SPEECH EMOTION RECOGNITION

A major project report submitted in partial fulfilment of the requirement for the award of degree of

Bachelor of Technology

in

Computer Science & Engineering / Information Technology

Submitted by

Vriti Sharma (201353)

Ritik Raushan (201440)

Under the guidance & supervision of

Dr. Pradeep Kumar Gupta

Professor

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology, Waknaghat, Solan-173234 (India)

TABLE OF CONTENTS

Candidate’s Declaration

The author hereby declares that the work that was submitted in this report, which was titled "**Speech Emotion Recognition,**" partially satisfies the requirements for the awarding of a Bachelor of Technology degree in Computer Science & Engineering / Information Technology. This work was completed under the supervision of **Dr. Pradeep Kumar Gupta**, who is a professor in the Department of Computer Science & Engineering and Information Technology at Jaypee University of Information Technology, Waknaghat. The work was conducted from August 2023 to June 2024.

There has been no application for any other degree or certificate pertaining to the subject matter of the report.

(Student Signature with Date)
Student Name: Vriti Sharma
Roll No.: 201353
Student Name: Ritik Raushan
Roll No.: 201440
This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Dr. Pradeep Kumar Gupta
Professor
Computer Science & Engineering and Information Technology
Jaypee University of Information Technology, Waknaghat

CERTIFICATION

This certifies that the work submitted in the project report "**Speech Emotion Recognition**" towards the partial fulfilment of requirements for the award of a B.Tech in Computer Science and Engineering, and submitted to the Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat, is an authentic record of work completed by "**Vriti Sharma(201353) and Ritik Raushan(201440)**" between August 2023 to June 2024, under the direction of Dr. Pradeep Kumar Gupta with the Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat.

Vriti Sharma (201353)
Ritik Raushan (201440)

The above statement made is correct to the best of my knowledge.

Dr. Pradeep Kumar Gupta
Professor
Computer Science & Engineering and Information Technology
Jaypee University of Information Technology, Waknaghat

Acknowledgement

First and foremost, I want to express my profound thanks and admiration to the all-powerful God for the heavenly gift that has allowed us to successfully complete the project work.

My sincere appreciation and responsibilities are owed to Dr. Pradeep Kumar Gupta, who serves as my supervisor in the Computer Science and Engineering Department at Jaypee University of Information Technology in Wakhnaghat. My supervisor has extensive expertise and a strong interest in deep learning, which will be invaluable as we carry out this research. We owe the completion of this project to his boundless patience, intellectual direction, encouragement, vigorous supervision, constructive criticism, helpful counsel, reading of several mediocre draughts and corrections at every level, and so on.

In addition, I would like to express my deepest gratitude to everyone who has helped me in any way, whether it be directly or indirectly, in order to ensure the success of our project. Considering the specifics of the case, I would want to express my gratitude to the numerous members of the staff, both teaching and non-teaching, who have provided me with useful assistance and made my pursuit possible.

Lastly, I must politely thank our parents for their ongoing assistance and patience.

Vriti Sharma (201353),
Ritik Raushan (201440)

ABSTRACT

With applications in customer service, mental health screening, and human-computer interaction, speech emotion recognition (SER) is a rapidly emerging field in study. Our goal in this project is to create a SER system that can reliably and accurately identify emotions in speech data. Using a labelled dataset of different vocal emotional expressions, we address key problems such feature extraction, model selection, data imbalance, robustness, and real-time inference. Our approach integrates deep learning techniques such as Convolutional Neural Networks (CNNs), Long Short Term Memory (LSTM), and Recurrent Neural Networks (RNNs) to extract complex forms from audio features. To ensure that emotions are fairly represented in the training set, we carefully handle data imbalance. We also improve the system's adaptability to different speaking situations. A confusion matrix, F1-score, and accuracy are some of the assessment metrics used to gauge the system's performance. We highlight the model's ability to generalise across various speakers, languages, and recording environments.

The outputs of this project include well-trained, documented SER models, thorough assessment findings, and a thorough report explaining the approach, difficulties, findings, and any improvements. Through subtle sentiment analysis of spoken language, this research has the potential to improve human-computer interactions, assist in mental health evaluations, and improve customer service experiences. Beyond the specifics, our work has enormous potential. We're paving the way for more intuitive and sensitive human-computer interactions by gently analysing the emotions in spoken language. We're assisting in the development of less intrusive, more precise mental health assessments. By enabling businesses to genuinely comprehend their consumers' wants and emotions, we are improving customer service experiences.In summary, this initiative aims to enhance our ability to communicate, comprehend, and empathise in a society that is becoming more and more reliant on technology. One word at a time, it aims to close the divide between technology and mankind.

LIST OF FIGURES

LIST OF TABLES

Chapter 01: Introduction

1.1 Introduction

In a world where digital communication is becoming a norm, it is vital to comprehend the subtle emotions that are part of human being. Emotion is the true essence of our conversations, and it is the main factor which shapes the characters of our relationships. Speech is the richest and most complex method of expressing emotions, of course there are other ways of communication too. The ability to detect emotional states in speech, or speech emotion recognition (SER), has become a major and significant area of research, and it is now a personal matter to me.

The main purpose of the introduction of emotions into SER is the identification and understanding of the emotional content of human speech. It has a great influence on lots of different situations. Through the measurement of customers' emotions and the assistance of mental health professionals in their assessments, SER is the bridge between technology and psychology that is going to improve customer service interactions. [1] Through this link, a better relationship where people and machines can be more sympathetic and receptive is created. Nevertheless, making a dependable and precise system for SER is not simple. These issues cover a wide range of topics, such as the technological difficulties of feature extraction and model selection, the fairness in data representation, the durability across different speaking conditions, and the real-time application. [2].

