature extraction deals with the identification of important acoustic features associated with the emotional state of the speaker. Commonly extracted features include Mel-frequency cepstral coefficients (MFCCs), pitch, energy, formants, and spectral properties. These features capture prosodic and spectral information of speech, crucial for discriminating between different emotions.

9.4 Classification

The classification stage is where the extraction features are mapped into a specific emotional category. Mostly, algorithms for machine learning and deep learning, such as Support Vector Machines (SVM), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN), are trained for this purpose. Emotions are generally classified according to basic models by Ekman: anger, disgust, fear, happiness, sadness, and surprise,

or according to dimensional models like arousal-valence frameworks.

10 Conclusion

It is evident from this chapter that speech emotion recognition is a promising field with great potential to enhance human-machine interaction. By employing advanced techniques such as deep learning and artificial intelligence, emotion recognition systems can offer innovative solutions in various areas. Despite numerous challenges, such as diversity in voice data and difficulties in adapting to noise, continuous research and innovation in this field will improve the ability of these systems to accurately and effectively understand human emotions. As technology progresses, these systems will play a key role in enhancing human interactions, making them more empathetic and responsive to individual needs in the future.

Chapter 2

Foundations of AI and Deep Learning for Emotion Recognition

1 Introduction

Artificial Intelligence (AI) has witnessed significant advancements in recent years, becoming a core component in many modern applications, particularly in machine learning (ML) and deep learning (DL). These technologies have revolutionized fields such as computer vision, natural language processing, and speech recognition. This chapter focuses on Convolutional Neural Networks (CNN), one of the most effective and widely used deep learning techniques, which has shown outstanding performance in tasks like speech emotion recognition. Additionally, the chapter delves into the datasets commonly used in this domain, which are critical for training and evaluating AI models. The combination of CNNs and rich, diverse datasets is key to the development of robust and accurate systems in speech emotion recognition.

2 Artificial Intelligence

AI alludes to the capacity of a machine to duplicate human-like behaviors, such as thinking, arranging, and inventiveness.

AI permits specialized frameworks to see their environment, oversee these discernments, unravel issues, and take activities to realize a particular objective. The computer gets information (pre-prepared or collected by means of its sensors—a camera, for illustration), analyzes it, and responds.

Frameworks prepared with AI are able to adjust their behavior (more or less) by analyzing the impacts of their past activities, working independently [10].

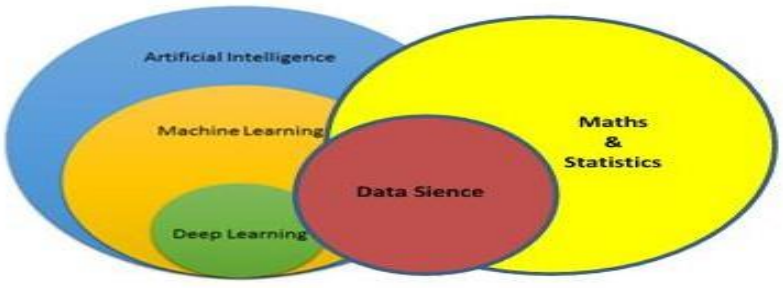


***Figure 1**:Relationship between AI and Data Science*

3 Machine Learning

Machine learning could be a subfield of manufactured insights (AI) that employments calculations prepared on information sets to form self-learning models able of anticipating results and classifying data

Without human mediation. Machine learning is utilized nowadays for a wide run of commercial purposes, counting recommending items to customers based on their past buys, foreseeing stock showcase changes, and interpreting content from one dialect to another [11].

4 Types of machine learning

Machine learning uses large volumes of data to learn, make predictions, find patterns, or classify information. While there are several types of machine learning, three of the most common kinds include supervised learning, unsupervised learning, and reinforcement learning. Here's what you need to know about each one.

- ***Supervised learning:*** Supervised learning is often used to create machine learning models used for prediction and classification purposes.
- ***Unsupervised learning:*** Unsupervised machine learning is often used by researchers and data scientists to identify patterns within large, unlabeled data sets quickly and efficiently.
- ***Reinforcement learning:*** Reinforcement learning is often used to create algorithms that must effectively make sequences of decisions or actions to achieve their aims, such as playing a game or summarizing an entire text [11].

5 Popular Algorithms in Machine earning

In machine and deep learning, several algorithms are implemented to classify and analyze data towards arriving at a solution for various problems. It is these techniques that enable machines to learn from the data and automatically enhance their performance over time, eliminating the need for constant human intervention.

Among the most typical algorithms in this domain include support vector machines, K-nearest

neighbors, and convolutional neural networks. These algorithms' innate characteristics often lead them to perform optimally in various expected environments. Therefore, this section of the report attempts a rudimentary discussion of these algorithms, how each works, and how it can be applicable in different machine learning contexts [13].

5.1 K-Nearest Algorithm (KNN)

The K-Nearest Neighbors algorithm classifies an object by looking at its closest neighbors. One application would be in streaming services, predicting which users are likely to cancel their subscription based on the ages and behaviors of other users. If most of the K nearest neighbors in age have canceled their subscriptions, then KNN will predict that this new user will also cancel. It is typically classified under lazy learning algorithms since it does not learn until classification is thrust upon it.

5.1.1 Steps of the KNN algorithm:

The K-Nearest Neighbors (KNN) algorithm predicts the outcome for a new data point by identifying its K closest neighbors in the training set and using their labels or values to make an informed prediction based on similarity.

- ***Step 1:*** Selecting the optimal value of K, K represents the number of nearest neighbors to consider during prediction.
- ***Step 2:*** Calculating distance, Euclidean distance is used to measure the similarity between the target points and the training data.
- ***Step 3:*** Finding nearest neighbors, the k points closest to the target point are found based on the calculated distance.
- ***Step 4:*** Voting for classification or taking average for regression, In the case of classification, the most common classifications among the nearest neighbors are considered to select the class. In the case of regression, the average values from the nearest neighbors are taken to determine the predicted value of the new point [14].

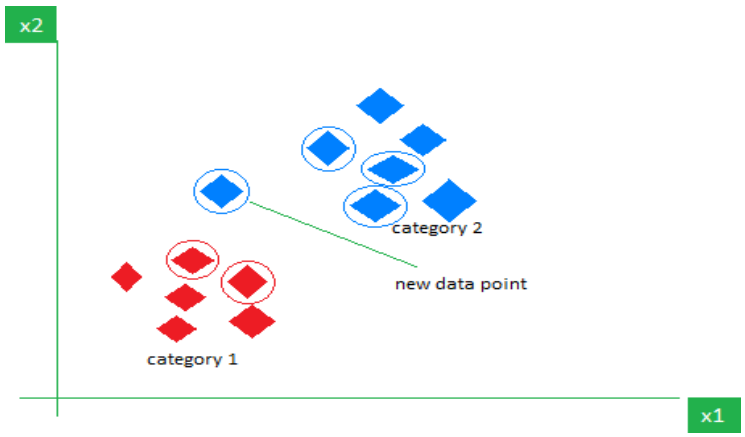


***Figure 2**: KNN Algorithm working visualization*

5.2 Support Vector Machine (SVM) Algorithm

The supervised machine learning method known as Support Vector Machine (SVM) is employed in regression and classification applications. SVM excels at classification assignments even though it can manage regression issues.

The goal of SVM is to divide data points into distinct classes by identifying the ideal hyperplane in an N-dimensional space. The technique seeks to increase the distance between the nearest points of different groups.

5.2.1 How does SVM work?

The key idea behind the SVM algorithm is to find the hyperplane that best separates two classes by maximizing the margin between them. This margin is the distance from the hyperplane to the nearest data points (support vectors) on each side.

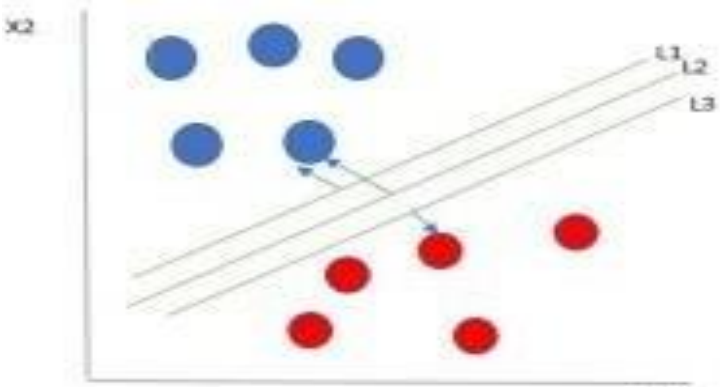


***Figure 3** Multiple hyperplanes separate the data from two classes*

The best hyperplane, also known as the “hard margin,” is the one that maximizes the distance between the hyperplane and the nearest data points from both classes. This ensures a clear separation between the classes. So, from the above figure, we choose L2 as hard margin.

5.2.2 How does SVM classify the data?

It’s simple! The blue ball in the boundary of red ones is an outlier of blue balls. The SVM algorithm has the characteristics to ignore the outlier and finds the best hyperplane that maximizes the margin. SVM is robust to outliers.

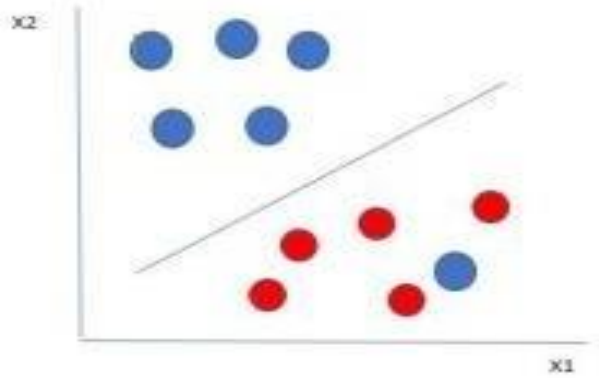


***Figure 4**: Hyperplane which is the most optimized one*

5.2.3 Types of Support Vector Machine Algorithm

Support Vector Machine algorithms can be classified based on how they separate data. This section briefly explains the two main types:

- ***Linear SVM:*** For linearly separable data, a line (in 2D) or a hyperplane (in higher dimensions) is used to divide the data points into distinct classes. The decision boundary maximizes the margin between the classes.
- ***Nonlinear SVM:*** Used when the data cannot be divided by a straight line. The data is mapped into a higher-dimensional space using a kernel function, where a linear decision boundary can then be applied to separate the classes. [14]

6 Deep learning

Deep learning is a branch of machine learning that is made up of a neural network with three or more layers:

- ***Input layer:*** Data enters through the input layer.
- ***Hidden layers:*** Hidden layers process and transport data to other layers.
- ***Output layer:*** The final result or prediction is made in the output layer.

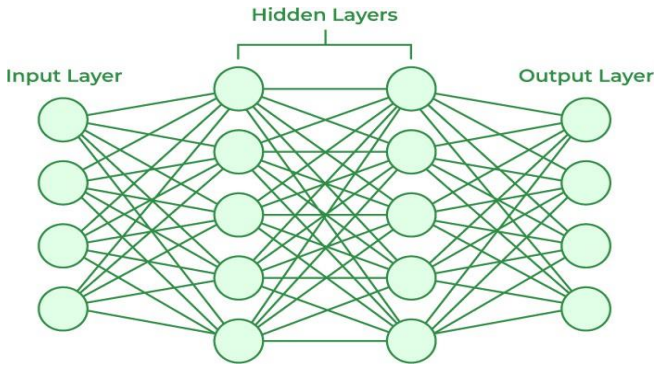


Figure 5: *Neural Networks Architecture*

Neural systems endeavor to demonstrate human learning by processing and analyzing enormous sums of data, moreover known as preparing information. They perform a given assignment with that information more than once, progressing in exactness each time. It's comparative to the way we ponder and hone to progress aptitudes [12].

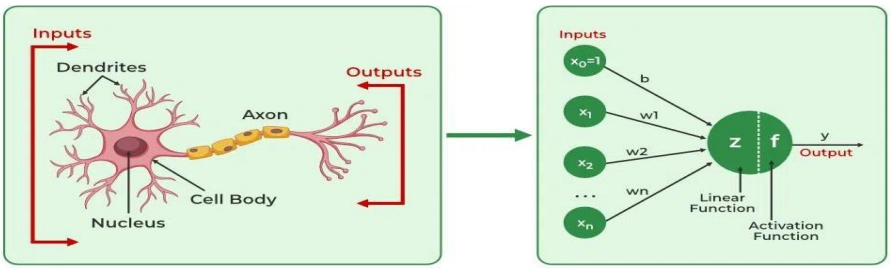


Figure 6: *From Biological Neurons to Artificial Neuron*

7 Popular Model in Deep learning

Deep learning uses more advanced versions of neural networks in trying to identify patterns from large datasets. This section champions a few of the popular algorithms and their major

applications across various domains.

