LEBANESE UNIVERSITY

FINAL YEAR PROJECT

Tell Me What You Feel

Authors:
Ismail KHODR KATTAR
Romy BOU ABDO

Supervisor(s):
Dr. Kassem RAMMAL
Mr. Elie DINA

A thesis submitted in fulfillment of the requirements for the degree of BSc. in Data Science

in the

Faculty of Information II

August 28, 2024

Declaration of Authorship

We, Ismail KHODR KATTAR, Romy BOU ABDO, declare that this thesis titled, “Tell Me What You Feel” and the work presented in it are our own. We confirm that:

- This work was done wholly or mainly while in candidature for a research degree at the Lebanese University, Faculty of Information II.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at the Lebanese University or any other institution, this has been clearly stated.
- Where we have consulted the published work of others, this is always clearly attributed.
- Where we have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely our own work.
- We have acknowledged all main sources of help.
- Where the thesis is based on work done by ourselves jointly with others, we have made clear exactly what was done by others and what we have contributed ourselves.

Signed: Romy Bou Abdo - Ismail Khoder Kattar

Date: 20/06/2024

"The mind is not a vessel to be filled but a fire to be kindled"

Plutarch

LEBANESE UNIVERSITY

Abstract

Faculty of Information II

BSc. in Data Science

Tell Me What You Feel

by Ismail KHODR KATTAR, Romy BOU ABDO

Speech Emotion Recognition (SER) is a task of speech processing and computational para-linguistics that aims to recognize and categorize the emotions expressed in spoken language. The goal is to determine a speaker's emotional state, such as happiness, anger, sadness, or frustration, from their speech patterns, such as prosody, pitch, and rhythm. Due to the crucial importance of emotions in daily life, the aim is to contribute to developing an emotion classifier that can be later on integrated into a useful tool to help people understand emotions since some difficulties are faced in this aspect (e.g. autistic people). Two approaches were conducted, Deep Learning (DL) models and Machine Learning (ML) Models. DL models outperformed ML models. In the DL models, a Convolutional Neural Network (CNN) is used since it can capture local information about features. Three models are tested, and each model will have three features from which the features are a combination of time domain and frequency domain. The frequency domain features used are the Mel Frequency Cepstral Coefficient(MFCC), Chroma-STFT (Chroma-Short Time Fourier Transform), and Spectral Centroid (SC). As for the time domain features, Zero Crossing Rate (ZCR), and Root Mean Square Error (RMSE) were used. The first model uses MFCC, ZCR, and RMSE as inputs, the second model uses SC instead of RMSE. As for the third model which outperformed the others uses Chroma-STFT, ZCR, and SC as features. The use of two frequency domain audio features and one time domain audio features performed better than the use of two time domain audio features and one frequency domain audio features. For the other approach, the goal is to implement different ML models and MFCC, an aspect of audio processing that enables efficient and effective extraction of speech features to achieve high accuracy rates. The different ML models used are Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest-Neighbors (KNN), Decision Tree (DT), Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost), and MultiLayer Perceptron (MLP). In, total, 367 Features and 9 classifiers were included in Grid Search with 3 folds of cross validation in this approach. SVM outperformed all the other ML models, and DT was the least successful model. All the proposed models have been evaluated on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) benchmark dataset, and data augmentation has been conducted on this dataset.

Acknowledgements

We are immensely thankful to our esteemed supervisors, Mr. Elie DINA and Dr. Kassem RAMMAL, for their invaluable mentorship and academic oversight. Their expertise and dedication to scholarly excellence have profoundly influenced our work and academic journey.

Our heartfelt appreciation goes to our families and friends, whose unwavering support and encouragement have been our pillars of comfort and motivation. Their faith in our potential has been the driving force that propelled us forward.

We also extend our gratitude to all the professors and colleagues at the university who have contributed to our learning experience and provided a nurturing environment for intellectual and personal growth.

Lastly, we extend our sincere thanks to all the individuals and organizations who have generously contributed data essential for our research. Their willingness to share knowledge and resources has been instrumental in the depth and quality of our study.

We are also grateful to those who have assisted us in discovering and accessing various courses that have enriched our understanding and skill set, necessary for the successful completion of this project. Their guidance has been invaluable.

Contents

List of Figures

List of Tables

List of Abbreviations

AdaBoost	**A**daptive **B**oosting
AI	**A**rtificial **I**ntelligence
ASR	**A**utomatic **S**peech **R**ecognition
Bi-LSTM	**B**idirectional **L**ong **S**hort-**T**erm **M**emory
CNN	**C**onvolutional **N**eural **N**etwork
DCT	**D**iscrete **C**osine **T**ransform
DL	**D**eep **L**earning
DFT	**D**iscrete **F**ourier **T**ransform
DT	**D**ecision **T**ree
EDA	**E**xploratory **D**ata **A**nalysis
ER	**E**motion **R**ecognition
FB	**F**eature **B**lock
FC	**F**ully **C**onnected
FFT	**F**ast **F**ourier **T**ransform
FT	**F**ourier **T**ransform
GDPR	**G**eneral **D**ata **P**rotection **R**egulation
HCI	**H**uman**C**omputer **I**nteraction
HMI	**H**uman-**M**achine **I**nteraction
IAA	**I**nter-**A**nnotator **A**greement
KNN	**K**-**N**earest **N**eighbor
LOG	Simple **Log**istic Regression
log-MFCC	**Log** **M**el-**F**requency **C**epstral **C**oefficient
LPCC	**L**inear **P**rediction **C**epstral **C**oefficient
LR	**L**ogistic **R**egression
MFCC	**M**el **F**requency **C**epstral **C**oefficient
ML	**M**achine **L**earning
MLP	**M**ulti **L**ayer **P**erception
MLR	**M**ultinomial **L**ogistic **R**egression
MTL	**M**ulti **T**ask **L**earning
RAVDESS	**R**yerson **A**udio-**V**isual **D**atabase of **E**motional **S**peech and **S**ong
REIS	**R**ecognition of **E**motion with **I**intensity **S**peech
ReLU	**Re**ctified **L**inear **U**nit
RF	**R**andom **F**orest
RMSE	**R**oot **M**ean **S**quare **E**rror
SC	**S**pectral **C**entroid
SER	**S**peech **E**motion **R**ecognition
SK	**S**pectral **K**urtosis
SMO	**S**equential **M**inimal **O**ptimization
STFT	**S**hort **T**ime **F**ourier **T**ransform
SVM	**S**upport **V**ector **M**achine
TDF	**T**ime **D**istribution **F**latten
TER	**T**ext **E**motion **R**ecognition

VER	**V**ocal **E**motion **R**ecogition
WPT	**W**avelet **P**acket **T**ransform
XGBoost	e**X**treme **G**radient **Boost**
ZCR	**Z**ero **C**rossing **R**ate

List of Symbols

A_0	Accuracy
A_e	Expected Argument
$a(k)$	Amplitude at the kth frame
$Chroma - STFT(t, \theta)$	Chroma-Short Time Fourier Transform at time t and Chroma bin θ
n_q	Total number of judgments for the category q
$n_{q,i}$	Number of judgments for the category q for annotator I
Q	Tags
R	Hop Size
RMS_t	Root Mean Square Error at frame t
SC	Spectram Centroid at frame t
$STFT[m, k]$	Short Time Fourier Transform at time m and frequency k
$s(k)$	Sample Value at k
$w(n)$	Windowing Function
$x(n)$	Signal Segment
ZCR_t	Zero Crossing Rate at frame t
ϕ	Phase

Dedicated to our Families and Friends

Chapter 1

Introduction

1.1 Project Description

Emotions are reactions that human beings experience in response to events or situations (Kendra Cherry, 2023). They differ from one another in several dimensions. For instance, some emotions manifest as immediate occurrences, such as panic, while others are long-term dispositions, like hostility. Emotions also vary in duration; anger might be fleeting, whereas grief can be prolonged. The cognitive processing involved can range from primitive, as in the fear triggered by a sudden threat, to complex, such as the anxiety over failing a test. Conscious emotions like pride when winning a prepared debate differ from unconscious ones, such as an underlying fear of failure. Additionally, motivations to act also vary, with rage often inspiring action, unlike sadness, which may not (Scarantino and Sousa, 2021).

Paul Ekman's notion of basic emotions is a fundamental theory in the study of emotions. Expanding upon Charles Darwin's initial studies on emotional expression, Ekman recognized a range of common emotions that he claimed were essential to both psychological and biological processes. Seven basic emotions were first put out by Ekman: fear, anger, joy, sadness, contempt, disgust, and surprise. Later, he excluded contempt and reduced this list to six items. Carroll Izard and Silvan Tomkins were two psychologists who aided and expanded Ekman in his work. They underlined that these basic emotions are evolutionary adaptations that are essential for social interaction and survival (Gu et al., 2019). Following Ekman's research, Plutchik, 1980 presented a model that classifies emotions into a wheel and shows how the primary emotions (anger, fear, sadness, disgust, surprise, anticipation, trust, and joy) relate to one another. Similar to how colors merge on a color wheel, Plutchik's wheel of emotions - Figure 1.1 illustrates how these fundamental emotions can come together to create complex emotional experiences.

Whilst widely acknowledged, there is continuous discussion on the exact number and nature of basic emotions. For example, Jack, Garrod, and Schyns, 2014 claimed that fear, anger, joy, and sadness are the only four basic emotions experienced by humans based on the overlap of certain emotions and facial expressions. Other researchers share this viewpoint, emphasizing how our understanding of human emotions is evolving.

Nowadays, the recognition of emotions is important for different reasons such as robots and Artificial Intelligence (AI) interaction with humans. For instance, Amazon has been working on making Alexa more emotionally intelligent. The goal is to make Alexa's interactions more empathetic and personalized by analyzing vocal tones, speech patterns, and contextual cues to detect users' emotional states. Thus, the development of Emotion Recognition (ER) as a subfield of AI. Researchers are discovering methods to teach computers to understand and respond according to human emotions. Speech Emotion Recognition (SER) is a blend of psychology and

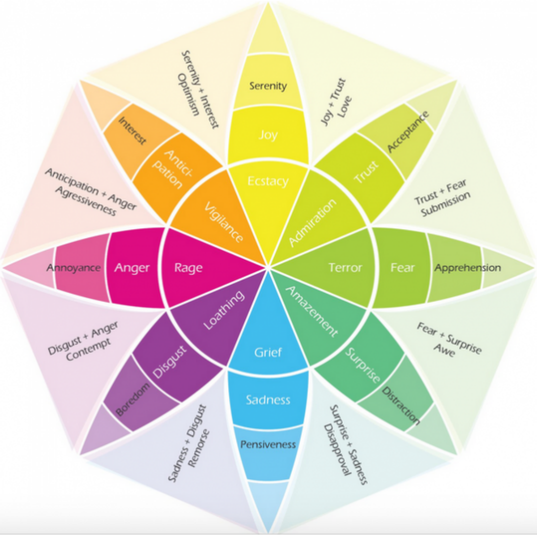


FIGURE 1.1: Plutchik Wheel Of Emotions

technology. It has the potential to revolutionize our world, from acquiring information to analyzing clinical records.

What is most exciting about this, is that those savvy technologies can change lives. For example, for people who struggle to grasp emotions (e.g. autistic people), those appliances can be life-changing tools. Because understanding and navigating social challenges is a real struggle for them.

SER is a task of speech processing and computational para-linguistics that aims to recognize and categorize the emotions expressed in spoken language. The goal is to determine a speaker's emotional state, such as happiness, anger, sadness, or frustration, from their speech patterns, such as prosody, pitch, and rhythm (*Papers with Code - Speech Emotion Recognition — paperswithcode.com* n.d.).

With the advancement of Deep Learning (DL), researchers endeavor to understand emotional portrayal from speech using deep neural networks like Convolutional Neural Networks (CNN). This model analyses various features of speech, such as tone, pitch, rhythm, ... to accurately detect and interpret emotional states.

In this work, two approaches are proposed that can effectively learn speech emotion representations. The first one is based on classifying emotions using CNN since it can capture local information about features. From this approach, three models are suggested, and each model will have three features from which the features are a combination of time domain and frequency domain. The frequency domain features used are the Mel Frequency Cepstral Coefficient(MFCC), Chroma-STFT (Chroma-Short Time Fourier Transform), and Spectral Centroid (SC). As for the time domain features, Zero Crossing Rate (ZCR), and Root Mean Square Error (RMSE) were used. For the other approach, the goal is to implement different ML models and MFCC, an aspect of audio processing that enables efficient and effective extraction of speech features to achieve high accuracy rates. Comparative experiments of these two Integration are conducted to demonstrate the effectiveness of the conducted approaches.

1.2 Ethical Concerns

Critical areas of ethical concern regarding SER deal with: Impact on neuro-divergent people like those with autism spectrum disorders(ASD), who might convey emotions in ways that are pretty different from neurotypical people, so inclusive data collection and models that can be adapted are critical to eschewing bias and exclusion. Moreover, there is a threat that SER tools can be used for emotional manipulation; businesses or individuals can exploit the tools to influence decisions or invade privacy through constant monitoring, for which clear ethical guidelines, user consent, and regulatory measures are called for. It is also essential for compliance with the General Data Protection Regulation (GDPR) because SER systems process personal and sensitive data, for which anonymization techniques and transparent data protection policies are required to ensure user privacy. Therefore, there is a need to justify the use of SER technology—its purpose, the balance of benefits against ethical risks, definition of beneficial use-cases like mental health support and enhanced customer services, and regular ethical reviews for realignment with societal values. These are the considerations that are very important for the development and deployment of SER technology responsibly and ethically.