7.1 Recurrent Neural Networks (RNNs)

Recurrent Neural Networks work a bit different from regular neural networks. In neural network the information flows in one direction from input to output. However in RNN information is fed back into the system after each step. Think of it like reading a sentence, when you're trying to predict the next word you don't just look at the current word but also need to remember the words that came before to make accurate guess.

RNNs allow the network to "remember" past information by feeding the output from one step into next step. This helps the network understand the context of what has already happened and make better predictions based on that. For example when predicting the next word in a sentence the RNN uses the previous words to help decide what word is most likely to come next.

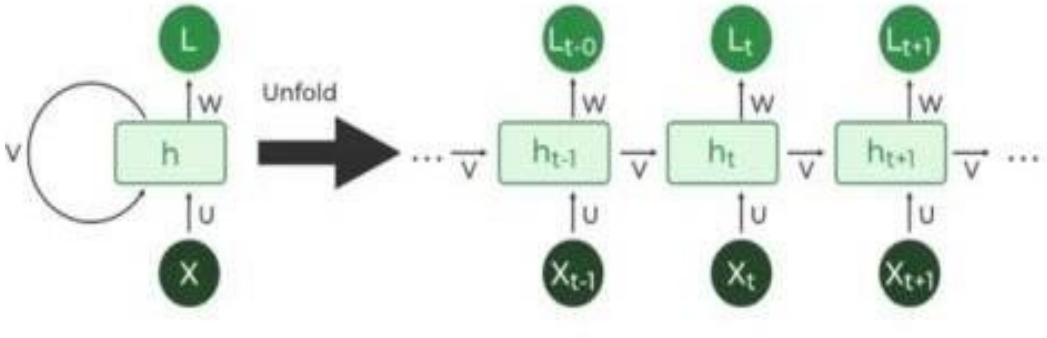


Figure 10: *The basic architecture of Recurrent Neural Network*

7.2 Differences Between FNNs and RNNs

Feedforward Neural Networks and Recurrent Neural Networks differ fundamentally in their data processing approach and memory capabilities, which determines their suitability for different types of tasks.

- ***Feedforward Neural Networks (FNNs):*** These networks process data in a single forward pass from input to output without retaining any previous information. They are best suited for tasks with independent inputs, such as image classification.

- ***Recurrent Neural Networks (RNNs):*** These networks contain loops that allow them to remember information from previous time steps, making them effective for sequential data and tasks requiring temporal context, like speech recognition and translation.

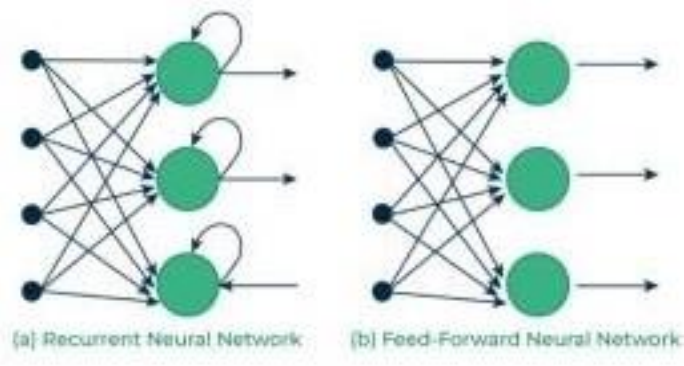


Figure 11: *Recurrent Vs Feedforward networks*

7.2.1 Types of RNNs

RNNs (Recurrent Neural Networks) are used to process sequential data, as they can remember information from previous time steps through a hidden state. Based on the number of inputs and outputs, there are four types of RNNs:

- ***One-to-One RNN:*** A single input produces a single output. Used for tasks like binary classification without sequential data.
- ***One-to-Many RNN:*** A single input generates multiple outputs over time, such as in Image captioning.
- ***Many-to-One RNN:*** A sequence of inputs generates a single output, useful for tasks like sentiment analysis.
- ***Many-to-Many RNN:*** A sequence of inputs generates a sequence of outputs, ideal for tasks like language translation [14].

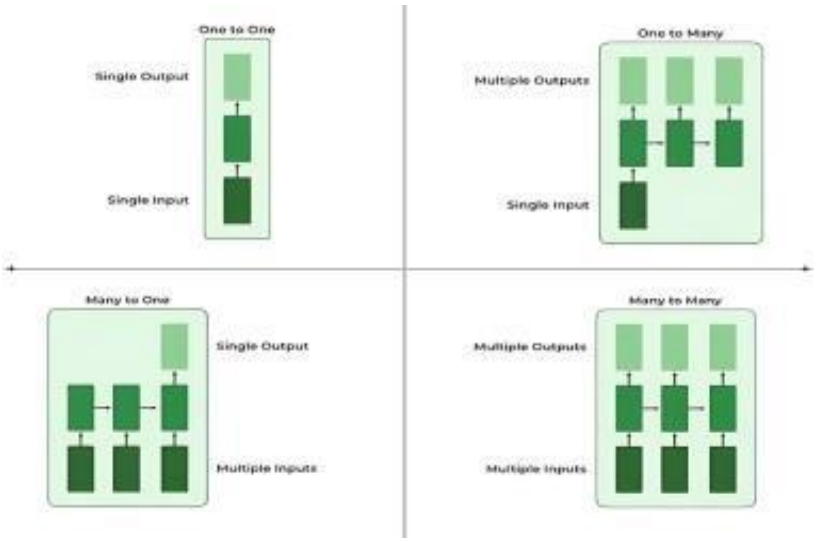


Figure 12: *Comparing RNN Architectures*

7.3 Convolutional Neural Networks (CNN)

Convolutional neural networks (CNNs) are an advanced type of artificial neural network (ANN), specifically designed to extract features from matrix data such as images or videos. These networks are among the most effective algorithms in computer vision and audio signal processing, due to their ability to recognize patterns and detect spatial relationships within the data. CNNs can also be used for speech recognition by converting audio signals into visual representations such as audio spectrograms and then analyzing them in the same way as images [14].

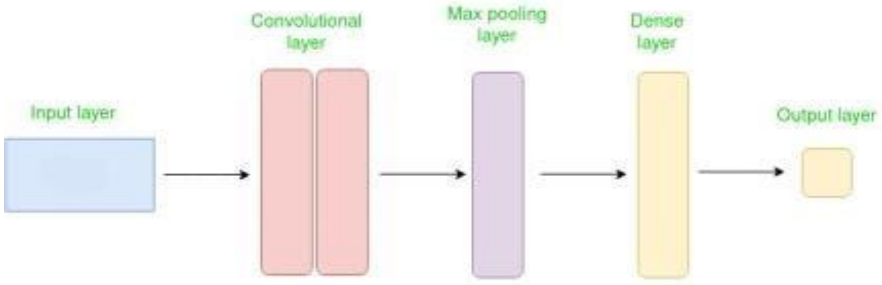


Figure 13: *Simple CNN architecture*

7.4 Main Components of a CNN

The main building blocks of CNNs are numerous essential components, each of which performs a crucial function in the processing and learning of image data. Knowing these components is crucial to understanding how CNNs work.

- ***Convolutional Layers:*** These layers apply convolutional operations to the input data, using filters (kernels) to detect features such as edges, textures, and more

  complex patterns. These convolutional operations help preserve spatial relationships between pixels.

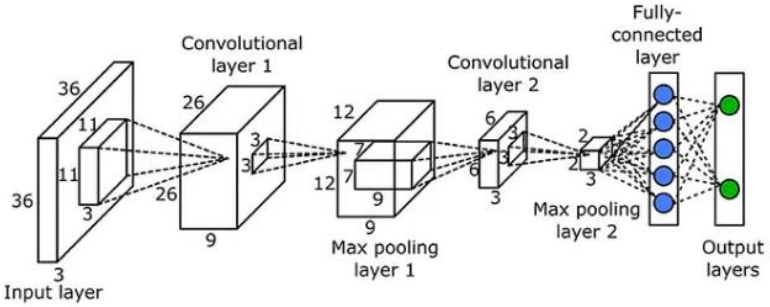


Figure 14: Convolutional Layer

- ***Pooling Layers:*** These layers reduce the spatial dimensions of the input data, reducing the computational complexity and number of parameters in the network. Max pooling is a common pooling operation, where the maximum value is selected from a set of neighboring pixels.

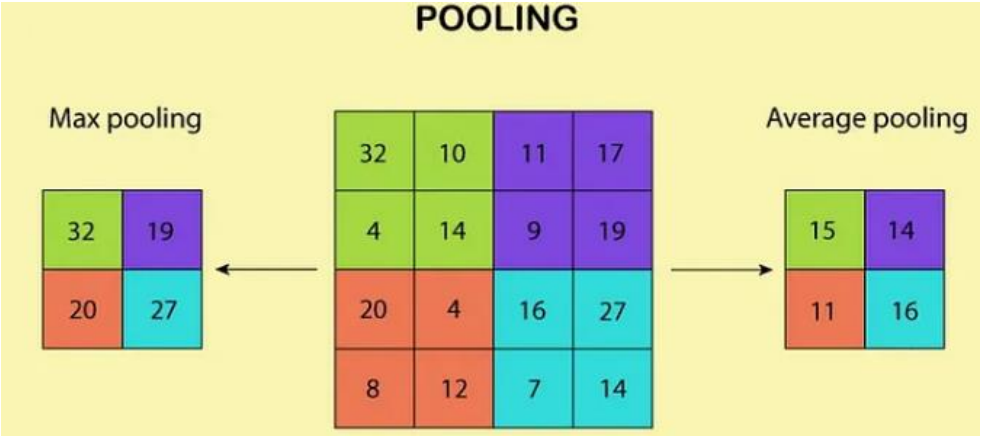


Figure 15: *MaxPooling Layer*

- ***Activation Functions:*** These add nonlinearity to the model, allowing it to learn more complex relationships in the data. One of the most common functions used is ReLU (Rectified Linear Unit).

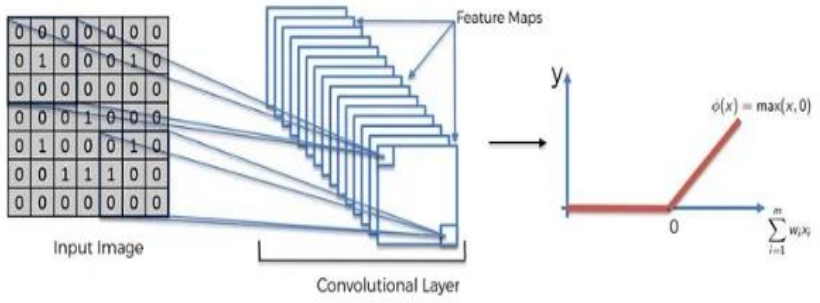


Figure 16: Activation Layer

- **Fully Connected Layers:** These layers are responsible for making predictions based on the high-level features learned by the previous layers. They connect each neuron in one layer to each neuron in the next layer.

  Dense and Flatten layers would come under this category.

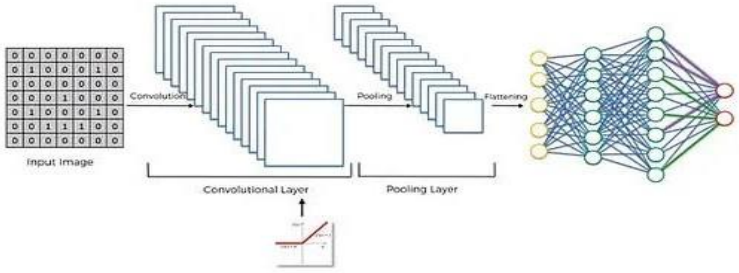


Figure 17: Dense Layer

7.4.1 How CNNs Work?

To understand the efficiency of Convolutional Neural Networks (CNNs), it is essential to explore how they process and analyze data through their various interconnected layers. The following points summarize the key steps in the operation of a CNN [15].

- ***Input Image:*** The CNN receives an input image, which is typically preprocessed to ensure uniformity in size and format.
- ***Convolutional Layers:*** Filters are applied to the input image to extract features like edges, textures, and shapes.

- ***Pooling Layers:*** The feature maps generated by the convolutional layers are down sampled to reduce dimensionality.
- ***Fully Connected Layers:*** The down sampled feature maps are passed through fully connected layers to produce the final output, such as a classification label.
- ***Output:*** The CNN outputs a prediction, such as the class of the image.