Therefore, a Google form was made where the scope of the work, the purpose of the data collected, whom it will be shared with, and how it will be used, was stated and the freedom of staying anonymous was given. As stated, the introduction of the form regulates with Article 5(1)(b-c) and Articles 13-14 (1)(a-b-c-d-e) data minimization, purpose limitation, and data subject rights. Articles 5(1)(a) and 6 implies lawfulness, fairness, and transparency which were also followed. And Articles 7 and 8 imply the consent part which is applied by making people give their consent by signing a paper that states their approval. Also, if a person is underage their parents need to approve *General Data Protection Regulation* n.d. The form's purpose is to collect audio from people to test them in this project. Only age is mandatory. The user can choose to record in any language specified (Lebanese, French, English), any of the 8 emotions stated, and whether they prefer to improvise the script or be given a written script.

Chapter 2

State of the Art

The work of Schuller, Rigoll, and Lang, 2004 will be remembered in the area of speech emotion recognition for initiating some innovative approaches toward fusion in acoustic and linguistic information for robustness in recognizing emotions from speech. Proper extraction of features with their analysis led to the elucidation of critical acoustic features that can estimate emotion. Still, it simultaneously raised a new concept: approving the linguistic content for extraction of cues about emotions. This is further supported by comparisons with another classification method, SVM, which proves its superiority in dealing with the tasks of emotion recognition complexity. Also, their integration through the fusion of acoustic and linguistic information has been considerably improved in performance through the soft decision approach. It reveals, therefore, that integrated systems such as these hold high potential for enabling better Human Computer Interaction(HCI), especially inside a car concerning safety and user experience, where knowledge of the driver's emotional state plays a vital role.

Besides this, Gao, Chu, and Kawahara, 2023 propose another novel two-stage fine-tuning approach that seeks to increase the embedding enhancement of wav2vec 2.0 through pre-training with Automatic Speech Recognition(ASR) and gender recognition for SER purposes. The methods mentioned above existed at two stages, outlining the pre-training attempt at the SSL model over ASR to handle the gradient conflict problem of multitask learning and then re-training for SER. Experimental results using the IEMOCAP database have shown that, about a naive multi-task learning implementation setup, ASR pretraining brings vast improvement in terms of SER performance. It further testifies to including gender recognition, offering more effective embeddings for SER based on existing end-to-end systems, with a tremendous absolute accuracy rate of improvement. However, through meticulous experimentation and comparisons, this paper stands in an excellent position to establish the effectiveness of the proposed two-stage fine-tuning methodology in terms of SER performance based on wav2vec 2.0 embeddings by outperforming previously running state-of-the-art approaches.

Bhangale and Kothandaraman, 2023 present various speech emotion recognition (SER) techniques in their study. They highlight the utilization of Mel Frequency Cepstral Coefficients (MFCC) spectrogram or coefficients as inputs to deep learning frameworks, while addressing their limitations in capturing lower-order characteristics and precise arousal and valence levels. The researchers advocate for improved distinctiveness of speech features, considering alternative approaches such as acoustic features. These include Mel frequency cepstral coefficients (MFCC), linear prediction cepstral coefficients (LPCC), wavelet packet transform (WPT), zero crossing rate (ZCR), spectrum centroid, spectral roll-off, root mean square error (RMSE), spectral kurtosis (SK), jitter, shimmer, pitch frequency, and formats.

Their comprehensive analysis reveals that these features offer valuable insights into various aspects of speech signal processing. For instance, MFCC computation involves pre-emphasis, frame alienation, Discrete Fourier Transform (DFT), Mel filter banks, and Discrete Cosine Transform (DCT), resulting in 39 features crucial for characterizing emotional transitions. Additionally, features like RMSE, ZCR, spectrum centroid, spectral roll-off, LPCC, spectral kurtosis, jitter, shimmer, pitch frequency, and formants provide essential information on loudness, signal transitions, spectral shape, and vocal characteristics, thereby enhancing SER accuracy. The study further demonstrates the efficacy of employing a compact 1-D Deep Convolutional Neural Network (DCNN) with 715 acoustic features, which significantly improves SER accuracy compared to raw speech and MFCC-based approaches across various emotion datasets like EMODB and RAVDESS. This research holds promise for real-time applications due to its lower computational complexity and reduced training time.

Another study was conducted by Dogdu et al., 2022 to show an analysis of ML algorithms for Vocal Emotion Recognition (VER) using four distinct feature sets on the Berlin Database of Emotional Speech (EMO-DB). Seven ML algorithms were evaluated: Multilayer Perceptron Neural Network (MLP), J48 Decision Tree (DT), Support Vector Machine with Sequential Minimal Optimization (SMO), Random Forest (RF), k-Nearest Neighbor (KNN), Simple Logistic Regression (LOG), and Multinomial Logistic Regression (MLR). The feature sets utilized were emobase, IS-09, GeMAPS, and eGeMAPS. Results demonstrated that SMO, MLP, and LOG exhibited superior performance, achieving accuracies of 87.85%, 84.00%, and 83.74%, respectively, while RF, DT, MLR, and KNN showed lower accuracies ranging from 73.46% to 58.69%. Notably, the emobase feature set outperformed others, indicating its efficacy in VER applications. These findings contribute to understanding the optimal ML algorithms and feature sets for VER, with implications for various fields such as clinical diagnosis, intervention, and human-computer interaction.

Lalitha et al., 2015 investigate the challenge of quantitatively predicting human emotions, particularly as facial expressions become less reliable indicators with age and experience. It highlights the limitations of traditional methods in recognizing nuanced emotions and proposes Speech Emotion Recognition (SER) as a solution, leveraging paralinguistic cues in speech. The research focuses on enhancing Human-Machine Interaction (HMI) through efficient SER, using Mel Frequency Cepstral Coefficients (MFCC) and Cepstrum for feature extraction and a Neural Network Classifier for emotion identification. Using the Berlin Emotional database, the study achieves an 85.7% accuracy in recognizing seven emotions, surpassing previous methods by 20%, and discusses avenues for future research to further improve recognition efficiency by considering a broader range of features and datasets.

Islam et al., 2022a's work focuses on identifying the extreme level of emotion (e.g., very sad or very angry) using emotional intensity (e.g., Normal, Strong) which is neglected in other (DL)-based SER models has a crucial influence on social activities. Once the extreme level of emotion is detected, an alarm notification may be forwarded to the supportive unit about the person to take necessary action(s). The automatic filtering of speech signals by Recognition of Emotion with Intensity from Speech REIS is the most challenging task addressed in this study.

The general architecture of the proposed REIS approach includes processing signals with MFCC, STFT, and Chroma-STFT. Those 2D transformed features are integrated into a 3D transformed feature. From here, two DL-based frameworks are conducted to find the most effective one for REIS, a single DL framework for classifying emotions with intensity together, and a cascaded DL framework that classifies in the first stage emotions from the 3D transformed feature and classifies in the second stage emotional intensity for the classified emotions from stage one. The cascaded

model showed the best results and outperformed other existing methods.
Single DL framework (for classification of emotion with intensity together). A proposed 3D CNN Bidirectional Long Short-Term Memory (Bi-LSTM), architecture for the simple and cascaded model uses four Feature Blocks (FB) containing each 3D Convolution, Batch Normalization, activation (RELU), and 3D Max-Pooling layers. The output of the convolved FB is put into a Time Distribution Flatten (TDF) layer which flattens the FB. Bi-LSTM takes the flattened FB as input and its outcome is fed into a Fully Connected (FC) layer which recognises emotion or intensity level. This final layer is fed to the Softmax layer, which computes output probabilities for all the classes. The complete 3D CNN model is trained to minimize the categorical cross-entropy loss function. 3D CNN is used for its suitability for feature extraction and Bi-LSTM for its capability to handle sequential data.
Speech audio RAVDESS dataset was used with 80:20 ratio training and testing sets. 4-fold cross-validation accuracy (average of four individual folds) for the single DL and cascaded DL frameworks are 76.66% and 85.34%, respectively.

Researchers, including Ekman, 1984, have proposed a fundamental model of emotions, suggesting that there exists a range of common emotions that he claimed were essential to both psychological and biological processes.These primary emotions include happiness, sadness, anger, fear, disgust, and surprise. From these core emotions, all other emotions are derived. According to this model, each basic emotion is linked to distinct facial expressions, physiological responses, and behavioral patterns.

Furthermore, this idea is further developed by Plutchik, 1980 wheel of emotions, which organises emotions into a circular model and illustrates how various feelings can interact to create more complex emotional experiences.

The table 2.1 presents a summary of some of the previous work done, mentioned above.

Approaches	Accuracy
Schuller, Rigoll, and Lang, 2004	92.95%
Gao, Chu, and Kawahara, 2023	76.10%
Bhangale and Kothandaraman, 2023	94.18%
Dogdu et al., 2022	87.85%,
Lalitha et al., 2015	85.70%
Islam et al., 2022a - Cascaded DL Framework	87.71%
Islam et al., 2022a - Stage 1: Emotions	90.06%

TABLE 2.1: Result of Earlier Approaches

Chapter 3

Data Description

3.1 Dataset Acquired

In this project, audio data was sourced from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). RAVDESS pertains to widely used datasets in research on SER. It is a comprehensive dataset with a total of 7,356 files. These files are divided into 3 main categories and each category contains a modality of speech or song:

- audio files
- video files
- full-AV

For this project, the main interest is speech audio files. The speakers consist of 24 professional actors, 12 males and 12 females. Two lexically matched statements were vocalized by each actor of the given statement in a neutral North American accent repeated twice. Every utterance is articulated with one of seven emotional feelings defined by Ekman, 1984 — neutral, calm, happy, sad, angry, fearful, surprise, and disgust— and is produced at two levels of emotional intensity: normal and strong, except for neutral which contains only normal. Thus, having for each emotion (= 2 statements × 2 repetitions × 2 emotional intensity × 24 actors) 168 samples except for neutral, by having (=2 statements × 2 repetitions × 1 emotional intensity × 24 actors) 96 samples. Such large variability of emotions allows for better generalization of models in emotion recognition in speech. Recorded utterances are intensity-matched and phonetically and acoustically balanced to preserve data integrity across the dataset. The RAVDESS is publicly available and validated in efficacy related to emotion recognition tasks (Livingstone SR, 2018).

3.2 Pre-Processing

In this project, the audio was organized like a directory tree, set up so that 24 folders existed, each representing an actor. Inside the folder of each actor were audio files, i.e. (`\Data Sets\Ravdess\audio_speech_actors_01-24 \Actor_01\03-01-01-01-01-02-01.wav`).

In the RAVDESS dataset, the filename follows a particular encoding format in which each part of the file path gives us information about different metadata. The filename structure includes:

- Modality (01 = full-AV, 02 = video-only, 03 = audio-only)

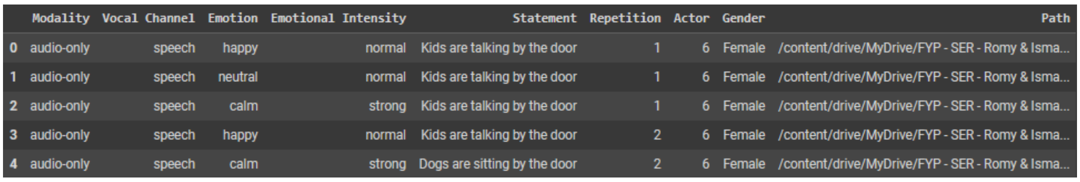
- Vocal Number (01 = speech, 02 = song)]
- Emotion (01: Neutral, 02: calm, 03: happy, 04: sad, 05: angry, 06: fearful, 07: disgust, 08: surprised)
- Intensity of One's Emotion (01 = Normal, 02 = Strong, and notice the neutral emotion, there is no strong intensity)
- Expression (01 = "Children are talking by the door", 02 = "Dogs are sitting by the door")
- Repetition (01 = 1st repetition, 02 = 2nd repetition)
- Actor (01 to 24; male for odd numbers and female for even numbers)

For example, the file 03-01-06-01-02-01-12.wav can be decoded as follows:

- Modality: Audio-Only (03)
- Vocal Channel: Speech (01)
- Emotion: Fearful (06)
- Intensity: Normal (01)
- Statement: "Dogs are sitting by the door" (02)
- Repetition: 1st Repetition (01)
- Actor: 12 (Female, as the actor ID number is even)

A processing techniques were applied to transform the audio files into a structured data frame. It first walks the directory structure in each folder for actors, reads the filenames, extracts the encoded metadata, and then constructs a data frame with the following fields: Modality, Vocal Channel, Emotion, Emotional Intensity, Statement, Repetition, Actor, Gender, Path.

The first section of the code enumerates all the folders; in other words, every folder is a respective actor. In each folder, we continue enumerating files. To extract metadata components from the filename for every file, a split operation using a delimiter '-' is performed. We label the gender for both males and females according to whether the ID is odd or even. Later, file metadata is compiled with its path in a list and appended to the dataset list. Finally, the list of data sets is converted into a data frame with the desired column names as shown in Figure 3.1. This important preprocessing step organizes and accurately labels the dataset so that practical analysis tasks can be successfully run.



	Modality	Vocal Channel	Emotion	Emotional Intensity	Statement	Repetition	Actor	Gender	Path
0	audio-only	speech	happy	normal	Kids are talking by the door	1	6	Female	/content/drive/MyDrive/FYP - SER - Romy & Isma...
1	audio-only	speech	neutral	normal	Kids are talking by the door	1	6	Female	/content/drive/MyDrive/FYP - SER - Romy & Isma...
2	audio-only	speech	calm	strong	Kids are talking by the door	1	6	Female	/content/drive/MyDrive/FYP - SER - Romy & Isma...
3	audio-only	speech	happy	normal	Kids are talking by the door	2	6	Female	/content/drive/MyDrive/FYP - SER - Romy & Isma...
4	audio-only	speech	calm	strong	Dogs are sitting by the door	2	6	Female	/content/drive/MyDrive/FYP - SER - Romy & Isma...

FIGURE 3.1: Data Set

The distribution of different features across audios was studied, and it shows no sign of bias. The emotions are distributed equally among each other, and emotional

intensity is also distributed equally among each emotion - Figure 3.2. Statements (statement 1 and statement 2 - Figure 3.3) and repetitions (1 and 2 - Figure 3.4) are as well distributed equally among emotions.
Neutral doesn't contain a strong intensity, which gives us half the number of other emotions.

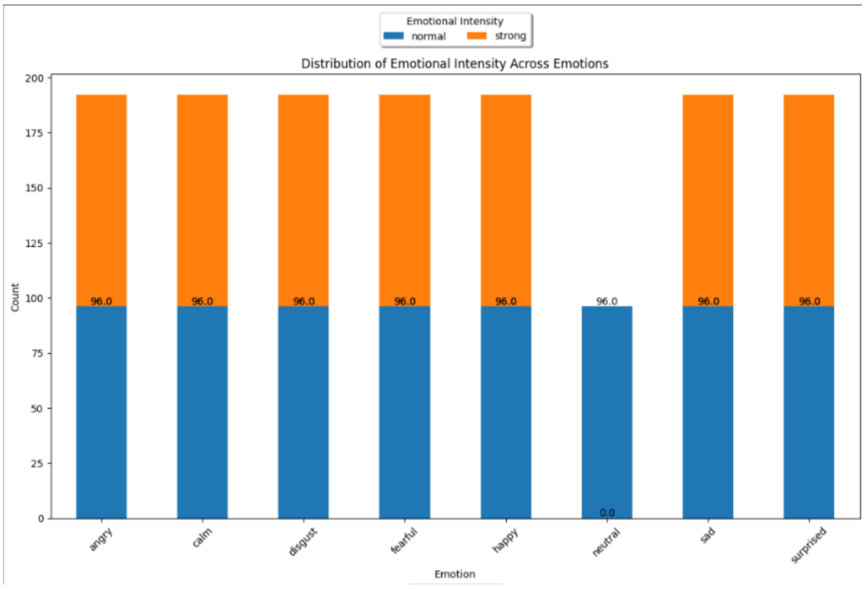


FIGURE 3.2: Emotion Distribution Across Intensity

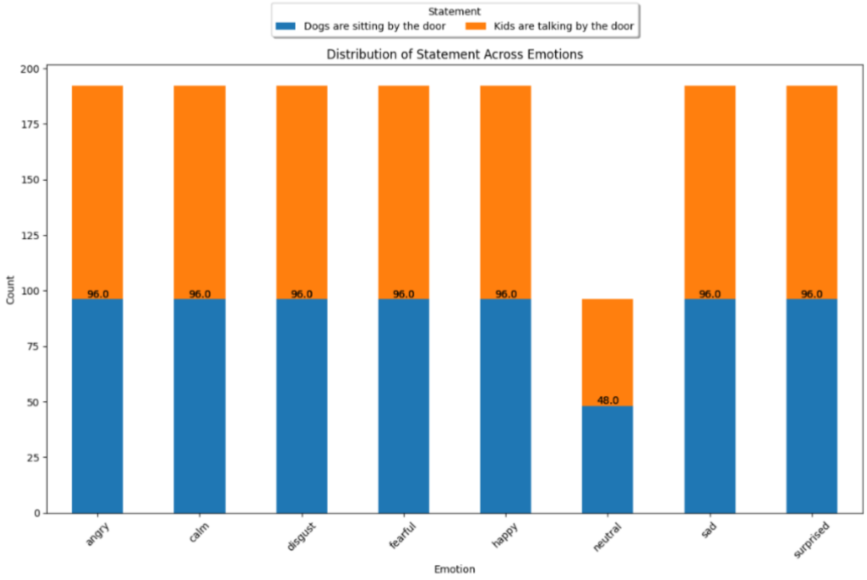


FIGURE 3.3: Distribution of Statement across emotion

Since the dataset size is relatively small (1440 audio samples) data augmentation techniques were required to increase the model's performance. Data augmentation is defined as artificially enlarging the dataset by transforming the existing examples in diverse ways; in effect, it provides the model with a diversified set of training instances. Such methods help not only to counter overfitting but also to improve the generalization capabilities of the model on unseen data.

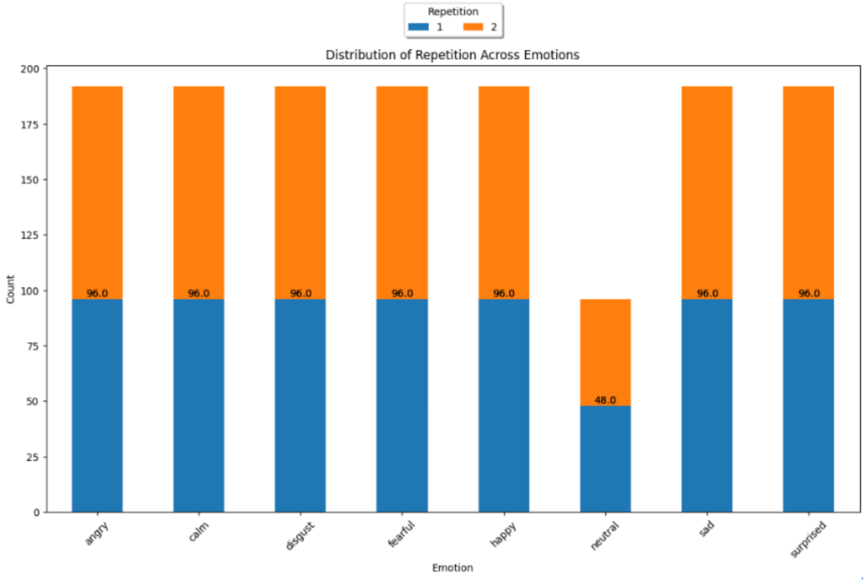


FIGURE 3.4: Repetition Distribution

3.2.1 Implementation of data augmentation:

Data augmentation uses pre-existing data to create new data samples that can improve model optimization and generalizability. It's used to improve accuracy and reduce overfitting by creating new instances from the current data.

1. **Noise Injection**: Random Gaussian noise were added to the audio signals. In that way, the model became more robust and resistant to noise while bringing variability closer to real-world environmental circumstances. Noise was injected into the audio signal y by adding random values drawn from a Gaussian distribution scaled by a factor of 0.005:

   $$y_{\text{aug}} = y + 0.005 \cdot \mathcal{N}(0, 1)$$

   Adding a small amount of noise ensures that the original characteristics of the audio signal are preserved while introducing sufficient variability to improve the model's robustness. As shown in Figure 3.5, the noise-injected signal maintains the overall shape of the original signal but includes minor perturbations.

2. **Time Stretching**: This variation can be achieved by augmenting audio signals. This is because time stretching is a change in tempo without a corresponding change in pitch; such augmentation introduces variations in the temporal domain through which the model can learn different temporal patterns.

   $$y_{\text{aug}}(t) = y\left(\frac{t}{0.8}\right)$$

   Time stretching was achieved by applying a stretch factor of 0.8, which slows down the audio signal. This adjustment helps the model recognize audio features that may vary in duration, such as prolonged sounds or speech patterns, thus enhancing its ability to generalize across different speaking rates and tempos. As shown in Figure 3.5, the time-stretched signal is longer in duration compared to the original, yet retains the same pitch.

3. **Pitch Shifting**: The pitches of audio signals were modulated by changing the pitch. This conversion changes the frequency content of the audio without changing its structure in time and, therefore, provides the model with stronger resilience to pitch modification.Pitch shifting was implemented by altering the pitch of the audio signal by 4 semitones.

$$y_{\text{aug}}(t) = y(t) \cdot 2^{\frac{4}{12}}$$

By shifting the pitch, the model is exposed to a range of frequency variations, enabling it to better handle different speakers, instruments, or tonal qualities in the audio data. As shown in Figure 3.5, the pitch-shifted signal exhibits a noticeable change in frequency while maintaining the same temporal structure.

Data augmentation techniques augment the dataset, increasing the dataset size to more than 5000 samples after data augmentation. This increment in dataset size provides more prosperous and more diverse training examples, thus enhancing the model's generalization ability, that results in better performance.

3.2.2 Feature Extraction

Chroma-STFT

Chroma features represent the twelve distinct pitch classes (C, C#, D, D#, E, F, F#, G, G#, A, A#, and B) in the Western musical scale in music analysis. By focusing just on whether a specific class is present within a given time period and disregarding the precise pitch within each class, the chroma characteristic offers a means of characterizing the harmonic content of the music.

Chroma-STFT combines Chroma features and STFT by first obtaining a spectrogram from STFT (which represents how the frequency content of the signal evolves over time). Then, extracting chroma feature from each time frame of the spectrogram.

In other words, Chroma-STFT is a representation of audio signals whose main focus is on the harmonic content and pitch of the sound rather than the exact time-domain waveform.

The Chroma-STFT mathematical equation for the STFT magnitude spectrum at time index t and frequency bin n $X[n, t]$ is defined as:

$$\text{Chroma-STFT}(t, \theta) = \sum_{n=0}^{N-1} X[n, t] \cdot w[n] \cdot \delta(\theta - \text{bin}(f[n]))$$

where θ represents the Chroma bin index, N the frame size, $\delta(.)$ the Dirac Delta function, and $bin(f[n])$ the function mapping n to the corresponding chroma bin index.

Mel-Frequency Cepstral Coefficient (MFCC)

In this project , the focus is on acoustic features especially MFCC's since RAVDESS does not contain any linguistic information.

Acoustics features of the speech signal represent the physical properties of the speech signal in terms of frequency, amplitude, and loudness (Bhangale and Kothandaraman, 2023).

MFCC provides the spectral information Of the speech and characterizes the human hearing perception (Bhangale and Kothandaraman, 2023)

Pre-emphasis is the first process during the extraction of MFCC coefficients, which further can normalize the raw speech signal and, hence, reduce noise and disturbances in the emotional speech input, for an audio signal lasting 2.6 seconds and sampled at some rate. In this application, each frame is applied with windowing using a Hamming window based on the frame representation of the closest frequency components. The time-domain speech signal is thus transformed by Fourier into its frequency-domain equivalent. Mainly, this transformation brings out the characteristics of the vocal tract in the speech. The signal is then filtered through filter banks, which provide perceptual information related to the human perception of frequency, denoted as Mel Frequency triangular filter banks. It effectively maps linear frequency into the scale of Mel, thereby focusing on frequencies that are more identifiable by the human ear. These log-filter bank energies are then transformed with a discrete cosine transform into a set of cepstral coefficients. For our specific configuration, we extract thirteen MFCC coefficients from the frame, making 13 features in total that capture representative characteristics of the speech signal. The process of MFCC is represented in Figure 3.6

Root Mean Square Energy (RMSE)

RMSE is a metric used in Audio segmentation to identify new segments or events. It is the square root of the mean sum of energy for all samples in frame t.

It is also an indicator of loudness, obtaining higher RMSE values correlates with the audio signal containing more noise or being more complex.

Its mathematical equation is represented by the following with $a(k)$ being the amplitude at the kth sample and $a(k)^2$ the energy:

$$\text{RMS}_t = \sqrt{\frac{1}{k} \sum_{k=tk}^{(t+1)k-1} a(k)^2}$$

where k is the frame size and t index of a frame.

Short Time Fourier Transform (STFT)

In order to understand STFT, the Fourier Transform (FT) process needs to be understood. FT decomposes a complex sound into its frequency components. In other words, it transforms time domain into the frequency domain for which each frequency contains a magnitude and a phase (ϕ). High magnitude indicates a high similarity between the signal and a sinusoid. The steps of the FT will be explained for better understanding.

A frequency is chosen by creating a sine wave out of it, then a phase is optimised in a way that gives maximum similarity between the chosen sine wave and the original signal. The magnitude is then calculated, and the process is repeated for each frequency. Since signals in machines are discrete and the FT is continuous, Discrete Fourier Transform (DFT) is applied to use the FT in a discrete manner. The Fast Fourier Transform (FFT) is an algorithm that efficiently computes the DFT. FT faces a slight problem, time is not taken into consideration. That is why STFT is going to be used in this model.

STFT is the application of FFT in small segments of the signal which is called a frame. Frame represents time. A windowing function is also applied to reduce spectral leakage. Mathematically, applying a windowing function $w(n)$ to a signal segment $x(n)$ involves element-wise multiplication:

$$x_w(n) = x(n) \cdot w(n)$$

where $x_w(n)$ is the windowed signal.

By applying a windowing function, a more accurate frequency domain representation of the signal is obtained, particularly when dealing with non-periodic or finite-length signals.
Hamming window is conducted in this project, and STFT is applied with a frame size of 2048 and a hop size of 512. The hop size determines the step size between successive frames or windows. This choice of frame size and hop size maintains a balance between the frequency and time resolution. This choice of frame and hop size will be used for all the following features.