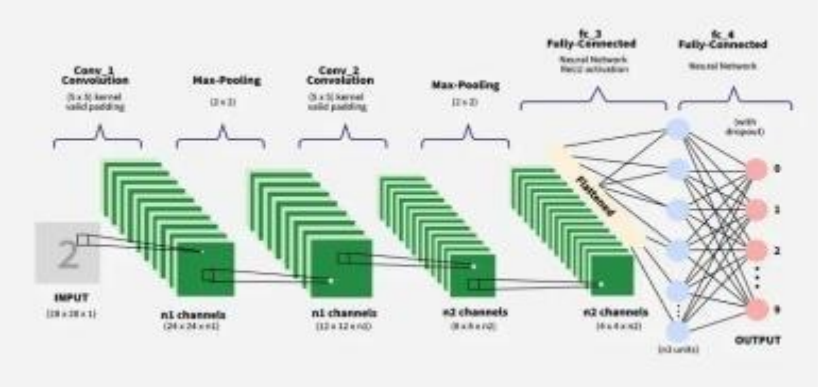


Figure 18: *The Mechanism of CNNs*

7.5 Advantages and Disadvantages of Convolutional Neural Networks

One of the most powerful tools in image and signal processing today is the Convolutional Neural Network (CNN). Like any other technology, when applied in real-world scenarios, CNNs present clear advantages and certain limitations that should be taken into account [16].

➢ ***Advantages of CNNs:***

- Excel at detecting patterns and key features in images, videos, and audio data.
- Maintain robustness against changes such as translation, rotation, and scaling.
- Facilitate end-to-end learning by eliminating the need for manual feature engineering.
- Capable of handling large datasets while maintaining a high level of accuracy.

➢ ***Disadvantages of CNNs:***

- Require significant computational resources and memory during training.
- Prone to overfitting when working with small datasets or without applying proper regularization techniques.

- Depend heavily on large volumes of labeled data for effective learning.
- Offer limited interpretability, making it difficult to precisely understand what the network has learned internally.

7.6 Applications of Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are used extensively in many areas, below are some of their most important applications:

- ***Object Detection:*** Identifying and locating objects within images.
- ***Video Analysis:*** Tracking objects and detecting events in video streams.
- ***Image Classification:*** Classifying images into categories (cats, dogs, cars...)
- ***Text Recognition (OCR):*** Extracting text from images or scanned documents. Art Generation: Creating or transforming images artistically [14].
- ***Speech Emotion Recognition:*** Analyze audio recordings or live speech to detect emotional states (happiness, sadness, anger, fear…).

  Used in:

  - Smart conversational chatbots
  - Customer service call analysis
  - Virtual voice assistants
  - Mental health and emergency helplines

8 Speech Emotion Recognition Datasets

Speech Emotion Recognition (SER) datasets play a considerable role in training and evaluating models intended for detecting emotion from vocal expressions. Such datasets consist of audio recordings wherein emotions are manually labeled, supplying critical examples for deep-learned models to learn from. Given these labeled audio files, models can learn to discriminate between various emotional states as expressed through speech, such as happiness, sadness, anger, or surprise. These datasets are crucial in the evolution of any speech emotion recognition system that calls for real-world applicability, such as HCI systems, virtual assistants, and emotion analysis [17].

9 Key Datasets in SER

Several well-known SER datasets are widely used for training emotion recognition models. The following are the most popular ones:

- ***RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song):*** It includes 24 professional actors (12 male and 12 female) performing two phrases with eight emotions: calm, happiness, sadness, anger, fear, surprise, disgust, and neutrality.

  The total number of audio files is 1440 [18].

- ***SAVEE (Surrey Audio-Visual Expressed Emotion):*** It consists of recordings from four British male speakers, each performing 120 phrases, covering seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutrality [19].

- ***CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset):*** It includes 7442 audio clips from 91 actors (48 male and 43 female), aged between 20 and 74 years. It covers six emotions: anger, disgust, fear, happiness, sadness, and neutrality, with varying levels of emotional intensity [20].

- ***TESS (Toronto Emotional Speech Set):*** It contains recordings from two women (aged 26 and 64) performing 200 target words within the phrase "Say the word," covering seven emotions: anger, disgust, fear, happiness, sadness, pleasant surprise, and neutrality. The total number of audio files is 2800 [21].

- ***CASIA (Chinese Academy of Sciences Emotional Speech Corpus):*** This corpus contains emotional utterances spoken by Chinese native speakers expressing different emotions (anger, fear, happiness, sadness, surprise, neutral, disgust). Although it is in Mandarin, it is still useful due to the universal nature of emotional expression in vocal signals [22].

10 Core Audio Features for Speech Emotion Detection

Speech emotion recognition systems extract a number of audio features that would represent the relevant acoustic characteristics of speech in a way intelligible and manipulable by machine learning models. These features describe the variations in tone, intensity, rhythm, and frequency,

which generally correlate to emotional states, including anger, sadness, happiness, fear, …. [23].

Below are the most commonly used features:

10.1 MFCC - Mel-Frequency Cepstral Coefficients

These shape the spectral footprint of the audio signal so as to reflect the patterns of human auditory perception. Probably the most used features in tasks specific to speech.

- ***Purpose:*** General Tone, rhythm, and devolution characteristics.

10.2 Chroma Short-Time Fourier Transform

Measures the tone of the 12 different pitch classes (semitones of the musical octave).

- ***Purpose:*** Useful when detecting pitch variation and harmonic content that can differ according to emotions.

10.3 ZCR- Zero Crossing Rate:

The number of times that a signal changes its sign: that is, crosses the zero amplitude axis.

- ***Purpose:*** For indicating whether a sound is noisy or sharp, frequently associated with emotions like anger or excitement.

10.4 RMSE- Root Mean Square Energy

Average power (loudness) of the signal.

- ***Purpose:*** to tell the difference between calm emotions like sadness,from powerful emotions like joy or anger.

10.5 Mel spectrogram

A spectrogram basically shows how content of frequency in time source varies. The same is converted to Mel Scale for mimicking the way in which humans hear sounds: It emphasizes the perceptual frequencies. Capture Timbre, Rhythm, and Tonal Quality.

11 Conclusion

In conclusion, this chapter provides a comprehensive overview of the theoretical and practical foundations of Artificial Intelligence, with a focus on CNN and its use in speech emotion recognition. It also reviews some of the most well-known datasets in the field and their significant

role in the development of intelligent models.

Chapter 3

System Implementation and Technologies Used

1 Introduction

This chapter delves into the technical implementation of the speech emotion recognition system. It begins by detailing the programming language and development tools used and then describes the essential supporting libraries for the building and training of the model. The chapter also describes how the dataset was prepared for the modeling work, including an augmentation procedure to boost the system's generalizability. Audio feature extraction, a crucially important step that allows the model to grasp emotional cues present in speech, is elaborated upon. Next, the principal machine learning algorithm of this study - CNN - is introduced and thoroughly explained in terms of model architecture. Finally, some training results are presented, and an analysis is performed concerning the model performance with regard to accuracy and loss metrics.

2 Programming Language (Python)

Python is a high-level language with object-oriented features and an interpreter being the main feature. It is easy to learn, with a simple syntax, dynamic typing, and a rich standard library; all of these characteristics together provide a friendly environment for rapid application development that reinforces code readability and reuse, and efficient debugging through fast edit-test cycles [24].

3 Development Tools Used

3.1 Visual Studio Code (VS Code)

Microsoft-developed shareware widely spread in usage, utilized for writing primarily Python and web-related code (HTML, CSS, JavaScript) in this project. Some features that boosted the efficiency in developing and maintaining the entire codebase include syntax highlighting, intelligent code completion, Git integration, and debugging tools [25].

3.2 Jupyter Notebook

Another interactive development environment used during the experimentation phase. Executing the code in separated cells enabled testing, visual assessment of audio features like waveforms and spectrograms, and tuning the emotion recognition model iteratively [26].

3.3 Anaconda Distribution

Anaconda is a distribution of Python geared toward scientific computing and data science. It was used for virtual environment management and for installing needed packages (including TensorFlow, Librosa, Pandas, etc.). It also includes Jupyter Notebook, which enhances launching

and controlling experiments in an enclosed environment [27].

3.4 Framework: Flask (for the backend web server)

Flask is a lightweight web framework written entirely in the Python language, used to build the backend to any web application. Being so, it can achieve routing, request handling, API setup, and machine-learning model deployment. The minimalist structure of Flask allows the construction of highly scalable, extremely complex web servers while giving the developers total control over the architecture [28].

3.5 Web Interface Technologies: HTML, CSS, and JavaScript

The web interface was developed using three front-end core technologies:

HTML, CSS, and JavaScript. These technologies work together for the structure, visuals, and interactivity for the end user's experience [29].

- ***HTML (HyperText Markup Language):*** is the standard markup language used to define the structure and content of web pages and such things as headings, paragraphs, links, images, and other components that form the layout of a webpage.

- ***CSS (Cascading Style Sheets):*** is a stylesheet language employed for telling the appearance and formatting of web document information, which includes colors, fonts, spacing, and layout. Every layout enhancement improves the user interface and visual sophistication.

- ***JavaScript***: is a dynamic and interactive scripting language at a higher level in the sense that it is interpreted and runs best in the browser, enabling developers to create responses to the actions of their users, real-time manipulation of HTML and CSS, and enhanced client-side user-friendly functionalities.

4 Libraries used in the Model

The speech emotion recognition system requires a battery of special Python libraries/tools for development to ensure strength and efficiency. Each of these was instrumental in data handling, feature extraction, model development, training, and evaluation:

4.1 NumPy

The library used for numerical operation and efficient array operation during data preprocessing and model input [30].

4.2 Pandas

Employed to structure and manipulate datasets using DataFrames, thus making the management of audio metadata and labels easier [31].

4.2.1 DataFrame (from Pandas)

A DataFrame is a two-dimensional, size-mutable, and heterogeneous tabular data structure with labeled axes (rows and columns). This project uses DataFrame to organize the model input and processing features and metadata in an efficient manner.

Two major DataFrames were created:

- **emotion_df:** It contained the emotion labels for the individual audio files.
- **path_df:** It contained the corresponding file paths for individual audio samples.

The emotion_df and path_df were then merged into one comprehensive DataFrame with two major columns:

	Emotions	Path
3635	surprise	C:\Users\Rania\Downloads\datasat\Train\Surpris...
3636	surprise	C:\Users\Rania\Downloads\datasat\Train\Surpris...
3637	surprise	C:\Users\Rania\Downloads\datasat\Train\Surpris...
3638	surprise	C:\Users\Rania\Downloads\datasat\Train\Surpris...
3639	surprise	C:\Users\Rania\Downloads\datasat\Train\Surpris...

Figure 19: *Tail View of the Dataset Entries*

4.2.2 Reasons for DataFrames for the Project

In the current project, DataFrames were created for the following reasons:

- ***Organization of Data:*** As the dataset consists of many audio files whose sound conveys different emotions, tabulating them makes them easily remaining available for any analysis.
- ***Easier Processing of Data:*** With the help of the DataFrames, we can apply the processing step conveniently, for instance, while adding noise, stretching, or pitch-shifting of audio files.

- ***Better Model Training:*** Supervised learning in the model requires data to be well organized and accurately labeled. Through the DataFrame, an assurance is provided that the signed data and the emotion labels are accurately paired, thereby forming an essential step in supervised learning.

Such structured format enables better organization of data, making the loading, shuffling, visualization, and feeding of data into the model easier during training and testing, thus improving the efficiency of data organization and preprocessing.

4.3 Librosa

A versatile library specializing in the analysis of audio and music, Librosa finds its application in extracting significant features such as MFCCs, Chroma, RMS, Zero-Crossing Rate, and Mel spectrograms from the speech signal [32].

4.4 Matplotlib

This library is called Matplotlib. With Matplotlib, you get a powerful library for working in Python that handles visualization in the form of static, animated, or even live, and interactive visualizations. Everything can be plotted to the figural level regarding how well one can make a plot, histogram, bar chart, error chart, among others [33].

In this creation, it was used to plot:

- Audio waveform signals that are in the time domain.
- Spectrograms indicating the frequency content of signals.
- Training curves such as accuracy and loss across epochs.
- Confusion matrices for a visual evaluation of model performance.

4.5 Seaborn

Seaborn is a high-level interface for Matplotlib and is therefore the Python data visualization library whose organization is based on Matplotlib but is much cleaner and offers more with respect to making more attractive and informative statistical graphics [34].

Drawing something very complicated such as heat maps, violin plots, and distribution plots will now be easier. In this project, Seaborn has its edge by:

- Presenting confusion matrices in a cleaner way.
- Improving the appearance of graphics from standard Matplotlib.

4.6 TensorFlow

TensorFlow is an open-source end-to-end machine learning-platform module development by Google. It consists of a complete ecosystem of tools, libraries, and community resources that enable researchers and developers to easily create and project ML-empowered applications [35].