The STFT of a signal for one frame $x[n + mR]$ for a signal of length N, at time index m and frequency index k is defined as:

$$STFT[m,k] = \sum_{n=0}^{N-1} x[n + mR] \cdot w[n] \cdot e^{-I\frac{2\pi}{N}nk}$$

where R represents the hop size.

Spectral Centroid (SC)

One of the key frequency-domain audio features is SC. It is presented as the center of gravity of the magnitude spectrum and indicates the frequency band where most of the energy is concentrated.

SC also measures the "brightness" of a sound, how dull/sharp a sound is. Let us take the changes in emotional state, it can affect the spectral properties of speech. For example, excited speech often has a higher SC compared to calm speech due to increased high-frequency energy.

The mathematical formula of SC at frame t is the weighted mean of the frequencies, given by:

$$\mathrm{SC}_t = \frac{\sum_{n=1}^{N} m_t(n) \cdot n}{\sum_{n=1}^{N} m_t(n)}$$

where n represents the frequency bins, and $m_t(n)$ the weight for n.

Zero Crossing Rate (ZCR)

ZCR is a metric representing the number of times a signal crosses the horizontal axis. It helps recognise percussive from pitched sounds. Percussive sounds have random ZCR patterns and pitched sounds contain a more stable ZCR due to their periodic nature.

ZCR is indirectly related to pitch, the higher the ZCR the higher the pitch. Also, if the audio is unvoiced (consonants or noise) it will contain a lower ZCR, while voiced audio (speech or singing) will obtain a higher ZCR due to the periodicity of vocal fold vibrations.

Its mathematical equation at frame t is presented by:

$$ZCR_t = \frac{1}{2} \sum_{k=tk}^{(t+1)k-1} |\text{sgn}(s(k)) - \text{sgn}(s(k+1))|$$

where k represents the index of the sample within the frame, $s(k)$ the sample value at index k, and sgn the sign function, which extracts the sign of a signal value $s(k)$ and returns +1 if $s(k) > 0$, 0 if $s(k) = 0$, and -1 if $s(k) < 0$.

3.3 Exploratory Data Analysis

A DataFrame was created in the Exploratory Data Analysis (EDA) phase to contain several attributions for each audio file: the modality, vocal channel, emotion, emotional intensity, statement, repetition, actor, gender, and file path as shown in figure 3.1. The maps presented the dictionaries used for transforming numerical codes to their initial meaning in diversified attributes, such as modality, vocal channel, emotion, emotional intensity, statement, and so on. Analysis of the emotion distribution test showed eight different emotions(Neutral, Happy, Calm, Sad, Surprised, Fearful, Angry and disgust), each of which - 192 audio files (96 of an average intensity level and 96 of a strong intensity level), except the emotion Neutral, which contained 96 audio files of a normal intensity only.

It has also been found that all the audio files in the dataset are of the "audio only" modality. The vocal channel is always labeled as "speech" with 2 different statements "kids are talking by the door" and "dogs are sitting by the door" with 2 repetitions."

The statistics for audio length were calculated and are displayed in Table 3.1. For instance the shortest audio length is of 2.94 sec and most of the audios ranges between 3.47 and 3.87 sec.

Stats	Mean	Standard Deviation	Minimum	Q1	Q2	Q3	Maximum
Values	3.70	0.37	2.94	3.47	3.67	3.87	5.27

TABLE 3.1: Audio statistics

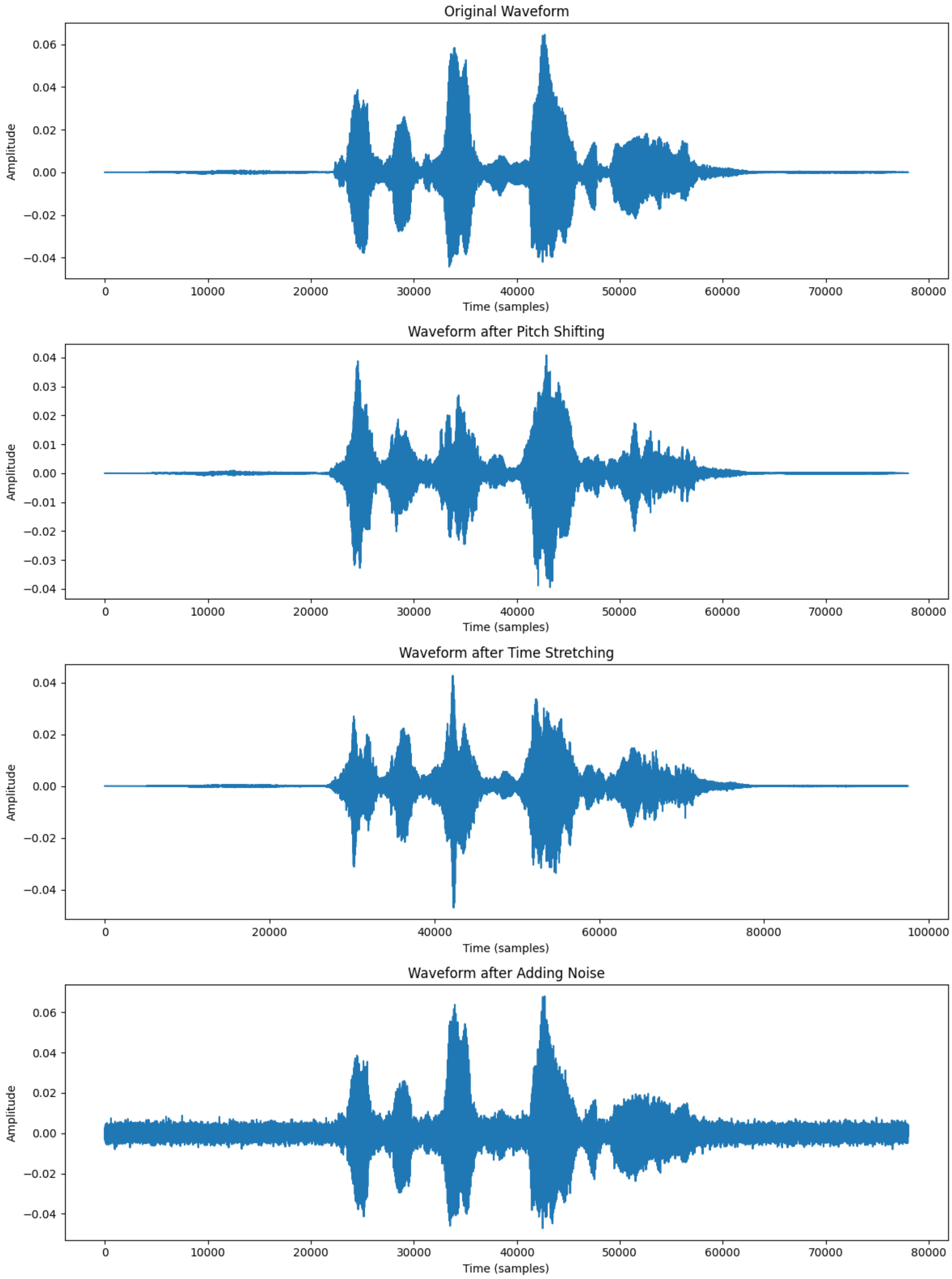


FIGURE 3.5: Augmentation Techniques

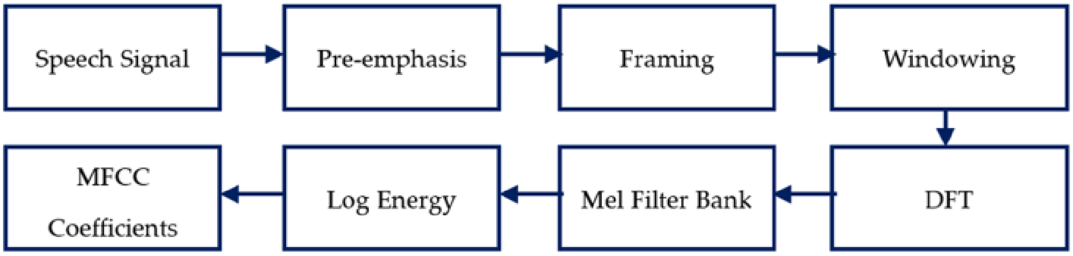


FIGURE 3.6: MFCC Work Flow

Chapter 4

Process Followed

4.1 Inter-Annotator Agreement (IAA)

IAA evaluates the consistency with which several annotators can annotate a given label category or class. It also perceives the quality of the annotation of a corpus.

In linguistic matters, correctness cannot be measured. There is no "ground truth" regarding the manual annotation, since it is based on people's judgment. But instead, the reliability of the annotation is measured (if the annotators agree on the annotations). To ensure reliability, each item is annotated by 2 different annotators (Annotator 1 and Annotator 2), which will be compared amongst each other and the Gold Data (assigned annotation) as reliability ensures correctness.

Some agreements may be accidental. To eliminate this factor, IAA calculations will be conducted to determine the "proportion" of agreement above chance. The S coefficient, π coefficient, and κ coefficient are metrics that will be calculated (showing immediately the results) and defined below.

126 audios were randomly chosen and annotated regarding the 8 emotions (Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, and Surprised), 2 emotional intensities (normal, strong), and their combination of 15 tags which will represent 9(=3x3) heat maps shown in Figure 4.1.

A_0 will present the Accuracy of each heat map. They will be executed each by the 3 coefficients mentioned.

The Expected Argument (A_e) which represents the expected value of the observed agreement is executed for the 3 coefficients:

- **The S Coefficient** :
  Give the same probability to all tags.
  With |Q| tags, what is the probability of both annotators picking the same tag?

$$A_e = \frac{1}{|Q|}$$

- **The π Coefficient** :
  Give different chances for different categories.

$$A_e = \sum_{q \in Q} \left(\frac{n_q}{2N} \right)^2$$

having **Q** as a set of possible categories, **N** as Number of items to annotate and n_q as Total number of judgments for the category q ($0 \leq n_q \leq 2N$).

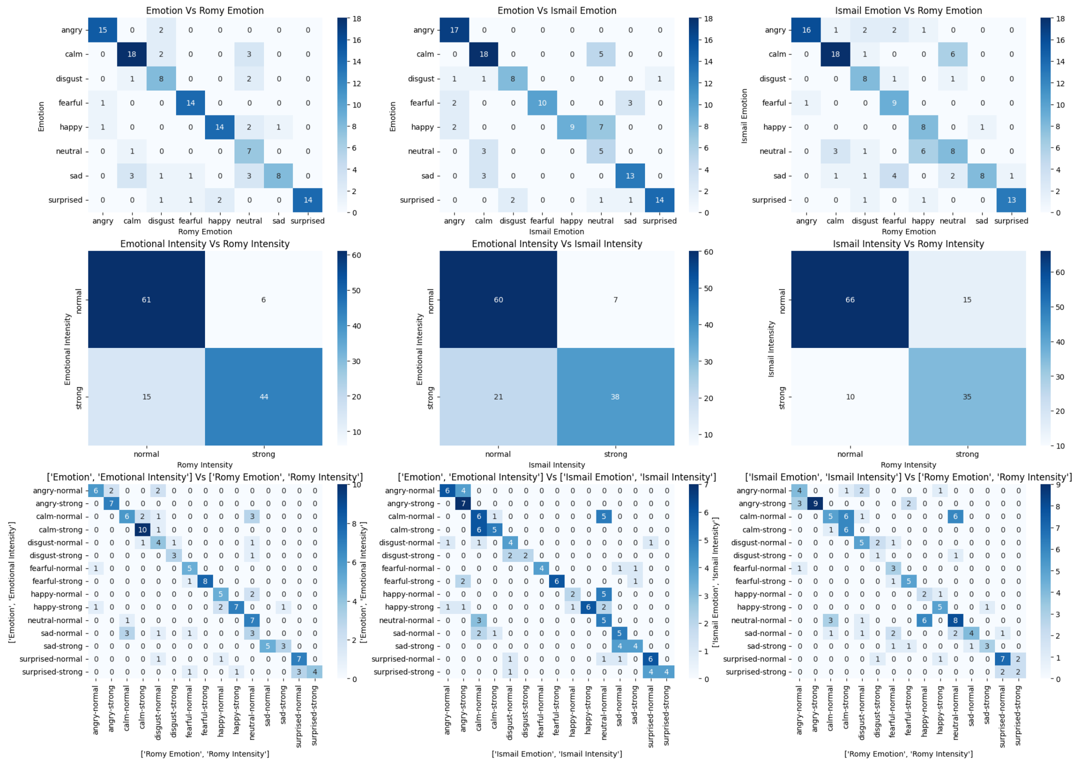


FIGURE 4.1: The Distribution of emotions, emotional intensities and their combination compared to Gold Annotation and Ismail and Romy's Annotation.

- **The κ Coefficient** :
  Give a different probability to each tag and each annotator.
  Different annotators have different interpretations of the instructions (bias).
  κ takes individual bias into account.

$$A_e = \frac{1}{N^2} \sum_{q \in Q} n_{q,1} \times n_{q,2}$$

with $n_{q,1}$: Number of judgments for the category q for annotator 1 and $n_{q,2}$ for annotator 2.

In the end, the calculation for each coefficient will be executed separately on Python following this formula:

$$IAA = \frac{A_0 - A_e}{1 - A_e}$$

The IAA results are represented in Tables 4.1, 4.2 and 4.3

	IAA for S Coefficient	IAA for π Coefficient	IAA for κ coefficient
Romy's Annotation vs. Gold Annotation	74.6%	74.44%	74.51%
Ismail's Annotation vs. Gold Annotation	70.97%	70.69%	70.82%
Romy's Annotation vs. Ismail's Annotation	65.53%	65.21%	65.33%

TABLE 4.1: IAA Results for Emotions

Note that the result's threshold for all the coefficients mentioned will be analyzed based on this image 4.2 *Inter-Annotator Agreement (IAA)* n.d.

	IAA for S Coefficient	IAA for π Coefficient	IAA for κ coefficient
Romy's Annotation vs. Gold Annotation	66.66%	66.04%	66.22%
Ismail's Annotation vs. Gold Annotation	55.55%	54.15%	54.73%
Romy's Annotation vs. Ismail's Annotation	60.31%	57.76%	57.83%

TABLE 4.2: IAA Results for Emotional Intensity

	IAA for S Coefficient	IAA for π Coefficient	IAA for κ coefficient
Romy's Annotation vs. Gold Annotation	62.58%	62.39%	62.5%
Ismail's Annotation vs. Gold Annotation	54.08%	53.66%	53.9%
Romy's Annotation vs. Ismail's Annotation	51.53%	50.83%	51.02%

TABLE 4.3: IAA Results for Emotion and Emotional Intensity

As we can see, the results obtained from the emotions are between 65.21% and 74.6% which gives us a substantial agreement. When looking closer, we can see that the agreement between the annotators and the gold annotation (around 70%) is much better than the agreement between the 2 annotators (around 60%).
Regarding the intensity and the combination of intensity and emotion, the results are between 50.83% and 66.66% which falls mostly in the moderate agreement category. That is why in this study, emotions only will be taken into consideration.

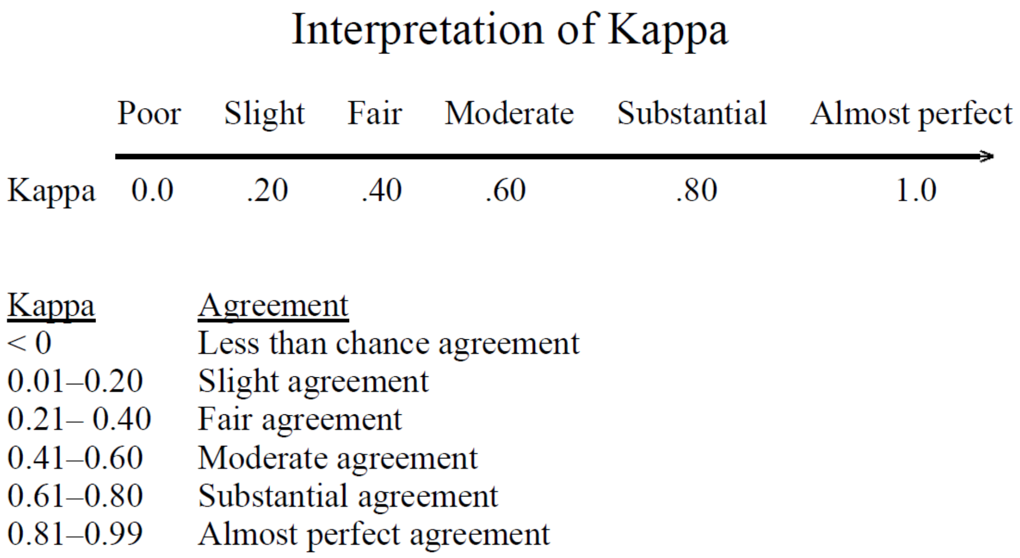
4.2 Models

SER is a trending research area introduced two decades ago. Several innovative strategies have been introduced to enhance the performance of SER. The two fundamental phases of SER are feature extraction and emotion classification. In the feature extraction phase, it can be a manually created feature or a learned feature using DLs. Additionally, classification can be carried out using ML, e.g., DL. Islam et al., 2022b

ML is a branch of AI that makes applications to have better and accurate predictions without hard coding it to do so. It uses historical data to predict new values. There are various types of ML: Iliev, 2023

- **Supervised Learning** : the machine is trained on a set of labeled data, which means that the input data is paired with the desired output. The machine then learns to predict the output for new input data. Supervised learning is often used for tasks such as classification (which is our case), regression, and object detection *Supervised and Unsupervised learning* n.d.
  There are 2 types of SL algorithms:
  - **Regression** : when the output variable is a continuous value such as price or weight.
  - **Classification** : when the output variable is a category such as emotion ('sad' or 'happy') or color ('red' or 'blue').
- **Unsupervised Learning** : Machine is trained on a set of unlabeled data, that is, on input data not represented by the desired output. The machine learns to find patterns and relationships in the data. Unsupervised learning is often used for tasks such as clustering, dimensionality reduction, and anomaly detection.*Supervised and Unsupervised learning* n.d.

Hyperparameters are customizable settings that influence the learning process of the model. By scrupulously selecting and tuning these hyperparameters, the model's ability to generalize and make accurate predictions on unseen data can be



Kappa	Agreement
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21– 0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

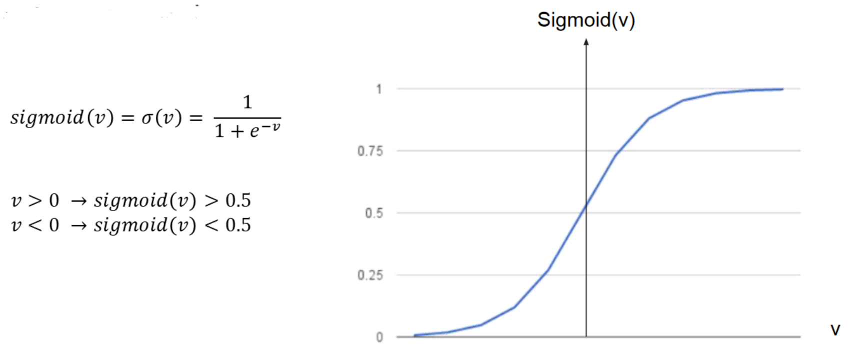
FIGURE 4.2: Interpretation of κ

enhanced. In order to get the ideal configuration that produces the greatest model performance, hyperparameter tuning includes systematically experimenting with combinations of hyperparameter values. Elie DINA, 2024

This section will introduce the different models used in this project and the hyperparameters used:

4.2.1 Logistic Regression (LR)

Logistic Regression is a way to map the values obtained from Linear regression to become between 0 and 1. The core of logistic regression is what we call the logistic function or the sigmoid function - Figure 4.3. This function maps values from all ranges to become between 0 and 1.



$$sigmoid(v) = \sigma(v) = \frac{1}{1+e^{-v}}$$

$$v > 0 \rightarrow sigmoid(v) > 0.5$$
$$v < 0 \rightarrow sigmoid(v) < 0.5$$

FIGURE 4.3: Sigmoid Function

The hyperparameter used is :

- C: which is the inverse of regularization strength; must be a positive float. This parameter was tuned on 0.01, 0.1 and 1.
- **Penalty**: The norm of penalties used are L1 or 'Lasso Regression' and L2 'Ridge Regression'.

- **Solver**: or the algorithm used in optimization problem. Solver has been tuned on 'lbfgs'.
- **max_iter**: is the maximum number of iterations taken for the solvers to converge. The numbers taken are 25,50 and 100.

4.2.2 Support Vector Machine (SVM)

SVM is a max-margin model that aims to find the best separator between several classes . The best separator is the hyperplane that maximizes the margin between the classes. The margin is the distance between the support vectors and the decision boundary. This process is shown in Figure 4.4
The hyperparameters used are:

- **C**: The inverse of regularization strength; must be a positive float. This parameter was tuned on 0.01, 0.1, 1, and 10.
- **Kernel**: Specifies the kernel type to be used in the algorithm. The kernels used are 'linear' and 'rbf'.

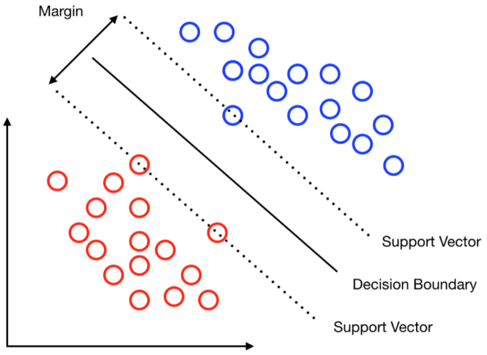


FIGURE 4.4: SVM

4.2.3 K-Nearest-Neighbors (KNN)

KNN looks at the k neighboring data points for a data example to classify and predicts the class based on the majority vote of the k neighbors - Figure 4.5.
The hyperparameter tuned in this case was:

- **n_neighbors**: The number of neighbors. This parameter was tuned on 10, 15, and 20.

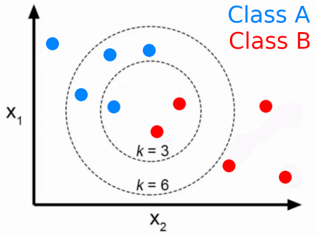


FIGURE 4.5: K-Neighbors Classifier

4.2.4 Decision Tree (DT)

DT is a type of flowchart used to visualize the decision-making process. DT creates a model that can predict the class or value of the target variable by learning simple decision rules during training (Entropy, Gini Index). It contains 3 types of nodes: root node and split node which both split based on condition and leaf node which give the final decision. An example of its architecture is presented in Figure 4.6
The hyperparameter used for this model is:

- **max_depth**: Responsible for the maximum depth of the decision tree, thus limiting the number of created splits. The values used are 3, 5, and 7.

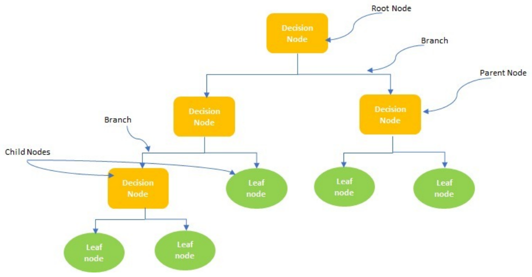


FIGURE 4.6: Decision Tree

4.2.5 Random Forest (RF)

RF generates many DT's and combines them for improved prediction accuracy and stability as shown in 4.7. It is a crowd of experts where trees can say their opinions but whose final decision will be taken by majority vote.
The hyperparameters used are:

- **max_depth**: Represents the maximum depth of the tree. The values used are 5, 10, and 25.

- **n_estimators**: Represents the number of trees in the forest. The values used are 30, 50, and 100.

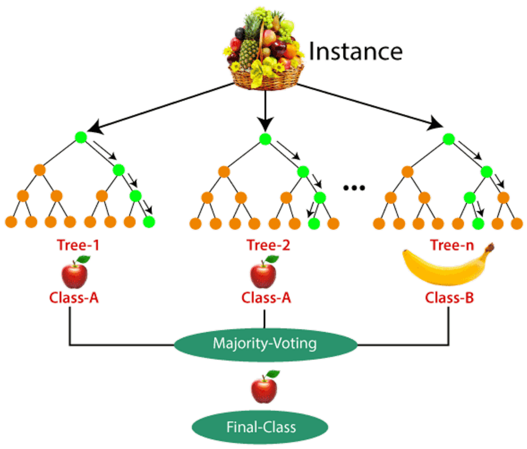


FIGURE 4.7: Random Forest

4.2.6 XGBoost

XGBoost stands for eXtreme Gradient Boosting and is implemented with ML algorithms within the Gradient Boosting framework. It has many advantages such as scalability, portability, and distributed solutions. The gradient boosting framework consists of 3 main phases: a loss function to be optimized, weak learners to make predictions, and a collective model to add weak learners to minimize the loss function *XGBoost* 2014. It is represented in Figure 4.6

The hyperparameters used are:

- **n_estimators**: the number of trees in the forest. This parameter was tuned with values 100, 200, and 500.
- **max_depth**: the maximum depth of the tree. This parameter was tuned with values 2, 3, and 10.
- **learning_rate**: or 'eta', is the step size shrinkage used to prevent overfitting. This parameter was tuned with values 0.01, 0.1, and 1.

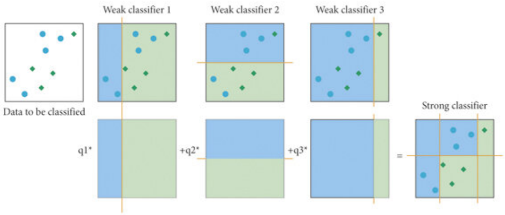


FIGURE 4.8: XGBoost

4.2.7 AdaBoost

AdaBoost stands for Adaptive Boosting and is classified as a boosting algorithm since it creates a strong classifier from a number of weak ones. It trains the weak learners sequentially and not in parallel like bagging algorithms. This is done by building a model from the training data, then creating a second model that attempts to correct the errors from the first model as show in Figure 4.9.

The hyperparameters used are:

- **n_estimators**: the number of weak learners to train. This parameter was tuned with values 50 and 100.
- **learning_rate**: controls the contribution of each weak learner. This parameter was tuned with values 0.01, 0.1, and 1.

4.2.8 MLP Classifier

MLP stands for MultiLayer Perceptron (MLP) and is a technique of feed-forward artificial neural networks using a backpropagation learning method to classify the target variable used for supervised learning. It consists of multiple layers of neurons where the number of neurons is defined by the user and it uses an activation function - Figure 4.10

The hyperparameters used are:

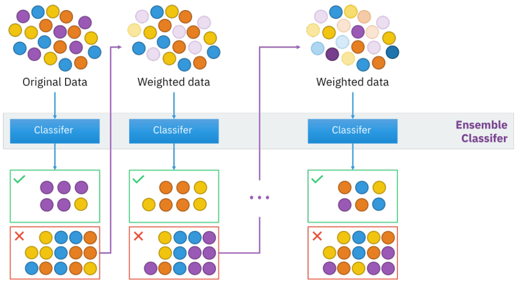


FIGURE 4.9: Ada Boost

- **hidden_layer_sizes**: represents the number of neurons in the hidden layer. This parameter was tuned with values (50,), (100,), and (200,).
- **activation**: the activation function used for hidden layers. This parameter was tuned with values 'relu' and 'tanh'.
- **solver**: the solver for weight optimization. This parameter was tuned with value 'adam'.
- **alpha**: the strength of the L2 regularization. This parameter was tuned with values 0.001, 0.01, 0.1, and 1.