TensorFlow is used as a basis for the implementation of deep learning operations and management in this project.

4.7 Scikit-learn

Used for data splitting, evaluation metrics, and generating reports like precision, recall, F1-score, and confusion matrices [36].

4.8 Os (Operating System Interface)

A standard Python module that allows interaction with the operating system, especially for file path operations such as listing the contents of directories and joining paths [37].

4.9 sys (System-specific Parameters and Functions)

A built-in module that provides access to system-specific parameters and functions. It was used in situations such as appending paths or handling runtime environments when required [38].

4.10 IPython.display. Audio

A Jupyter utility that allows embedding and playing audio files directly in the notebook interface.

This was used for manually inspecting and validating audio samples during data exploration [39].

5 Dataset Structuring and Preparation

In this study, a hybrid dataset has been formed by fusing three benchmark emotional speech corpora**: RAVDESS, SAVEE, iEMOCAP**, and **CASIA**. Each of these datasets brings its own peculiarities regarding speaker diversity, emotion expression, and/or acoustic variation, thereby giving the model a better generalization power over a range of different speech acts and emotional cues.

5.1 Audio Data Preparation for Emotion Recognition

To provide linguistic and structural consistency, all audio samples in English were placed in folders that matched the format in the datasets. Each emotion category was assigned an emotion

label and balanced to contain 520 audio files in total. The table in the follow describes the data organization.

Table 1: Audio Data Preparation for Emotion Recognition

Emotion	Label (File Symbol)	Number of Audio Files
Anger	a/ such as a01.wav, a02.wav, ..., a520.wav.	520
Disgust	d/ such as d01.wav, d02.wav, ..., d520.wav.	520
Fear	f/ such as f01.wav, f02.wav, ..., f520.wav.	520
Happiness	h/ such as h01.wav, h02.wav, ..., h520.wav.	520
Neutral	n/ such as n01.wav, n02.wav, ..., n520.wav.	520
Sadness	sa / such as sa01.wav, sa02.wav, ..., sa520.wav.	520
Surprise	su / such as su01.wav, su02.wav, ..., su520.wav.	520
Total	-	3640

5.2 Audio Feature Visualization

Audio signal visualization is an important way to understand the structure and distribution of the speech data. There were two main ways to visualize this project.

- **Waveplot:** The signal is presented in the time domain, representing amplitude in relation to time [40].

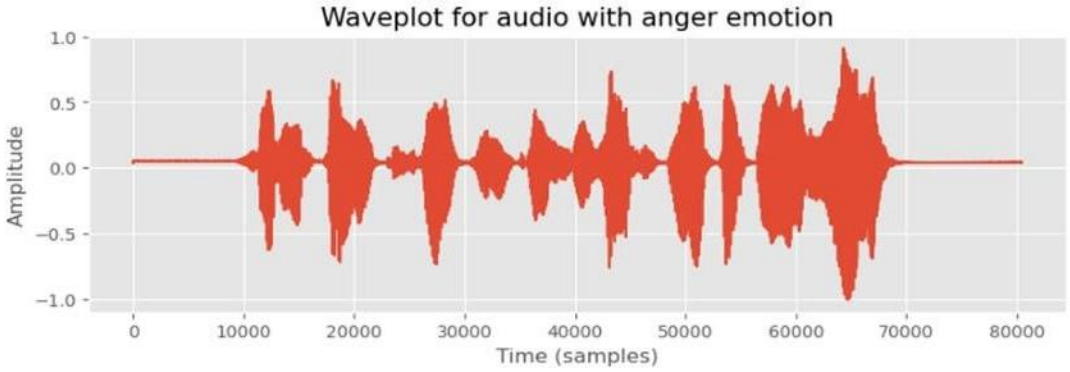


Figure 20: *Waveform of Angry Speech Signal*

- **Spectrogram:** The contents of the signal frequencies are displayed as a function of time,

depicting patterns that distinguish the emotions.

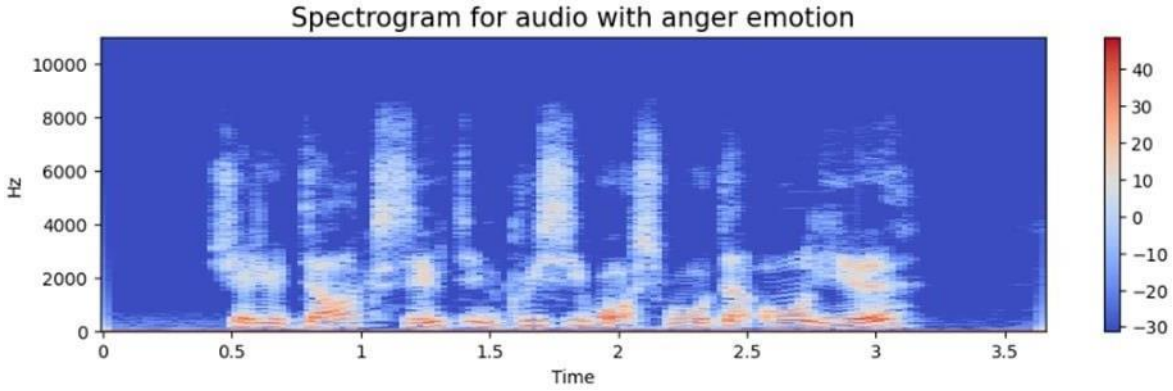


Figure 21: *Spectrogram of a Speech Signal Expressing Anger*

6 Data Augmentation

Data augmentation is the introduction of controlled perturbations to the original training data. Data augmentations generate new synthetic data samples. For this project, multiple audio-based augmentation methods were applied on the dataset to enhance robustness and generalization capability of the resultant model.

The above transformations were imputed on audio data:

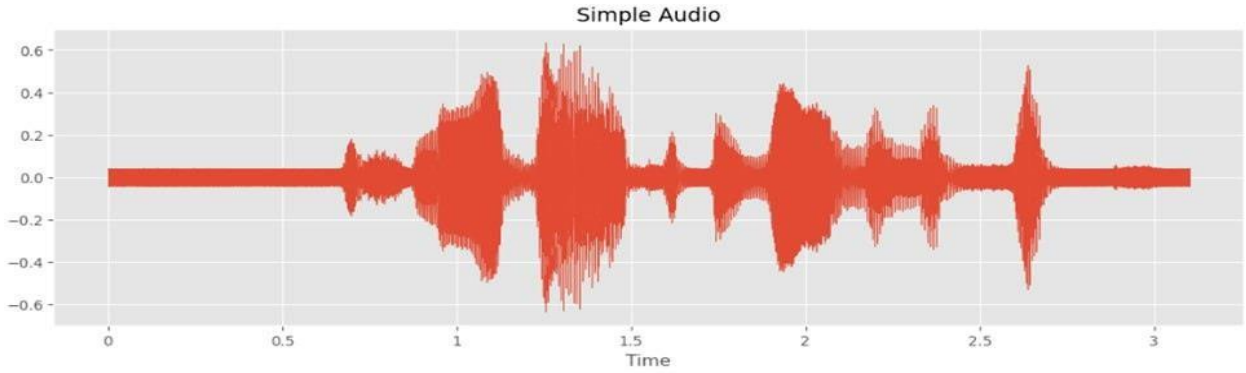


Figure 22: *Waveform Simple Audio*

6.1 Noise Injection:

Random Gaussian noise is added to protect emotional content just like real audio distortions would do, which gives much resilience against background interference for the model.

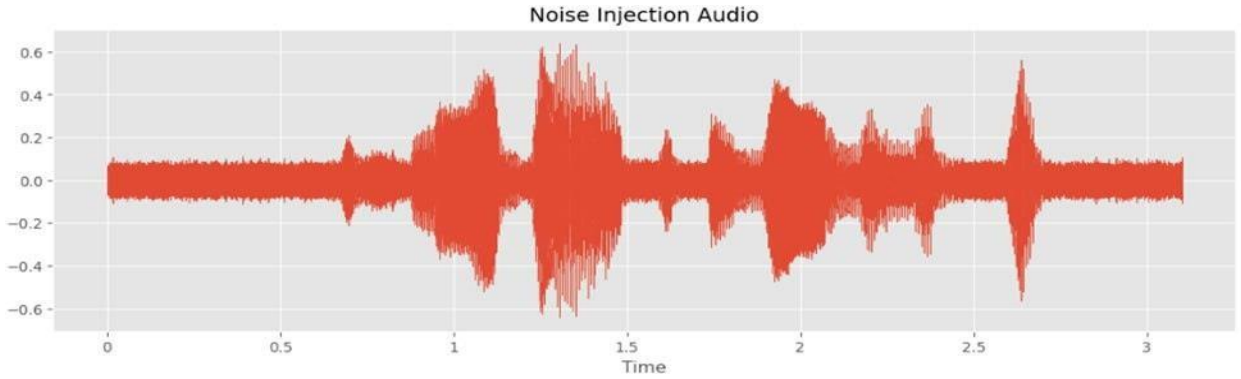


Figure 23: *Waveform of Speech Signal with Noise Injection*

6.2 Time Stretching:

the speed of the audio by changing the rhythm without changing the pitch would make the model better able to adapt for change in tempo by which speech would occur and articulation.

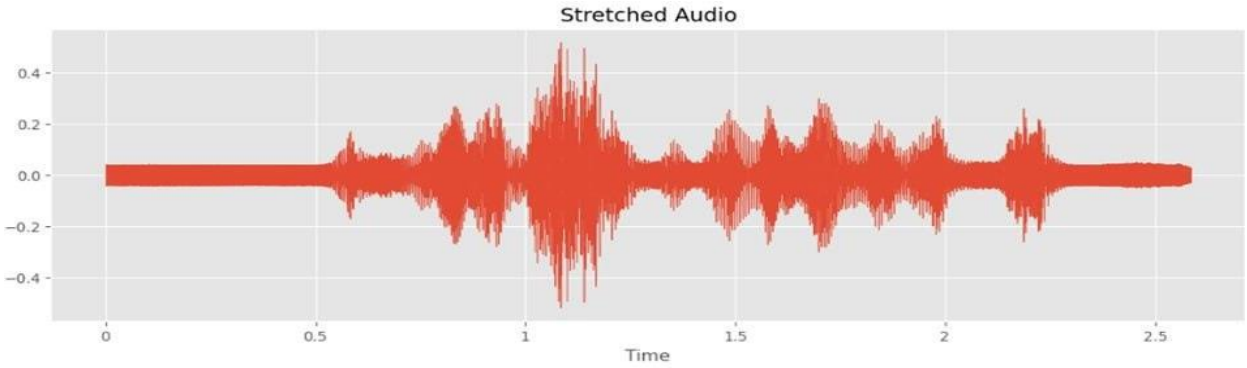


Figure 24: *Waveform of signal with Stretched*

6.3 Pitch Shifting:

This pitch shifting would take the audio pitches up or down in respect to original tempo, and this also helps the model to understand almost all emotional cues in different tones of the voice.

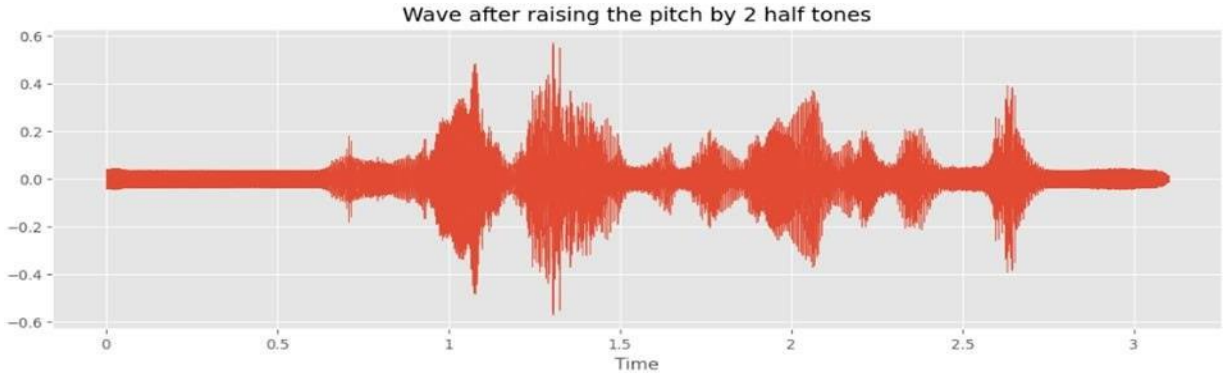


Figure 25: *Waveform of signal with Pitch Shifting*

6.4 Temporal Shifting:

Audio signals were shifted forwards or backwards within the time axis, helping models generalize over temporal misalignments or timing variations in speech delivery.