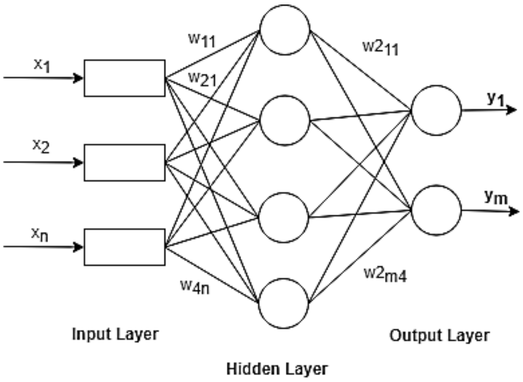


FIGURE 4.10: MultiLayer Perceptron

4.2.9 DL Models

In the following, three models are inspired by Hamid, 2023's work from Kaggle, from where some changes were conducted to some parameters and different feature extractions were tested for each model. Seven emotions are going to be calculated (neutral, sad, disgust, angry, happy, surprise, and fear) by combining calm with neutral. the ratio taken will be 80:20 training and testing sets. The data is scaled with Sklearn's Standard scaler. The following functions were used:

- `ModelCheckpoint Callback`: Saves the model's weights to 'best_model1 _weightsh5' based on the validation accuracy, ensuring only the best model is saved.
- `EarlyStopping Callback`: Monitors the validation accuracy. It automatically stops training if the validation accuracy doesn't improve for

10 epochs and restores the model weights from the epoch with the best validation accuracy.

- `ReduceLROnPlateau Callback:` Monitors the validation accuracy. If the validation accuracy doesn't improve for 3 epochs, the learning rate is reduced by a factor of 0.5, with a minimum learning rate of 0.00001.

The following CNN architecture conducted contains five FB layers, each FB has a 1D Convolution with the "Rectified Linear Unit (ReLU)" activation function which will be defined in Section 4.2.9, "same" padding, 1D Max-Pooling with stride equal to two and "same" padding, and a dropout layer with a dropout rate of 0.2. The first three FBs contain for the 1D convolution a Kernel Size of five, and for the 1D Max-Pooling a pooling size of five. As for the fourth Kernel Size, it will be changed to three, and the final FB will have a Kernel Size of five for the 1D convolution and a pool size of 3 for the 1D Max-Pooling. This output will go through an FC layer with "ReLU" activation, following a Batch Normalization, which will finally go through FC with "softmax" activation and 7 units output.

The model was trained using the training x and y(target) variables for 40 epochs. During training, the validation data was used. Each batch during training consisted of 64 samples, and callbacks including the ones mentioned were employed.

ReLU activation

The ReLU activation function is differentiable at all points except at zero. For values greater than zero, the maximum of the function is taken (Krishnamurthy, 2024)

It follows the equation:

$$\text{ReLU}(x) = \max(0, x)$$

Its graphic presentation is presented in Figure 4.11

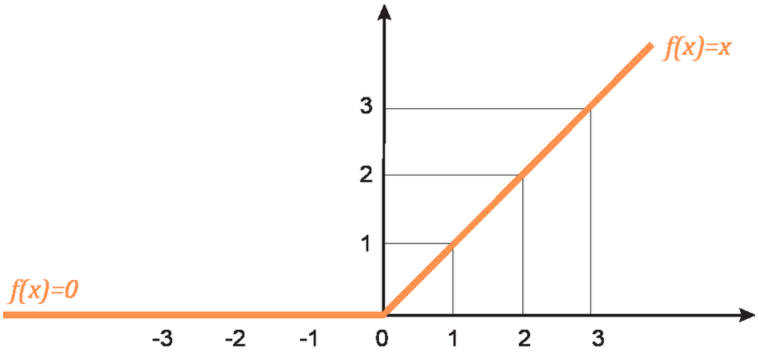


FIGURE 4.11: ReLU Graph Presenation

Model 1

The first model consists of having as input one frequency audio domain feature MFCC and two, time domain audio features ZCR and RMSE. The output shape and parameters are presented in Table 4.4

Layer (type)	**Output Shape**	**Param #**
conv1d_5 (Conv1D)	(None, 2376, 512)	3072
batch_normalization_6 (BatchNormalization)	(None, 2376, 512)	2048
max_pooling1d_5 (MaxPooling1D)	(None, 1188, 512)	0
conv1d_6 (Conv1D)	(None, 1188, 512)	1311232
batch_normalization_7 (BatchNormalization)	(None, 1188, 512)	2048
max_pooling1d_6 (MaxPooling1D)	(None, 594, 512)	0
dropout_3 (Dropout)	(None, 594, 512)	0
conv1d_7 (Conv1D)	(None, 594, 256)	655616
batch_normalization_8 (BatchNormalization)	(None, 594, 256)	1024
max_pooling1d_7 (MaxPooling1D)	(None, 297, 256)	0
conv1d_8 (Conv1D)	(None, 297, 256)	196864
batch_normalization_9 (BatchNormalization)	(None, 297, 256)	1024
max_pooling1d_8 (MaxPooling1D)	(None, 149, 256)	0
dropout_4 (Dropout)	(None, 149, 256)	0
conv1d_9 (Conv1D)	(None, 149, 128)	98432
batch_normalization_10 (BatchNormalization)	(None, 149, 128)	512
max_pooling1d_9 (MaxPooling1D)	(None, 75, 128)	0
dropout_5 (Dropout)	(None, 75, 128)	0
flatten_1 (Flatten)	(None, 9600)	0
dense_2 (Dense)	(None, 512)	4915712
batch_normalization_11 (BatchNormalization)	(None, 512)	2048
dense_3 (Dense)	(None, 7)	3591
Total params		**7193223 (27.44 MB)**
Trainable params		**7188871 (27.42 MB)**
Non-trainable params		**4352 (17.00 KB)**

TABLE 4.4: Model 1 Summary

Model 2

In the second model and for the rest of the models, two frequency domain features will be used with one, time domain audio feature to test its efficiency. This model consists of MFCC and SC as frequency-domain audio features, and ZCR as time domain audio feature. The learning rate is automatically adjusted by the callback function to reach approximately 1.56e-05. The output shape and parameters are presented in Table 4.5

Model 3

In this model, Chroma-STFT and SC will be used as frequency domain audio features, and ZCR as time domain audio features. ReduceLROnPlateau callback function reduced the learning rate to approximately 6.25e-05. The summary of this model is presented in Table 4.6

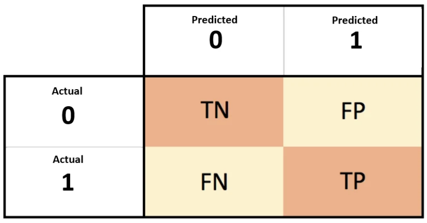
4.3 Metrics

The classification models were further validated by multiple classification metrics, which assess the effectiveness of the models in making predictions. Among the key measures that shed light on different performance components are Accuracy, Precision, Recall, and F1-Score in models. Having understood these metrics, it becomes easy to arrive at a general understanding of how good classification models have performed, swaying decisions on effectiveness in solving classification problems.

Layer (type)	**Output Shape**	**Param #**
conv1d (Conv1D)	(None, 2376, 512)	3072
batch_normalization (BatchNormalization)	(None, 2376, 512)	2048
max_pooling1d (MaxPooling1D)	(None, 1188, 512)	0
conv1d_1 (Conv1D)	(None, 1188, 512)	1,311,232
batch_normalization_1 (BatchNormalization)	(None, 1188, 512)	2048
max_pooling1d_1 (MaxPooling1D)	(None, 594, 512)	0
dropout (Dropout)	(None, 594, 512)	0
conv1d_2 (Conv1D)	(None, 594, 256)	655,616
batch_normalization_2 (BatchNormalization)	(None, 594, 256)	1024
max_pooling1d_2 (MaxPooling1D)	(None, 297, 256)	0
conv1d_3 (Conv1D)	(None, 297, 256)	196,864
batch_normalization_3 (BatchNormalization)	(None, 297, 256)	1024
max_pooling1d_3 (MaxPooling1D)	(None, 149, 256)	0
dropout_1 (Dropout)	(None, 149, 256)	0
conv1d_4 (Conv1D)	(None, 149, 128)	98,432
batch_normalization_4 (BatchNormalization)	(None, 149, 128)	512
max_pooling1d_4 (MaxPooling1D)	(None, 75, 128)	0
dropout_2 (Dropout)	(None, 75, 128)	0
flatten (Flatten)	(None, 9600)	0
dense (Dense)	(None, 512)	4,915,712
batch_normalization_5 (BatchNormalization)	(None, 512)	2048
dense_1 (Dense)	(None, 7)	3591
Total params		**7,193,223**
Trainable params		**7,188,871**
Non-trainable params		**4,352**

TABLE 4.5: Model 2 Summary

This section will review each of these metrics by discussing what they measure and providing detailed explanations and their formulas. The confusion matrix is presented in Figure 4.12.



	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

FIGURE 4.12: Confusion Matrix

1. **Accuracy**: Accuracy shows the models' overall performance in making correct predictions across all classes (Elie DINA, 2024).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2. **Recall**: Recall is a metric that quantifies the ability of a classification model to correctly identify positive instances out of all actual positive instances (Elie DINA, 2024).

$$Recall = \frac{TP}{TP + FN}$$

Layer (type)	**Output Shape**	**Param #**
conv1d (Conv1D)	(None, 1512, 512)	3072
batch_normalization (BatchNormalization)	(None, 1512, 512)	2048
max_pooling1d (MaxPooling1D)	(None, 756, 512)	0
conv1d_1 (Conv1D)	(None, 756, 512)	1,311,232
batch_normalization_1 (BatchNormalization)	(None, 756, 512)	2048
max_pooling1d_1 (MaxPooling1D)	(None, 378, 512)	0
dropout (Dropout)	(None, 378, 512)	0
conv1d_2 (Conv1D)	(None, 378, 256)	655,616
batch_normalization_2 (BatchNormalization)	(None, 378, 256)	1024
max_pooling1d_2 (MaxPooling1D)	(None, 189, 256)	0
conv1d_3 (Conv1D)	(None, 189, 256)	196,864
batch_normalization_3 (BatchNormalization)	(None, 189, 256)	1024
max_pooling1d_3 (MaxPooling1D)	(None, 95, 256)	0
dropout_1 (Dropout)	(None, 95, 256)	0
conv1d_4 (Conv1D)	(None, 95, 128)	98,432
batch_normalization_4 (BatchNormalization)	(None, 95, 128)	512
max_pooling1d_4 (MaxPooling1D)	(None, 48, 128)	0
dropout_2 (Dropout)	(None, 48, 128)	0
flatten (Flatten)	(None, 6144)	0
dense (Dense)	(None, 512)	3,146,240
batch_normalization_5 (BatchNormalization)	(None, 512)	2048
dense_1 (Dense)	(None, 7)	3591
Total params		**5,423,751**
Trainable params		**5,419,399**
Non-trainable params		**4,352**

TABLE 4.6: Model 3 Summary

3. **Precision**: Precision is a metric that measures the ability of a classification model to correctly predict positive instances out of all instances predicted as positive (Elie DINA, 2024).

$$Precision = \frac{TP}{TP + FP}$$

4. **F1 score**: F1-Score is a metric that combines precision and recall into a single measure, providing a balanced evaluation of a classification model's performance (Elie DINA, 2024).

$$F1Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Chapter 5

Results and Discussion

5.1 DL Models

The training and testing accuracy for the three given models - Figures 5.1, 5.2, and 5.3 shows almost convergence to 1 on the 40th epoch for the training and testing accuracy, with a bit of fluctuation in the first proposed model, considered normal in DL tasks. More epochs could have given better accuracy and should have been used, but due to GPU limitations and power only 40 epochs were performed. It can be also said that the training and testing accuracy follow each other and are almost similar which indicates good performance. Regarding the training and testing losses, it can be seen that all models almost converge to 0 on the 40th epoch, with Model 1 having the most fluctuations and Model 3 being the closest to 0. Converging to 0 is a good and necessary sign for DL models.

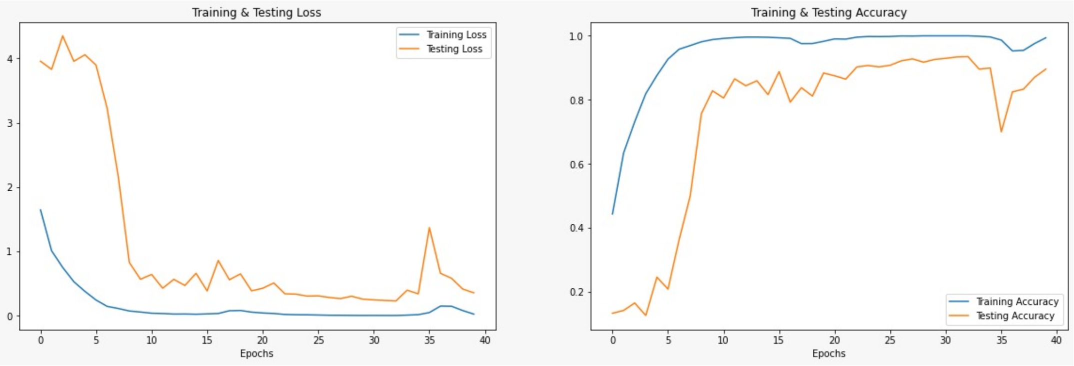


FIGURE 5.1: Training and Testing Loss and Accuracy for Model 1

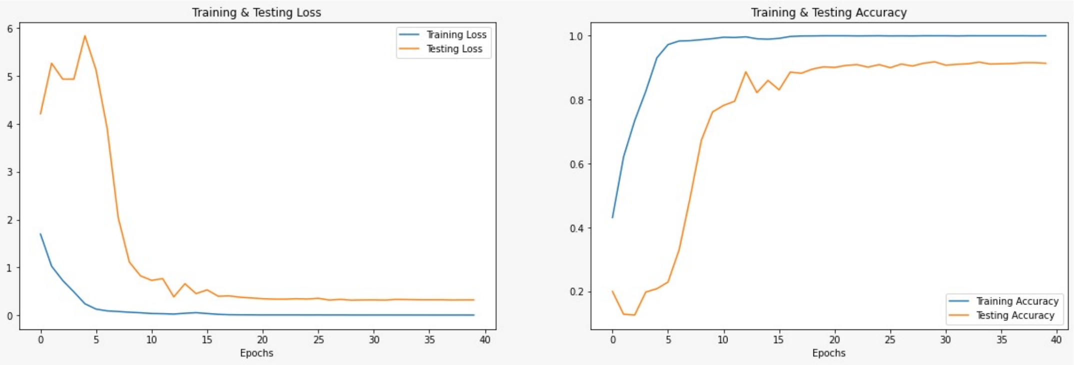


FIGURE 5.2: Training and Testing Loss and Accuracy for Model 2

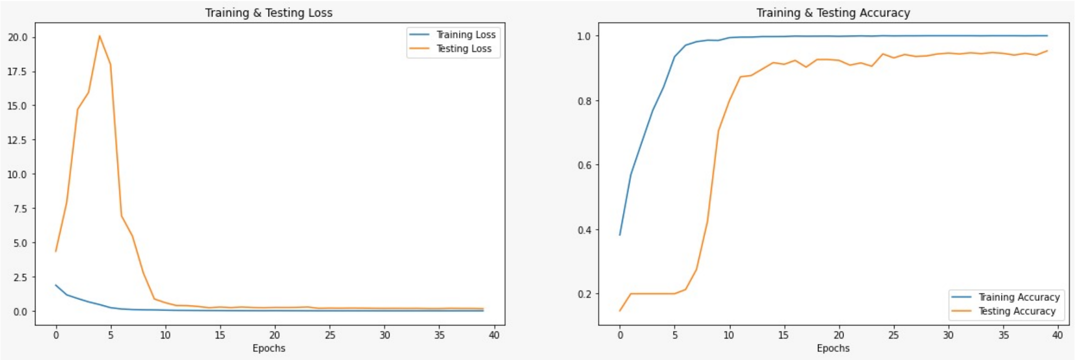


FIGURE 5.3: Training and Testing Loss and Accuracy for Model 3

Regarding Model 1 containing features MFCC, ZCR, and RMSE, a remarkable accuracy for the test data is perceived with 89.58%. The prediction of emotions is shown in Figure 5.4 and their calculated Precision, Recall, and F1-Score are presented in Table 5.1. It is shown that Disgust is the most recognised emotion with an F1-Score of 92% where 153 of 166 emotions were recognised and 13 out of 166 were misclassified. Followed by Angry, Neutral, and Surprise as the second-best performers with 91% F1-Score. However, the angry emotion showed a better performance in Precision with 97% while getting a Recall of 87%, which means that it predicted more positive outputs than it exists. The same can be said for surprise which contained similar results, but neutral emotion has 89% Precision and 93% Recall which means that 89% were correctly classified and 93% of the actual neutral class were identified. The worst performance predicted is happy with an 87% F1-Score from where 16 out of 149 were misclassified.

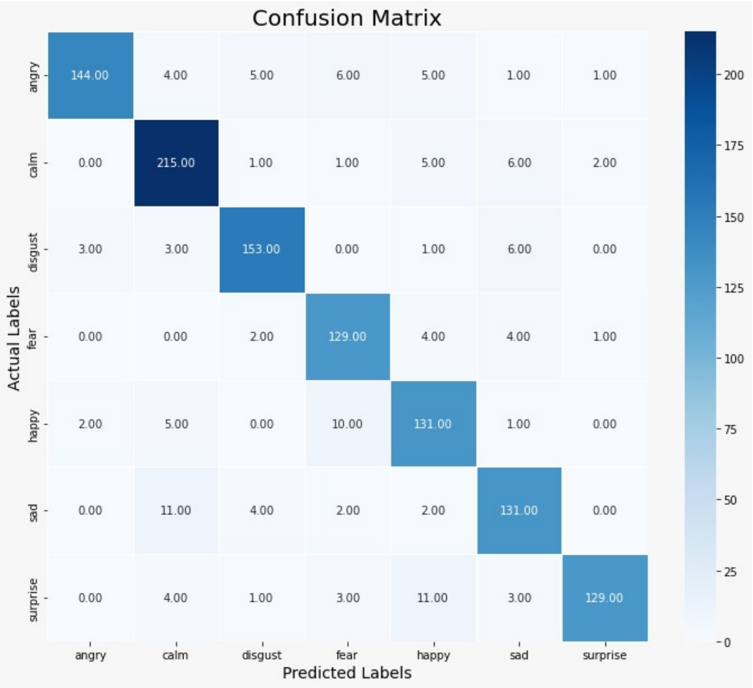


FIGURE 5.4: Heatmap by Model 1 for Emotion Recognition

Model 2 outperformed Model 1 by obtaining an accuracy on the test data of 91.84%. Figure 5.5 presents the predicted and actual classes of each emotion, and

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Angry	0.97	0.87	0.91	166
Neutral	0.89	0.93	0.91	230
Disgust	0.92	0.92	0.92	166
Fear	0.85	0.92	0.89	140
Happy	0.82	0.88	0.85	149
Sad	0.86	0.87	0.87	150
Surprise	0.97	0.85	0.91	151
Accuracy			0.90	1152
Macro Avg	0.90	0.89	0.89	1152
Weighted Avg	0.90	0.90	0.90	1152

TABLE 5.1: Precision, Recall, F1-Score, and Support For Each Emotion Class in Model 1.

the results are thoroughly explained in Table 5.2. It is shown that the Disgust, Neutral, and Surprise classifications outperformed others by obtaining a 93% F1-Score. Neutral differs in Precision (88%) and Recall (98%) while the other two stated emotions are almost similar in Precision and Recall. The worst performers are Happy and Sad with a 90% F1-Score.

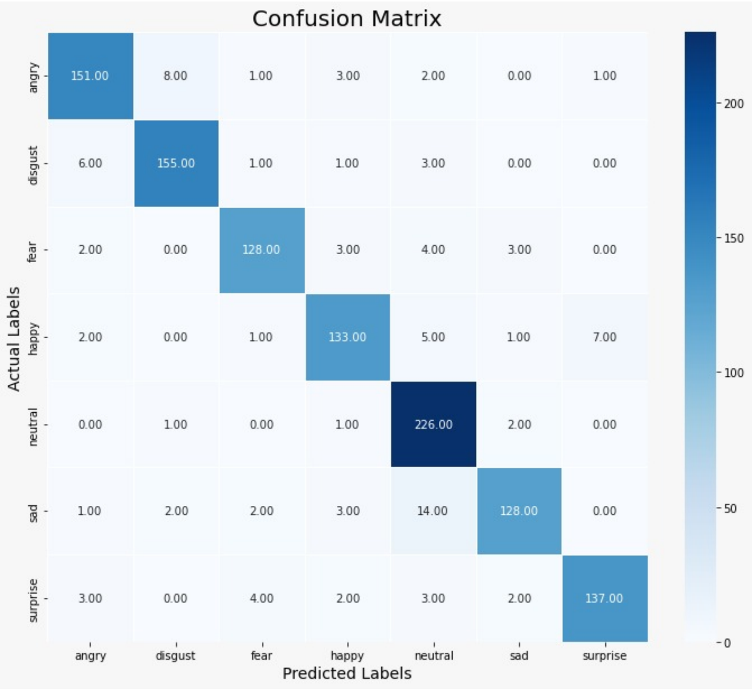


FIGURE 5.5: Heatmap by Model 2 for Emotion Recognition

For the last proposed model (Model 3), as before the confusion matrix of the actual and predicted emotions is shown in Figure 5.6, and their metrics are detailed in Table 5.3. Model 3 outperformed all the other proposed models with an accuracy of 95.31% on the test data. Neutral got the best prediction with an F1-Score of 97% and Recall of 100% which signifies that all the actual neutral classes were correctly classified. It is followed by Surprise, Disgust, and Angry with a 96% F1-Score. The worst performer is Happy with an F1-Score of 92%, by having 9 falsely classified out of 149.

Class	Precision	Recall	F1-Score	Support
Angry	0.92	0.91	0.91	166
Disgust	0.93	0.93	0.93	166
Fear	0.93	0.91	0.92	140
Happy	0.91	0.89	0.90	149
Neutral	0.88	0.98	0.93	230
Sad	0.94	0.85	0.90	150
Surprise	0.94	0.91	0.93	151
Accuracy			0.92	1152
Macro Avg	0.92	0.91	0.92	1152
Weighted Avg	0.92	0.92	0.92	1152

TABLE 5.2: Precision, Recall, F1-Score, and Support For Each Emotion Class in Model 2

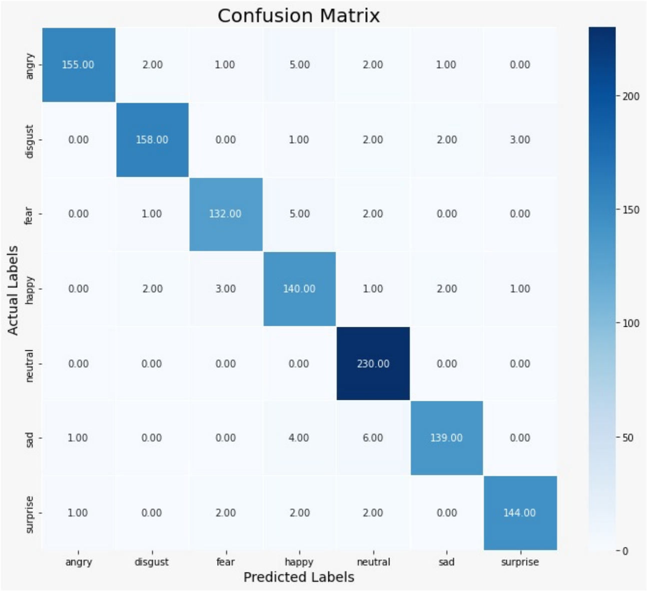


FIGURE 5.6: Heatmap by Model 3 for Emotion Recognition

For the collected Dataset containing audios of random statements lasting 3-5 seconds for different ages and three languages (Lebanese, French, and English), a prediction was made using the proposed model. 5/15 were correctly predicted from the English audios, 4/19 were correctly predicted and a total of 8 from different emotions were wrongfully predicted as sad. For the Arabic audios, 7/48 were correctly classified with also a big number of wrong predictions to different emotions assigned sad. No direct meaning can be extracted from these results since the data is still biased. This data ranges from ages 18-25 mostly, still needs more processing and adjusting. However, some hypothesis can be drawn, and one of the main ones is that non-English speakers (Lebanese and French) are more likely to have any emotion sounding like a sad person, which may be due to the difference in linguistic pronunciation.

Overall, the performed models could have been trained on more epochs for better results. However, these many epochs used confidentially illustrate a good performance. It is observed that Disgust and Neutral got the best predictions out of all the models, and Happy got the most misclassification. Also, Model 3 outperformed Model 1 and 2 with their accuracy results illustrated in Table 5.4

Class	Precision	Recall	F1-Score	Support
Angry	0.99	0.93	0.96	166
Disgust	0.97	0.95	0.96	166
Fear	0.96	0.94	0.95	140
Happy	0.89	0.94	0.92	149
Neutral	0.94	1.00	0.97	230
Sad	0.97	0.93	0.95	150
Surprise	0.97	0.95	0.96	151
Accuracy			0.95	1152
Macro Avg	0.95	0.95	0.95	1152
Weighted Avg	0.95	0.95	0.95	1152

TABLE 5.3: Precision, Recall, F1-Score, and Support For Each Emotion Class in Model 3

Model	Accuracy
Model 1	89.58%
Model 2	91.84%
Model 3	95.31%

TABLE 5.4: Accuracy Results for the Performed DL Models

5.2 ML Models

	LR	SVM	XGBoost	AdaBoost	DT	RF	KNN	MLP
Acc Train (%)	69.49	98.82	100.00	35.20	33.53	96.40	69.46	99.91
Acc Test (%)	48.26	70.03	53.44	31.13	26.67	40.32	44.16	67.08
Prec Train (%)	69.51	98.84	100.00	36.74	35.10	96.92	71.43	99.91
Prec Test (%)	48.14	70.26	53.55	28.71	27.73	38.12	44.41	66.94
Rec Train (%)	69.49	98.82	100.00	35.20	33.53	96.40	69.46	99.91
Rec Test (%)	48.26	70.03	53.44	31.13	26.67	40.32	44.16	67.08
F1Train (%)	69.39	98.82	100.00	33.55	32.38	96.43	69.21	99.91
F1 Test (%)	47.91	69.97	52.80	29.34	25.26	38.50	43.18	66.77

TABLE 5.5: ML Models Performance

In, total, 367 Features and 9 classifiers were included in Grid Search with 3 folds of cross validation in this study. (see also Table 5.5 for all prediction performance evaluations in the supplementary materials). Overall, the most accurate model and feature set combination was the SVM classifier with a hyperparameter of C = 10 with an RBF kernel.(ACC = 70.03%). SVM (ACC = 70.03%) and MLP (ACC = 67.08%)classifiers showed relatively better performance compared to XGBoost(ACC = 53.44%), LR (ACC = 48.26%), KNN (ACC = 44.16%), RF(ACC = 40.32%), AdaBoost(ACC = 31.13%) and DT(ACC = 26.67%) classifiers.
Overall, the worst accuracy performances were detected on DT (ACC = 26.67%) and AdaBoost (ACC = 31.13%) classifiers. However, XGBoost shows more improvement than the other ensemble methods such as RF and AdaBoost.
As shown in figure B.1, MLP had a better performance in predicting the following emotions, Happy(112 out of 152), Sad(98 out of 152), Angry(125 out of 154), Fearful(101 out of 145) and Surprised(95 out of 144). While SVM takes the lead for the following emotions Neutral(43 out of 72), Calm(123 out of 150), and disgust(113 out of 152).