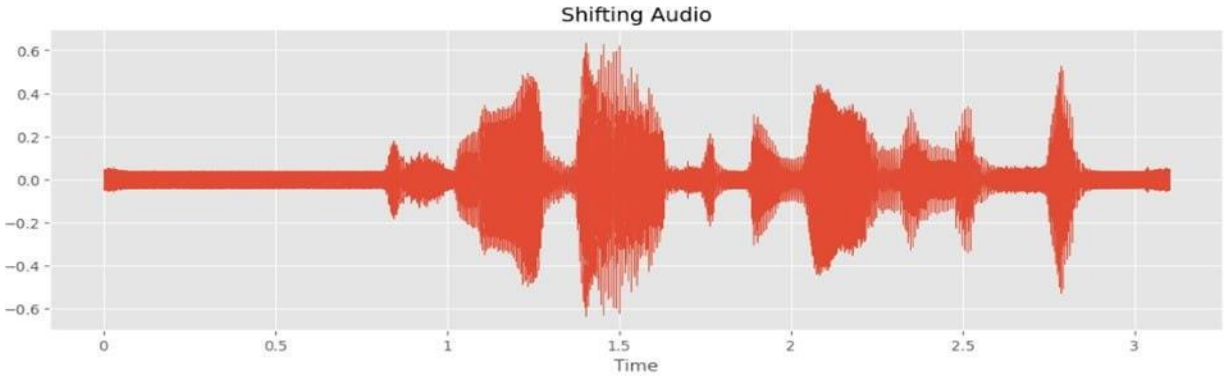


Figure 26: *Waveform of signal with Temporal Shifting*

Careful design augmentations resulted in maintaining the original emotion label for each sample,which added number and diversity in the training dataset without compromising the integrity of the labels. This results in a stronger and more invariant system in emotion recognition to different real- world audio scenarios.

All extracted features (from the original and augmented audios) were standardized to the same shape (182, 40) to ensure consistency during training and prediction.

7 Audio Feature Extraction for Emotion Recognition

In this specific project, the focus was placed on MFCC features due to their proven effectiveness in speech-based tasks. To improve the robustness and generalization of the model, data augmentation techniques were applied before extracting the features.

- **MFCC – Mel-Frequency Cepstral Coefficients:**

In this project, only MFCCs were utilized as the main input feature due to their effectiveness in capturing essential speech characteristics. Their ability to represent the human auditory perception made them suitable for recognizing emotional patterns in voice signals.

Extracted from:

- The original audio
- Noise-injected version
- Pitch-shifted version

- **MFCC Extraction Process:**

➢ Segment the audio into short frames.

➢ Apply FFT to each frame.

➢ Use Mel Filter Banks to mimic human hearing.

➢ Apply logarithmic scaling.

➢ Use DCT to decorrelate the coefficients.

➢ Final Output Shape: (182 frames, 40 coefficients)

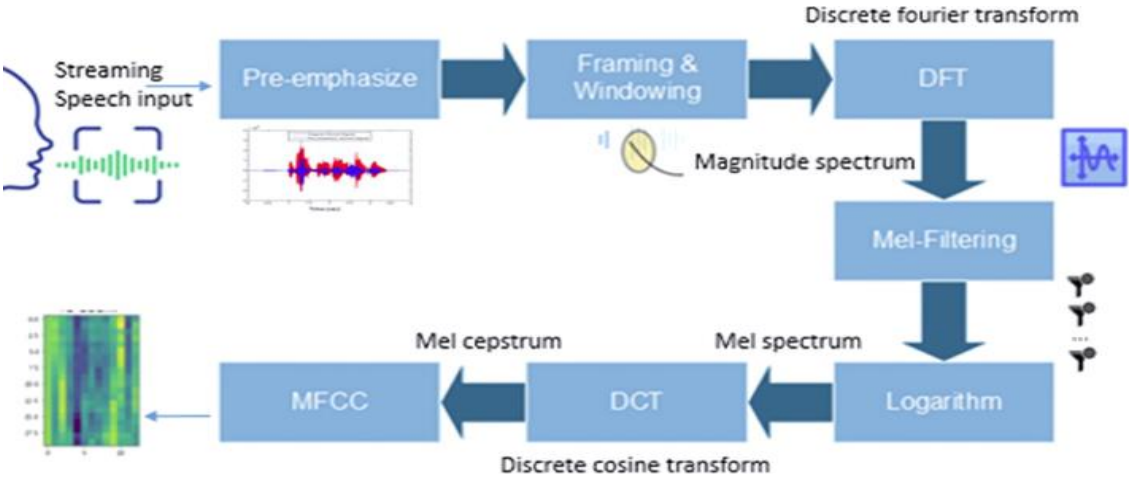


***Figure 27**: MFCC Feature Extraction Process for Speech Signals*

8 Algorithm Used in This Approach

In this work, Convolutional Neural Network (CNN)-based architecture is adopted to classify emotions in the speech. CNNs are very useful for structured data which have spatial or temporal relationships which are important for audio signals where the representation of patterns over time is significant.

- **Why Chose CNN ?**

➢ CNN automatically extracts meaningful features from raw input, which eliminates the task of doing handcrafted feature engineering.

➢ It is good at modeling local patterns in time series data, such as tone or pitch changes in audio.

➢ The architecture is efficient and scalable and performs well even on large datasets of audio.

9 Model Architecture and Layer Description

The CNN model used in this project possesses a sequential architecture consisting of the

following layers:

9.1 Conv1D Layers

To extract temporal patterns from the audio signal, the model uses 1D convolutional layers that act as feature detectors across time. These layers are configured in the model as follows:

- One-dimensional convolutional filters (Conv1D) are applied to extract temporal features from the audio signals.
- The model includes three Conv1D layers:
  - The first layer has 256 filters to extract low-level features from the raw signal.
  - The second layer has 128 filters to learn more abstract, mid-level representations.
  - The third layer has 64 filters, focusing on high-level features.
- All layers use ReLU activation and 'same' padding to preserve the output dimensions**.**

9.2 BatchNormalization Layers

To stabilize and accelerate the training process, BatchNormalization layers are applied to normalize the outputs of convolutional layers. Their application in the model is as follows:

- Applied after each Conv1D layer to normalize the output.
- Helps stabilize and accelerate training by reducing internal covariate shift.
- Improves overall model performance and reduces the sensitivity to initialization and learning rates.

9.3 MaxPooling1D Layers

To reduce the temporal dimension and retain the most relevant features, MaxPooling1D layers are used after normalization. These layers are integrated into the model as follows:

- Applied after each BatchNormalization layer.
- Reduces the temporal dimension by taking the maximum value over a pool size of 2.
- This downsampling reduces computational cost while retaining the most important features.

9.4 Dropout Layers

To prevent overfitting and improve the model's generalization, Dropout layers are included during training. The dropout strategy is applied as follows:

- Dropout layers are applied with a rate of 0.3.
- Randomly deactivate a fraction of the neurons during training.

- Helps prevent overfitting and enhances the model's generalization capability.

9.5 Flattening Layer

To prepare the data for the fully connected layers, the multidimensional output from the previous layers is flattened into a 1D vector. This transformation is handled as follows:

- Converts the 3D output of the convolutional layers into a 1D vector.
- Prepares the data for input into the fully connected (Dense) layers.
- Acts as a transition from spatial processing to classification.

9.6 Dense Layer with L2 Regularization

To perform classification, the model uses dense layers with L2 regularization to reduce overfitting. These layers are structured in the following way:

- The first Dense layer consists of 64 units with ReLU activation and L2 regularization to mitigate overfitting.
- Followed by a Dropout layer with a rate of 0.3 for additional regularization.
- The final Dense layer has 7 units with softmax activation.

9.7 Output Layer (Softmax)

To produce the final prediction, a softmax output layer converts the network's output into probabilities for each emotion class. This final step is implemented as follows:

- It produces a probability distribution for 7 emotion classes.
- This soft-max activation function is used in this layer to ensure that the class sums to one.

Model: "sequential"

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 182, 256)	51,456
batch_normalization (BatchNormalization)	(None, 182, 256)	1,024
max_pooling1d (MaxPooling1D)	(None, 91, 256)	0
conv1d_1 (Conv1D)	(None, 91, 128)	163,968
batch_normalization_1 (BatchNormalization)	(None, 91, 128)	512
max_pooling1d_1 (MaxPooling1D)	(None, 46, 128)	0
dropout (Dropout)	(None, 46, 128)	0
conv1d_2 (Conv1D)	(None, 46, 64)	41,024
batch_normalization_2 (BatchNormalization)	(None, 46, 64)	256
max_pooling1d_2 (MaxPooling1D)	(None, 23, 64)	0
flatten (Flatten)	(None, 1472)	0
dense (Dense)	(None, 64)	94,272
dropout_1 (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 7)	455

Figure 28: *Model Architecture Summary for Speech Emotion Classification*

10 Training Results and Performance

The performance of the Convolutional Neural Network (CNN) model used for emotion classification was monitored during training for 16 epochs. The graphs below show the loss and accuracy curves for the training and testing phases:

10.1 Training & Testing Loss (Left Plot)

- The training loss (blue line) is seen to decrease smoothly, showing that the model is learning properly.
- The testing loss (red line) initially mimics the training loss but starts to show slight fluctuations after epoch 11, suggesting slight overfitting from hereon.

- The final testing loss stabilizes around 0.3, while the final training loss is below 0.2.

10.2 Training & Testing Accuracy (Right Plot)

- Training accuracy (blue line) shows a steady increase and almost reaches 0.97, which is a good indication for the strong learning capability of the model on the training set.
- Testing accuracy (red line) improves significantly and reaches ~0.96, indicating good generalizability over unseen data.
- Some small fluctuations on the testing accuracy after epoch 11 could be attributed to the variability of the validation data.

10.3 Interpretation

- The model went on to achieve high accuracy (~96%) alongside low testing loss, which indicates well training and good generalization.
- Some overfitting can be seen from epoch 11, suggesting that Infusing methods such as EarlyStopping, Regularization, or extra data augmentation would be valuable for foreseeable improvements.

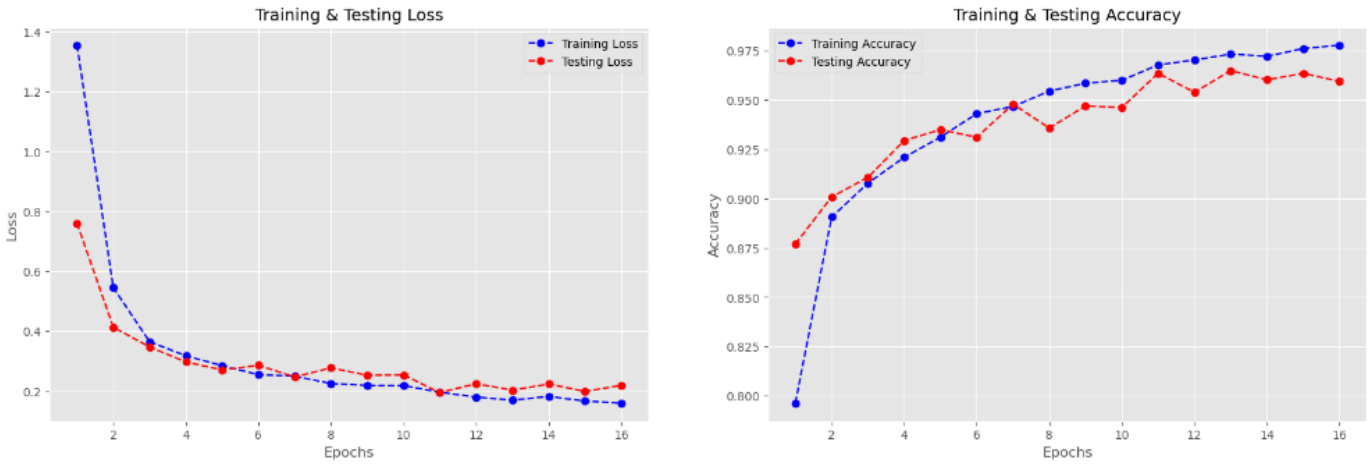
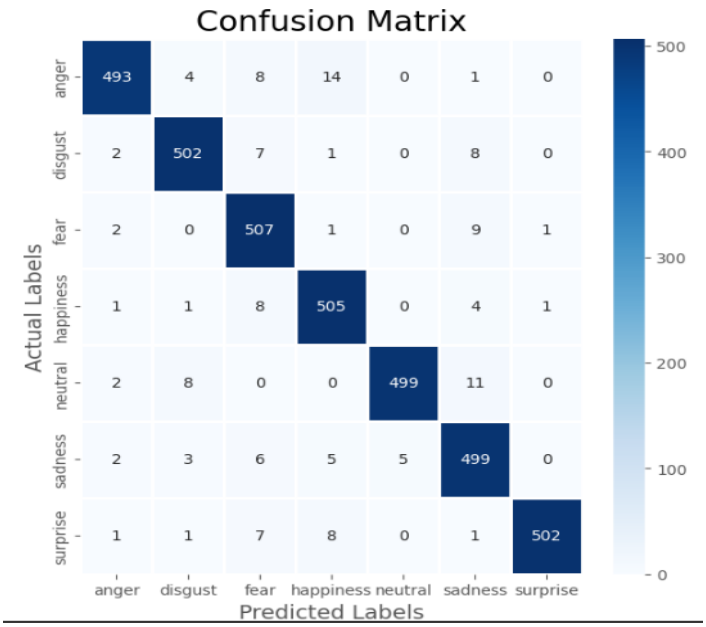


Figure 29: *Model Training & Testing Loss and Accuracy Curves*

	Predicted Labels	Actual Labels
0	surprise	surprise
1	surprise	surprise
2	happiness	happiness
3	surprise	surprise
4	neutral	neutral
5	surprise	surprise
6	anger	anger
7	surprise	surprise
8	neutral	neutral
9	happiness	happiness

Figure 30: *Predicted vs. Actual Emotion Classifications*



	precision	recall	f1-score	support
anger	0.98	0.95	0.96	520
disgust	0.97	0.97	0.97	520
fear	0.93	0.97	0.95	520
happiness	0.95	0.97	0.96	520
neutral	0.99	0.96	0.97	520
sadness	0.94	0.96	0.95	520
surprise	1.00	0.97	0.98	520
accuracy			0.96	3640
macro avg	0.96	0.96	0.96	3640
weighted avg	0.96	0.96	0.96	3640

Figure 31: *Confusion Matrix and classefication report*

11 Conclusion

In summary, I have provided a detailed pipeline for designing a deep learning system for emotion recognition from speech. Each part of the pipeline was crafted with the utmost attention, from tool and technology choice to proper dataset structure and enhancement, always

keeping in view the optimization of model performance. The use of CNNs was believed to help in learning important features from audio, and experimental evidence suggests the model possesses very good generalization capabilities. Such outputs set the perfect foundation for further enhancements and eventual implementation in practical applications like affective computing and HCI.

Chapter 4

SER System Implementation & Some User Interfaces

1 Introduction

This chapter takes you through the detailed design of user interfaces for the Sound Emotion Web Application. The system was modular in construction, where each page is designed to perform a specific task in the workflow of emotion recognition from audio. In essence, the web application uses Flask as a backend server, while modern front-end technologies—HTML, CSS, and JavaScript—are integrated to maintain responsiveness and intuitive feel for the user. Each interface, from training, testing, and prediction, is designed for a smooth interaction experience for the beginners as well as advanced ones. The subsequent sections describe each page's visual layout, interactivity, and functional behaviors.

2 Interface Description: Sound Emotion Web Application

2.1 Home Page

The home page is where the project is initiated on a tasteful note alongside a dark view enhanced by a geometric wireframe face intended to symbolize voice-emotion recognition systems.

And with that, there is a navigation bar carrying links to all major sections, namely: Home, Training, Testing, Prediction, and Contact.

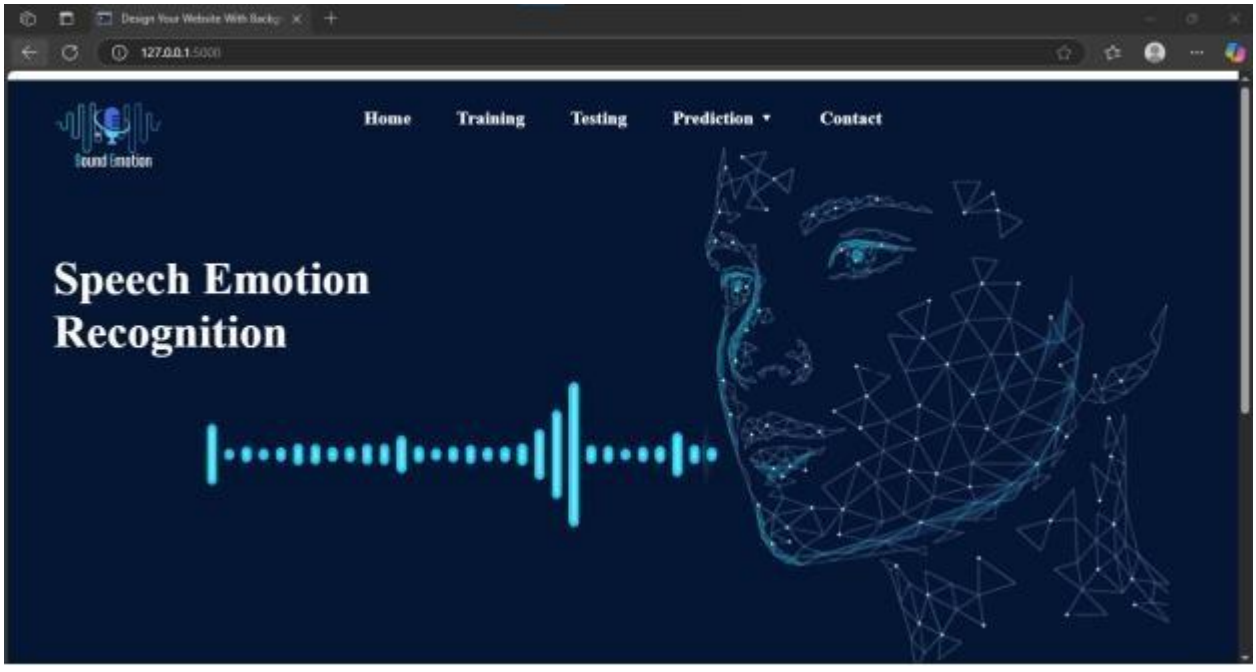


Figure 32: *Home Page*

2.2 Training Interface

This page received from the user the ability to train on the model with the dataset provided as a ZIP file.

- Only files with .zip extensions are accepted.
- The training dataset can be uploaded by drag-and-drop or through buttons labeled "Choose ZIP File."
- The user can select any number of training epochs for conducting training
- The "Train" button will be activated once the user uploads a ZIP file and selects an epoch value.
- A progress bar is then displayed to visualize the training status.
- After the training has been completed, the results will be displayed, such as:
    - *Training Accuracy*
    - *Validation Accuracy*

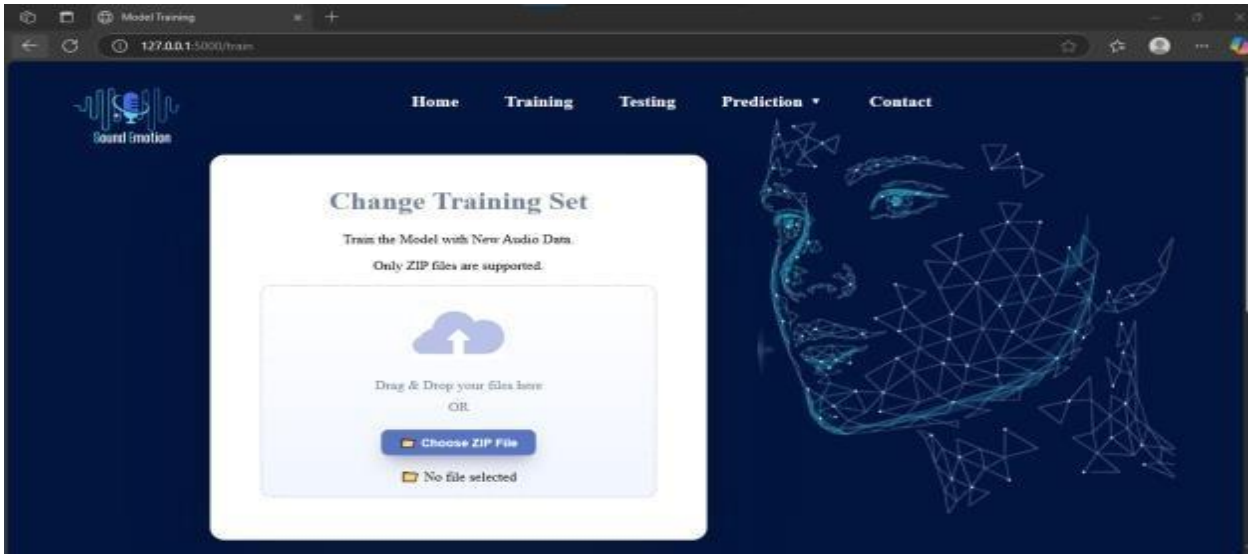


Figure 33: *Training Interface*

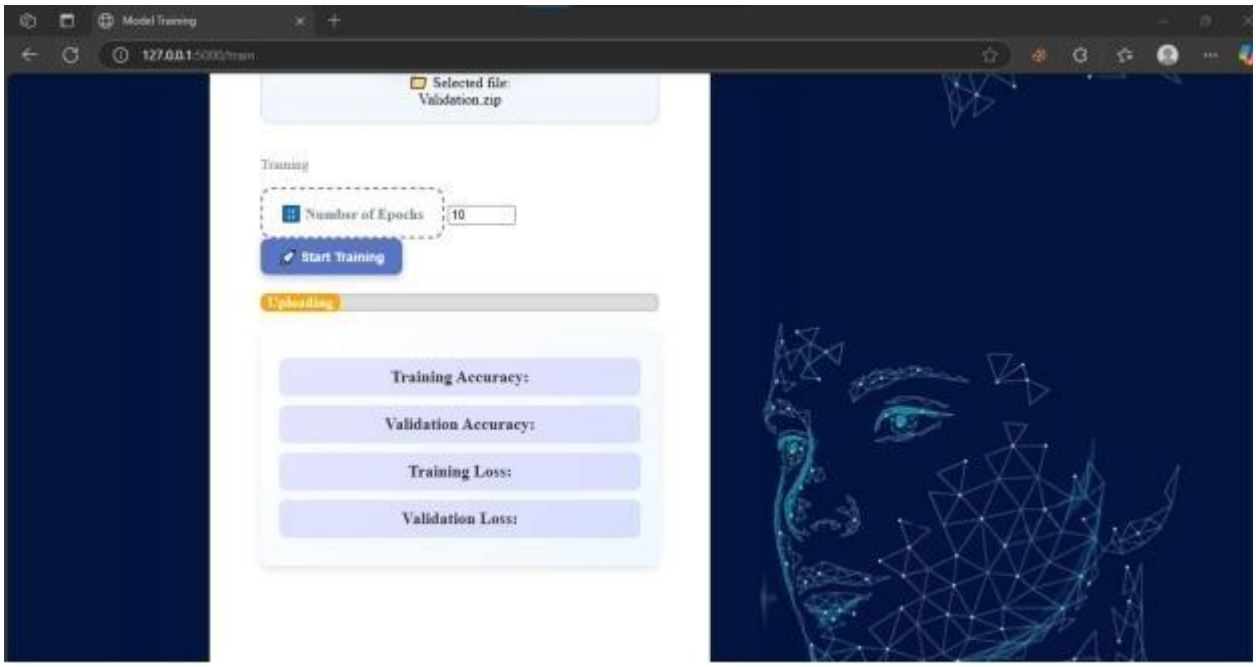


Figure 34: *Start Training interface*

Figure 35: *Result Training interface*

2.3 Testing Interface

This page permits the user to test the trained model using a test data ZIP file.