Chapter 6

Conclusion & Future Studies

6.1 Conclusion

This study addressed the recognition of emotion from speech (SER) to help people with difficulties in understanding emotions (e.g. Autism) by contributing to the development of a better model for recognising emotion that can be later on integrated into a tool targeting this purpose. Two approaches were followed in the following project, where the DL approaches outperformed the ML approaches. The highest accuracy result achieved is 95.31% despite using one small dataset (RAVDESS dataset) with data augmentation. From the DL models, the best result gotten out of the three suggested ones, are the ones with two frequency domain audio features and one time domain audio feature. the DL models worked on a CNN architecture due to its suitability for feature extraction. Moreover, the use of Chroma-STFT got better results than MFCC. The best model integrated Chroma-STFT, ZCR, and SC as features. However, from the proposed ML models, the best accuracy achieved is 70.03% by SVM.

Furthermore, extended studies might be interested in using features other than (MFCC, ZCR, SC, and Chroma-STFT), and future studies might also be interested in developing the model's architecture and USING e.g. CNN+LSTM.

6.2 Future Studies

For future studies, it is imperative to train the model on a diverse set of linguistic data to enhance its generalization capabilities across different languages, dialects, and accents. This will ensure that the model is not biased towards a particular linguistic group and can perform accurately in various linguistic contexts. Integrating the wav2vec model, which excels in unsupervised learning of speech representations, with a convolutional neural network (CNN) that classifies data from spectrograms could significantly improve the accuracy and efficiency of both speech recognition and emotion detection. This hybrid approach leverages the strengths of both models, with wav2vec handling raw audio input and CNNs effectively interpreting visual data representations.

Furthermore, the preprocessing phase can be further developed for better accuracy, and the use of a 3D CNN with the combination of Bi-LSTM can be implemented, where 3D CNN is used for its suitability for feature extraction and Bi-LSTM for its capability to handle sequential data. Inspired by the work of (Islam et al., 2022a).

In addition to this, expanding the dataset to include a broader range of speech samples is crucial for thoroughly testing and validating the model's generalization ability. This should involve collecting data from different demographics, including varying ages, genders, and cultural backgrounds, to ensure the model's robustness and reliability in real-world applications.

Moreover, future research should explore the integration of the model with other modalities, such as text emotion recognition (TER) models and facial expression-based emotion models. Combining these different sources of emotional cues can create a comprehensive multi-modal emotion recognition system. Such an integrated system would be capable of analyzing and synthesizing information from speech, text, and facial expressions, leading to a more holistic understanding of human emotions. This multi-modal approach is likely to yield a more robust and versatile model, capable of accurately interpreting and responding to human emotions in a wide variety of contexts, including but not limited to virtual assistants, customer service bots, and therapeutic applications. By pursuing these research directions, a more nuanced and effective emotion recognition system that aligns more closely with the complexities of human emotional expression can be developed.

Appendix A

Data Usage Agreement

This appendix contains the format of the data usage agreement signed by those who registered their audios.

I hereby accept that audio files I provide are to be treated for the sake of the "Speech Emotion Recognition" project, held by Mrs. Romy BOU ABDO and Mr. Ismail KATTAR.

I acknowledge the fact that such audio files are going to be shared with their supervisors DINA Elie and their assigned university supervisor, and agree that it will remain confidential and not be shared with anyone outside of the project team, without explicit consent or as required by law.

Last Name	First Name	Signature

TABLE A.1: Data Usage Agreement

Appendix B

Models' Results

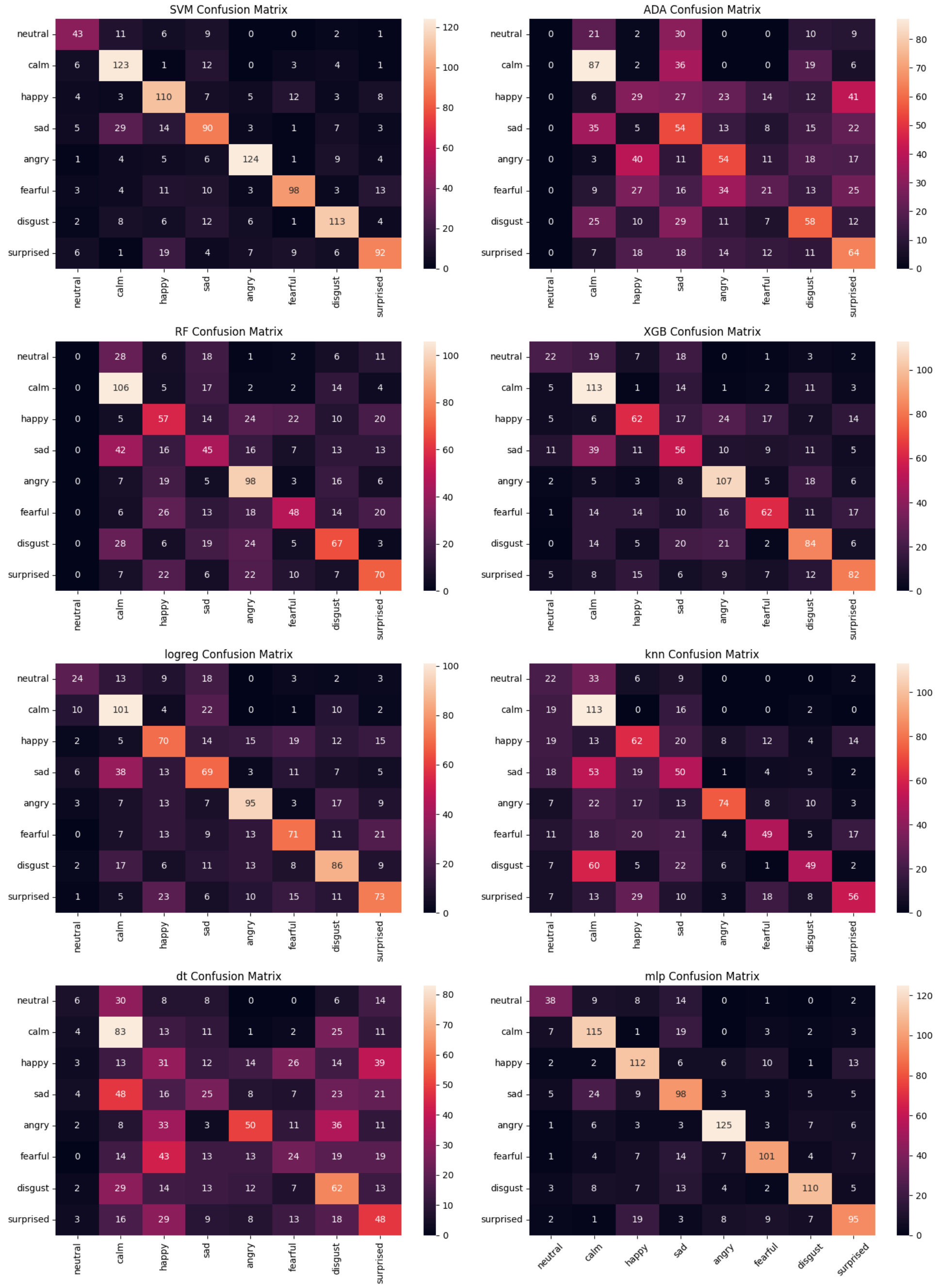


FIGURE B.1: ML Models Heatmaps

Bibliography

Bhangale, Kishor and Mohanaprasad Kothandaraman (Feb. 2023). "Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network". In: *Electronics* 12, p. 839. DOI: 10.3390/electronics12040839.

Dogdu, Cem et al. (Oct. 2022). "A Comparison of Machine Learning Algorithms and Feature Sets for Automatic Vocal Emotion Recognition in Speech". In: *Sensors* 22, p. 7561. DOI: 10.3390/s22197561.

Ekman, Paul (1984). "Expression and the Nature of Emotion". In: *Approaches to Emotion*. Ed. by Klaus Scherer and Paul Ekman. Hillsdale, NJ: Lawrence Erlbaum, pp. 319–344.

Elie DINA Aine DRELINGYTE, Axelle GAPIN (2024). "Sarcasm Detection". In.

Gao, Yuan, Chenhui Chu, and Tatsuya Kawahara (Aug. 2023). "Two-stage Finetuning of Wav2vec 2.0 for Speech Emotion Recognition with ASR and Gender Pretraining". In: pp. 3637–3641. DOI: 10.21437/Interspeech.2023-756.

General Data Protection Regulation (n.d.). https://gdpr-info.eu/.

Gu, S. et al. (2019). "A Model for Basic Emotions Using Observations of Behavior in Drosophila". In: *Frontiers in Psychology* 10. URL: https://doi.org/10.3389/fpsyg.2019.00781.

Hamid, Mostafa Abdul (2023). *CNN-Model*. https://www.kaggle.com/code/mostafaabdlhamed/speech-emotion-recognition-97-25-accuracy/.

Iliev, Alexander I. (2023). "Perspective Chapter: Emotion Detection Using Speech Analysis and Deep Learning". In: *Emotion Recognition*. Ed. by Seyyed Abed Hosseini. Rijeka: IntechOpen. Chap. 2. DOI: 10.5772/intechopen.110730. URL: https://doi.org/10.5772/intechopen.110730.

Inter-Annotator Agreement (IAA) (n.d.). https://towardsdatascience.com/inter-annotator-agreement-2f46c6d37bf3/.

Islam, Md. Riadul et al. (2022a). "Recognition of Emotion with Intensity from Speech Signal Using 3D Transformed Feature and Deep Learning". In: *Electronics* 11.15. ISSN: 2079-9292. DOI: 10.3390/electronics11152362. URL: https://www.mdpi.com/2079-9292/11/15/2362.

Islam, Md. Riadul et al. (July 2022b). "Recognition of Emotion with Intensity from Speech Signal Using 3D Transformed Feature and Deep Learning". In: *Electronics* 11, p. 2362. DOI: 10.3390/electronics11152362.

Jack, Rachael E., Oliver G.B. Garrod, and Philippe G. Schyns (2014). "Dynamic Facial Expressions of Emotion Transmit an Evolving Hierarchy of Signals over Time". In: *Current Biology* 24.2, pp. 187–192. ISSN: 0960-9822. DOI: https://doi.org/10.1016/j.cub.2013.11.064. URL: https://www.sciencedirect.com/science/article/pii/S0960982213015194.

Kendra Cherry, MSEd (2023). *Emotions and Types of Emotional Responses*. https://www.verywellmind.com/what-are-emotions-2795178/.

Krishnamurthy, Bharath (2024). *ReLU Graph*.

Lalitha, S. et al. (2015). "Emotion Detection Using MFCC and Cepstrum Features". In: *Procedia Computer Science* 70. Proceedings of the 4th International Conference on Eco-friendly Computing and Communication Systems, pp. 29–35. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2015.10.020. URL: https://www.sciencedirect.com/science/article/pii/S1877050915031841.

Livingstone SR, Russo FA (2018). "RAVDESS". In: *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English.* https://doi.org/10.1371/journal.pone.0196391.

Papers with Code - Speech Emotion Recognition — paperswithcode.com (n.d.). https://paperswithcode.com/task/speech-emotion-recognition. [Accessed 11-05-2024].

Plutchik, R. (1980). "A general psychoevolutionary theory of emotion". In: *Emotion: Theory, research, and experience: Vol. 1. Theories of emotion*. Ed. by R. Plutchik and H. Kellerman (Eds.) Kellerman. New York: Academic, pp. 3–33.

Scarantino, Andrea and Ronald de Sousa (2021). "Emotion". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University.

Schuller, Björn, Gerhard Rigoll, and Manfred Lang (June 2004). "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture". In: vol. 1, pp. I –577. ISBN: 0-7803-8484-9. DOI: 10.1109/ICASSP.2004.1326051.

Supervised and Unsupervised learning (n.d.). https://www.geeksforgeeks.org/supervised-unsupervised-learning/.

XGBoost (2014). https://xgboost.readthedocs.io/en/stable/.