- The user uploads the test data in a .zip format.
- Once uploaded, a "Test" button appears.
- During the test, a progress bar shows the status.
- Upon finishing testing, a "Show Results" button shows up. Results are shown in a pop-up modal, including:
  - *Test Accuracy*
  - *Confusion Matrix for detailed visualization of performance.*

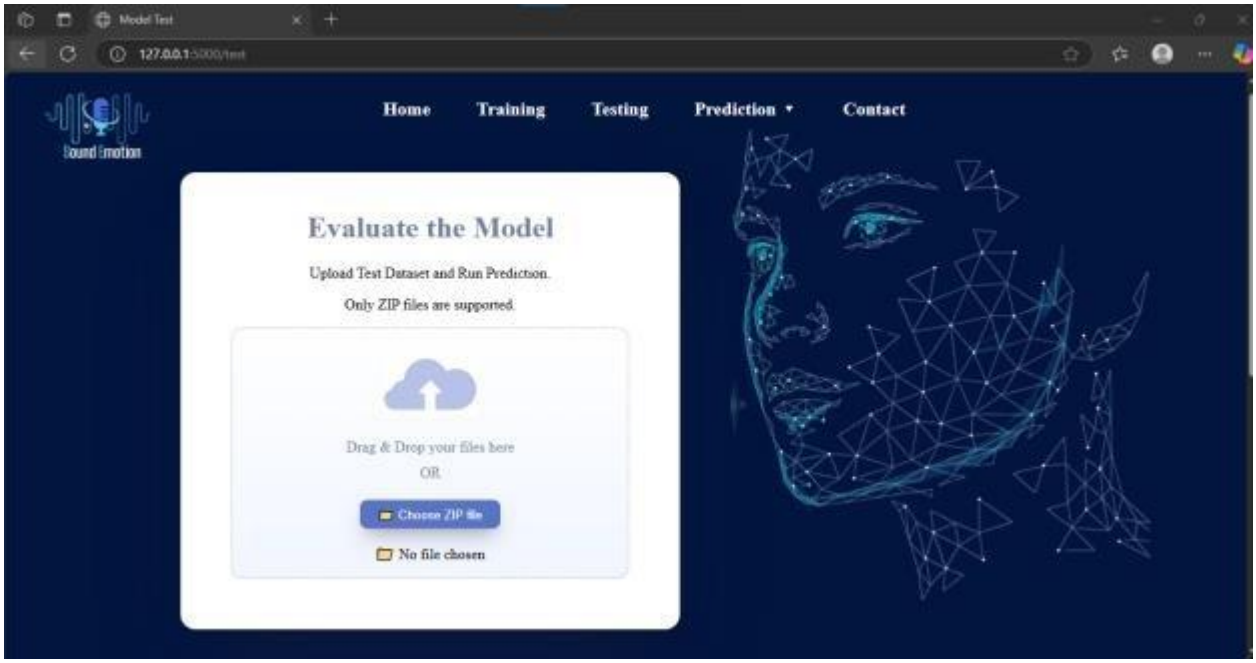


Figure 36: *Testing Interface*

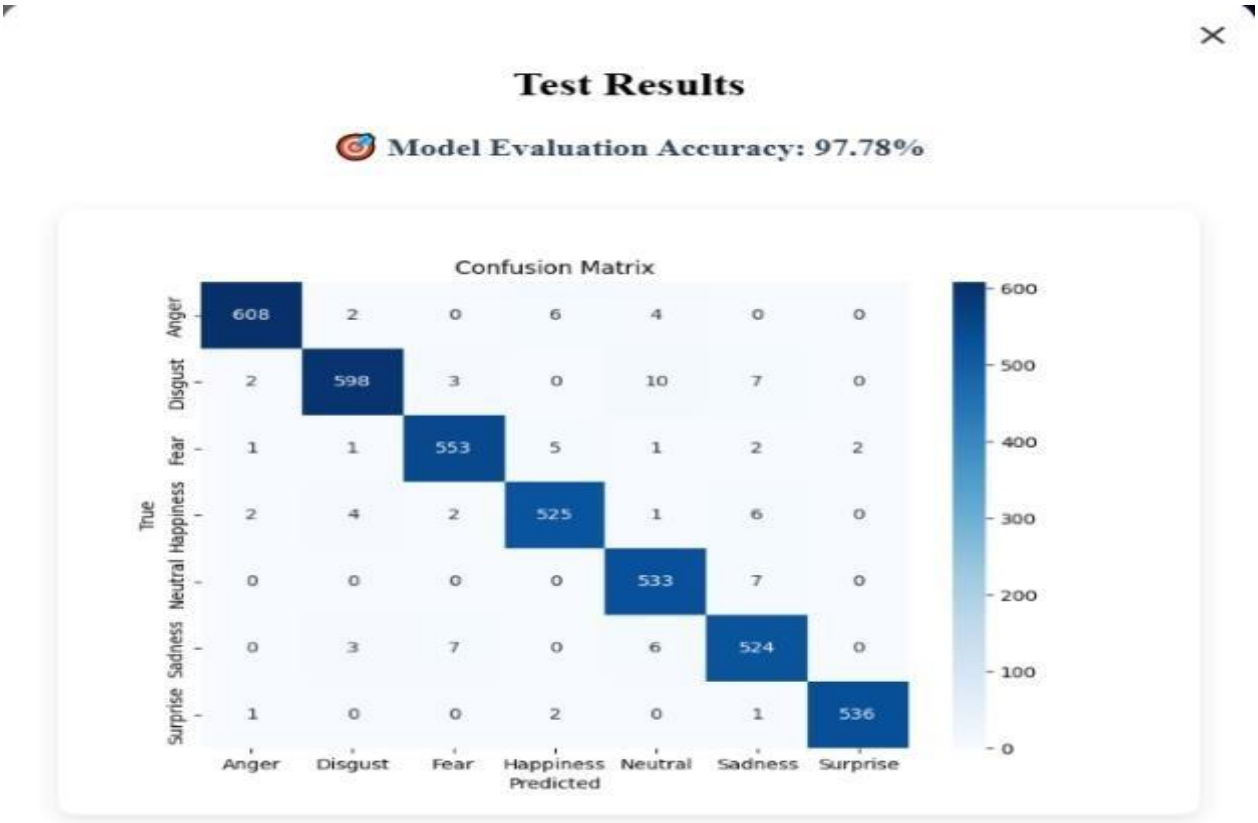


Figure 37: *Testing Results*

2.4 Prediction Interface

Two methods are presented here for making predictions:

2.4.1 Upload Audio Files

The user can upload one or more audio files. A "Predict" button appears after upload.

Once the "Predict" button is clicked, the files are processed by the app and the results are shown in a table with 3 columns:

- File Name
- Predicted Emotion
- True Emotion (if known)

Each row has an icon to view the waveform of the respective audio in a popup window.

There is a "Re-upload Files" button that clears the currently selected files and allows the user to upload a new set of files.

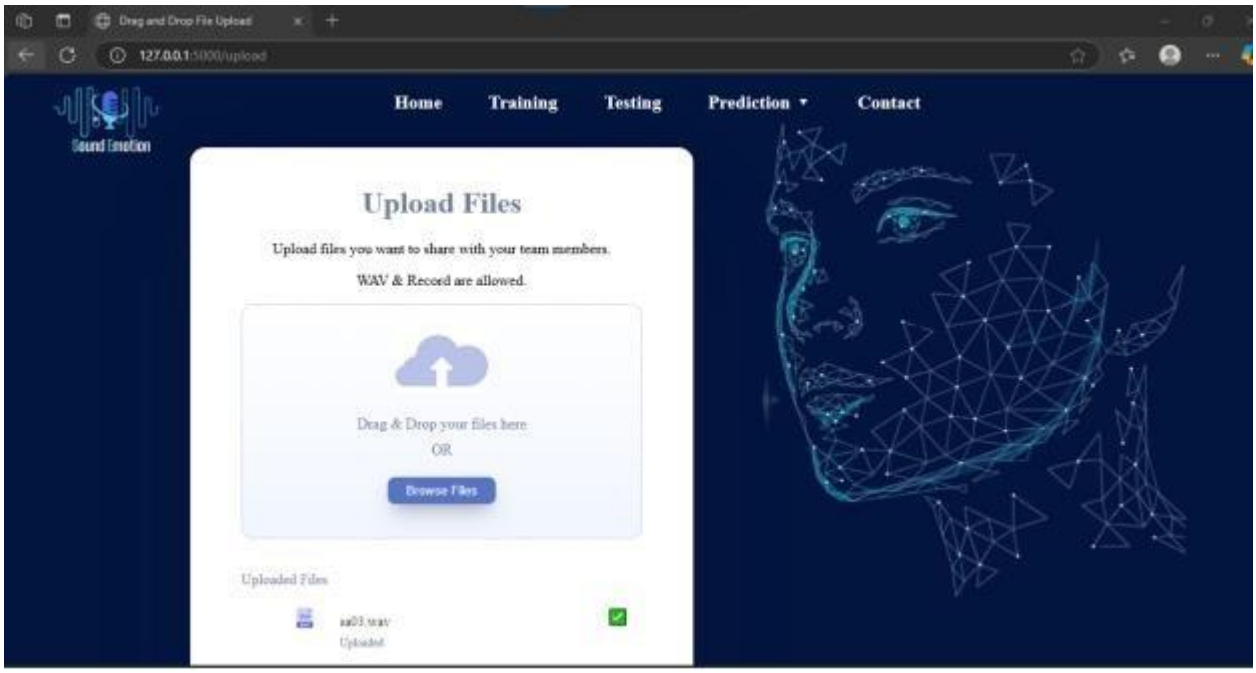


Figure 38: *Audio Upload Prediction*

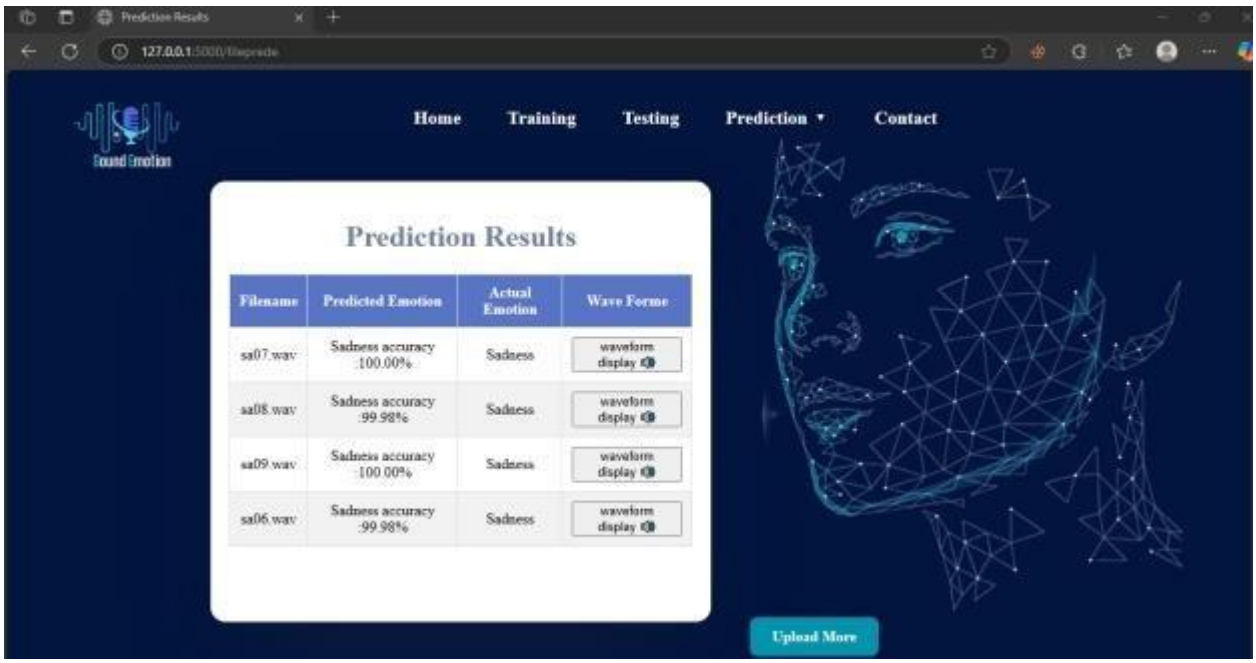


Figure 39: *Prediction results interface*

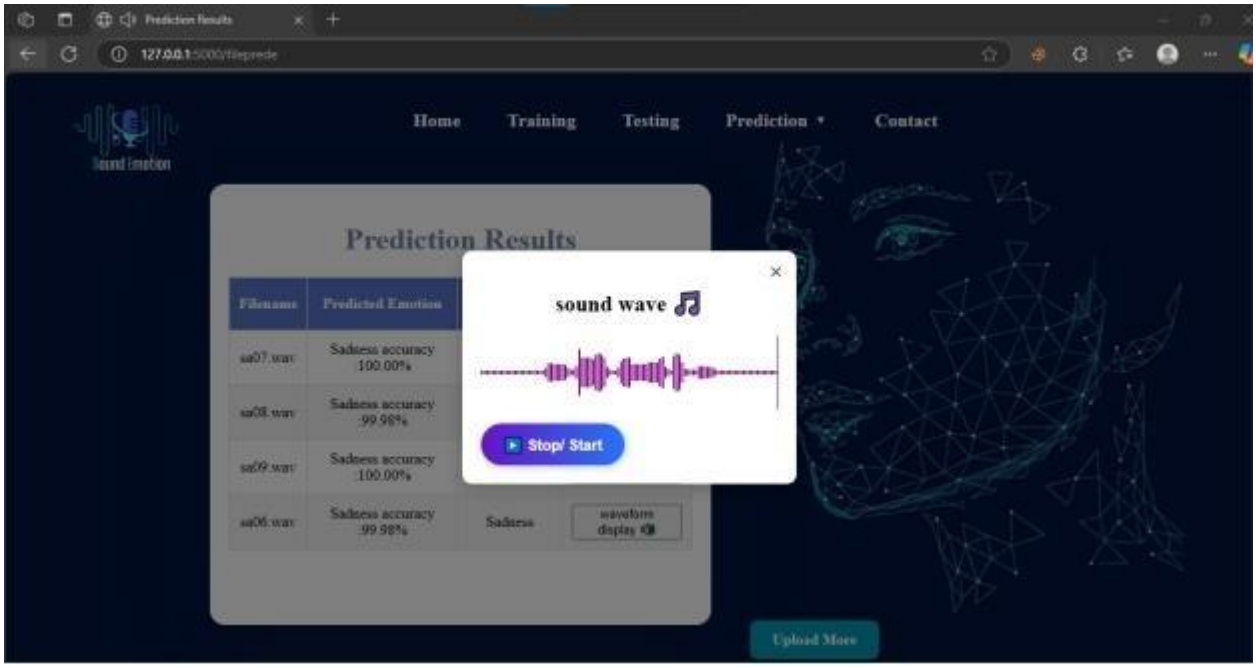


Figure 40: *waves results interface*

2.4.2 Record Audio

- Users record their voice via the microphone.
- The waveform of the audio is shown once recording has finished.
- The "Predict" button is then shown to classify the emotion. Results appear in a table with:
- File Name
- Predicted Emotion
- Each row has an option to delete the entry:

The "Clear All" option allows the user to clear the entire table.

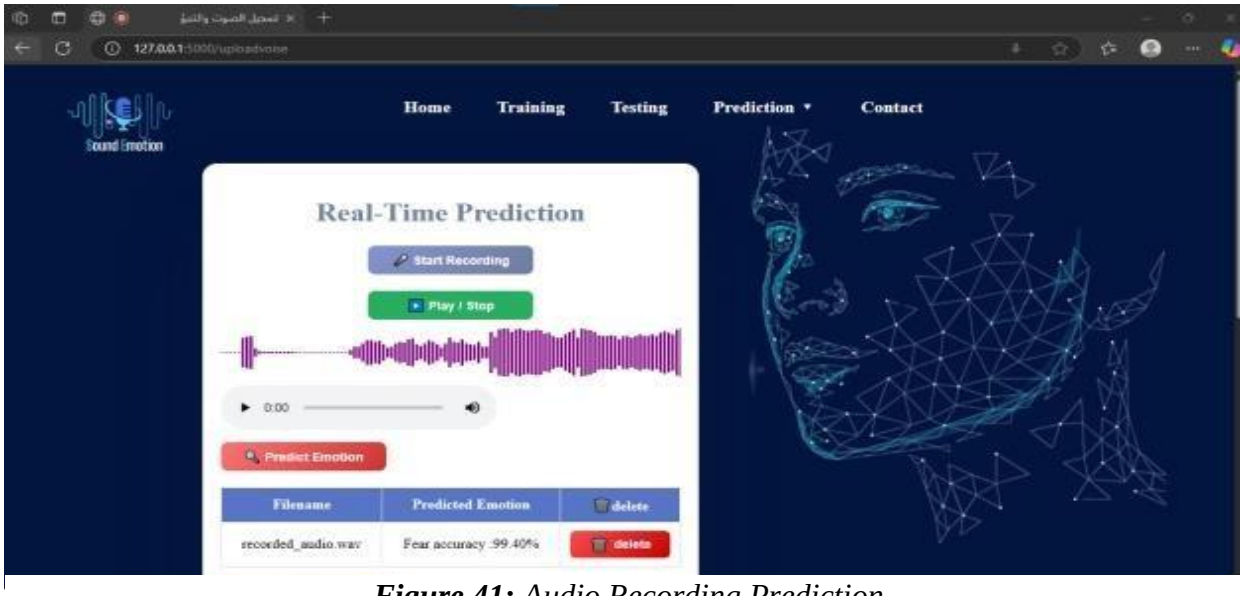


Figure 41: *Audio Recording Prediction*

2.5 Contact Section

When clicking on "Contact" in the navigation menu, the page slides smoothly to the footer section, giving the user access to contact information.

The footer is a modular animated component, which complements the aesthetic of the app.

Figure 42*: Footer*

3 Conclusion

To conclude, the web interfaces of the Sound Emotion Recognition Application were created having usability and technical depth as the basis. Every single page from the home page to training,

testing, prediction, and contact encourages an integrated workflow that supports model development, evaluation, and real-time usage. This system offers a robust and friendly environment for an interactive audio-based emotion analysis due to clear design, interactivity, and feedback mechanisms like progress bars and result popups. These interfaces provide an underlying framework for deploying the machine learning model into web-based real-world environments.

General Conclusion

In summary, Speech Emotion Recognition (SER) is an important and promising field that can contribute to human interaction and in bridging the gap that exists between men and machines. Through this thesis, we aim to conduct a comprehensive study on this subject, beginning with the theory behind emotions, the need for vocal expression, and ending in the actuality of implementing an emotion recognition system.

The results obtained using CNNs to analyze acoustic features evinced that they are quite suitable for and effective in classifying the basic emotions. We created a model that could learn intricate patterns embedded in the audio signals to provide accurate prediction results, hence proving the viability of the considered deep-learning techniques. To give shape to an otherwise abstract theoretical notion and demonstrate its actual practical value, we developed an easy-to-use web interface. The interface also emphasizes the potential that the emotion recognition system may have in real life through its training, testing, and real-time prediction capabilities.

Despite the challenges still holding back this domain, such as limited availability and variability of data, a significant influence of noise, and the inherently complexity of the human emotional expressions, there is hope from the recent improvements in the field of AI machine and deep learning (DL) along with the increasing availability of large datasets for the future of SER.

It is hoped this work will be a significant contribution to the SER knowledge and may be used as the groundwork for future research and systems development in the development of more robust, adaptable, and emotionally intelligent systems. This will enable radically new systems that support human-technology interaction, empower machines to understand and keep pace with human cognition and behavior.

Bibliography

[1] P. Ekman, An argument for basic emotions, Cognition & Emotion, 6(3-4), 169–200, 1992.

[2] C. Darwin, The Expression of the Emotions in Man and Animals, 1872.

[3] C. Ekman et W. V. Friesen, Facial Action Coding System (FACS), 1978.

[4] J. S. Lerner et al, Emotion and decision making, Annual Review of Psychology, 66, 799–823, 2015.

[5] J. J. Gross, The emerging field of emotion regulation: An integrative review, Review of General Psychology, 2(3), 271–299, 1998.

[6] P. Ekman, Emotions Revealed, Henry Holt and Company, 2003.

[7] K. Yao, Challenges in speech emotion recognition: Current trends and future directions, Journal of Signal Processing Systems, 2021.

[8] Z. Yin et J. Zhao, Deep learning for emotion recognition in speech: Review and future directions, IEEE Access, 7, 111347–111364, 2020.

[9] B. Schuller et A. Batliner, Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing, Wiley, 2014.

[10] «Theme parlement europeen,» 07 09 2020. [En ligne]. Available: https://www.europarl.europa.eu/topics/fr/article/20200827STO85804/intelligence-artificielle-definition-et-utilisation.

[11] S. Coursera, «coursera,» 20 05 2025. [En ligne]. Available: https://www.coursera.org/articles/what-is-machine-learning.

[12] «GeeksforGeeks,» 04 04 2025. [En ligne]. Available: https://www.geeksforgeeks.org/artificial-neural-networks-and-its-applications/.

[13] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, O'Reilly Media, 2019.

[14] «GeeksforGeeks,» 14 05 2025. [En ligne]. Available: https://www.geeksforgeeks.org/k-nearest-neighbours/.

[15] Ian Goodfellow, Yoshua Bengio et Aaron Courville, Deep Learning, MIT Press, 2016.

[16] I. Goodfellow, Y. Bengio et A. Courville, Deep Learning, MIT Press, 2016.

[17] R. Lotfian et C. Busso, Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings, Computer Speech & Language, 2019.

[18] F. A. Russo et D. Livingstone, The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), PLoS ONE, 2018.

[19] University of Surrey, SAVEE Database, University of Surrey, UK.

[20] n. predefined, Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D.

[21] University of Toronto, TESS Dataset, University of Toronto.

[22] Institute of Automation, Chinese Academy of Scienc, CASIA Chinese Emotional Corpus, Institute of Automation, Chinese Academy of Sciences.

[23] F. Eyben, M. Wöllmer et B. Schuller, OpenEAR – Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit, In Proceedings of the 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction (ACII), 2009.

[24] P. S. Foundation, Python Language Reference, Python Software Foundation, online.

[25] Microsoft, Visual Studio Code, Microsoft, open source.

[26] T. Kluyver et La, Jupyter Notebooks – a publishing format for reproducible computational workflows,

Proceedings of the 20th International Conference on Electronic Publishing, 2016.

[27] A. Inc, Anaconda Distribution, Anaconda Inc..

[28] M. Grinberg, Flask Web Development: Developing Web Applications with Pytho, O'Reilly Media, 2018.

[29] J. Duckett, HTML and CSS: Design and Build Websites, Wiley, 2011.

[30] C. R. Harris, Array programming with NumPy, Nature, 2020.

[31] W. McKinney, Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference, 2010.

[32] B. McFee, Audio and music signal analysis in Python, Proceedings of the 14th Python in Science Conference, 2015.

[33] J. D. Hunter, A 2D Graphics Environment, Computing in Science & Engineering, 2007.

[34] M. Waskom, Statistical data visualization, Journal of Open Source Software, 2021.

[35] T. Developers, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, Google, 2015.

[36] F. Pedregosa, Machine Learning in Python, Journal of Machine Learning Research (JMLR), 2011.

[37] Python Software Foundation, Python os — Miscellaneous operating system interfaces, Python Software Foundation.

[38] P. S. Foundation, Python sys — System-specific parameters and functions, Python Software Foundation.

[39] F. Pérez et B. E. Granger, IPython: A System for Interactive Scientific Computing, Computing in Science & Engineering, 2007.

[40] Brian McFee, Colin Raffe, Dawen Liang, Daniel P et Ellis, Audio and music signal analysis in Python, Proceedings of the 14th Python in Science Conference (SciPy), 2015.

[41] «CHEAVD: a Chinese natural emotional audio–visual database,» vol. 8, n° %12017, p. 913–924, 2016.

[42] Microsoft.

ABSTRACT

This project addresses the recognition of human emotions through speech using artificial intelligence. A CNN-based model was developed to classify emotions like happiness, anger, and sadness from features such as MFCC. Datasets like RAVDESS, SAVEE, and CASIA were combined to enhance diversity. Data augmentation techniques improved model generalization. The model achieved over 96% accuracy. A Flask-based web interface enables training, testing, and real-time prediction. The system offers promising applications in mental health, education, and customer service. This work lays the groundwork for emotionally aware and intelligent interaction systems.

الملخص

يتناول هذا المشروع موضوع التعرف على المشاعر من خلال الصوت باستخدام تقنيات الذكاء الاصطناعي. يعتمد النظام على شبكة عصبية التفافية (CNN) لتحليل الخصائص الصوتية مثل MFCC وتمييز المشاعر كالفرح، الغضب والحزن. تم استخدام قواعد بيانات متعددة مثل CASIA, SAVEE, RAVDESS , لتحقيق التنوع في البيانات, فاعتمدنا على تقنيات التحسين مثل إضافة الضوضاء، وتغيير السرعة والنعمة لتحسين أداء النموذج.

حقق النموذج دقة تجاوزت 96. كما تم تطوير واجهة ويب تفاعلية باستخدام Flask تمكّن المستخدم من رفع أو تسجيل صوت لتحليل المشاعر في الزمن الحقيقي.

يظهر هذا النظام إمكانيات تطبيقية كبيرة في مجالات مثل الصحة النفسية التعليم، وخدمة العملاء، ويمهد الطريق نحو أنظمة أكثر ذكاء وتفاعلا عاطفيا مع المستخدم

RÉSUMÉ

Ce mémoire traite de la reconnaissance des émotions à partir de la voix en utilisant l’intelligence artificielle. Un modèle basé sur les CNN a été conçu pour classer des émotions comme la joie, la colère ou la tristesse à partir de caractéristiques audio (MFCC). Les bases de données RAVDESS, SAVEE et CASIA ont été fusionnées. Des techniques d’augmentation ont renforcé la robustesse du modèle, qui a atteint une précision de plus de 96 %. Une interface web avec Flask permet l'entraînement, les tests et la prédiction en temps réel. Ce système trouve sa place dans la santé mentale, l'éducation et les services interactifs.