This project will seek to fully overcome these obstacles. Therefore, to replicate the various patterns that are present in speech which convey emotions, we look into the most advanced deep learning techniques comprising LSTM networks, CNNs, and RNNs. [3].

We are really committed to the idea of making sure that our model accurately reflects all emotions and we have worked hard to tackle the issue of imbalanced data. We prioritize the matter of robustness, and our SER system is made to work in real-life speaking situations, no matter whether it is a noisy or if the recording is of poor quality. Besides, our model is brought to life for inference in real-time, thus, it is a great tool for immediate feedback and interactions[4].

1.2 Objectives

1. The main objective of this study is to create a trustworthy system for speech emotion recognition (SER) that will be able to accurately recognize different emotions in speech samples.
2. Model Selection and Implementation: Evaluate and use the latest SER-specific deep learning and ML models including GRUs (Gated Recurrent Unit) networks, LSTMs (Long Short-Term Memory) networks and others.
3. Evaluation and Performance Metrics: Enhance the model for real-time inference and speed up the emotion detection; assess the model generalizability to different people, languages, and recording conditions.

If the project is successful in achieving these goals, it will have made a significant contribution to the development and practical application of Speech Emotion Recognition technology, which will have improved human-computer interaction and emotional understanding in technologically driven configurations[5].

1.3 Motivation

The context and objectives of the project dictate the motivation, which may be complex and diverse. A few typical reasons to launch a SER project are as follows

1. Enhanced Human-Computer Interaction: The application of SER can increase human contact. Machines like devices, virtual assistants, and chatbots will learn to recognize emotions using the emotion identification system.
2. Market Research and Customer Feedback: One example of how a company may use SER is to comb through customer reviews and comments. The making of better decisions and product development can be achieved by getting a better understanding of the reactions of the clients to the goods, services, or marketing campaigns through this information.
3. Accessibility and Assistive Technology: Individuals with disabilities, especially those who are confronted with speech deficits, are the ones that can be helped by SER. It serves their well-being by aiding them in the process of better communication and the expression of feelings.
4. Security and Surveillance: Emergency situations, criminal investigations and access control systems are only a few of the many examples of how SER can be used in security and surveillance applications to identify elevated emotions in audio data.
5. Research and Scientific Understanding: The scientists in the fields of neuroscience and psychology can use SER to investigate the association between emotional states and the vocal expressions. The use of this methodology helps students understand how emotions are expressed linguistically.
6. Education and Training: By analyzing the effectiveness of the teaching methods, the granting of the constructive criticism to the learners and the improvement of the online learning systems, SER can make a great contribution to the field of education.

7. Cultural and Social Analysis: By the way, the emotion of the people in countries can be understood better by the academics and policymakers through the analysis of the speech emotions in a large dataset which will give them the cultural and social trends.

1.4 Tools & Techniques

Python is a great option for Speech Emotion Recognition (SER) projects since it is highly adaptable, has a large number of libraries, and has a dynamic community of machine learning and natural language processing professionals who are always there to assist you. A student could use Python to make a project that would be the best project example for the project:

1. Data Collection and Preprocessing: It is for certain that you can gather audio data and make it ready for analysis using Python. Audio data is like a friend who helps you with any problems you face, and when you need to do some extra work on it, libraries like 'pyaudio,' 'soundfile,' and 'librosa' become your best friends.
2. Machine Learning and Deep Learning: When it comes to building SER models, Python's vast array of deep learning and ML libraries is truly a wonderful thing. The old machine learning methods have the regular libraries of scikit-learn and also the deep learning frameworks like TensorFlow and PyTorch.
3. Feature Extraction: Python is a fantastic tool that enables you to get all the information that you need from audio signals, like spectrograms, chroma characteristics, and even Mel-frequency cepstral coefficients (MFCCs). In the world of feature extraction, 'librosa' library is usually the first option to choose.
4. Model Training: Machine learning and deep learning models that are trained in Python using audio attributes as input data can really help us in various ways. You can develop and train a whole bunch of models for the SER problem, such as CNNs, RNNs, and even hybrid models.

5. Real-time Prediction: Python may be the tool for real-time audio input capture and processing if your project is going to include real-time emotion identification from live audio streams. Through libraries like pyaudio, you can easily turn your dreams into reality.
6. Visualization: Python is equipped with a lot of libraries for data visualization that can be used for visualizing audio data, model training progress and results. Libraries like matplotlib and seaborn have gained popularity for their use in visualization.

1.5 Technical Requirements

1. Hardware Requirements:
    a. Processing Power: Depending on the complexity of the machine learning models, a computer with a CPU or GPU capable of handling the training and inference processes efficiently.

    b. Memory: Sufficient RAM is essential for loading and processing large datasets, especially for using deep learning models.

    c. Storage: Storage space for audio datasets, model checkpoints, and other project-related files.
2. Software Requirements:
    a. Python: Most SER projects are conducted in Python simply because of the abundance of libraries that are available for machine learning, signal processing, and audio analysis.

    b. Development Environment: A Python IDE like Google Colab or Kaggle helps you to write and debug your code while also allowing you to learn new coding techniques.

    c. Machine Learning Frameworks: You are now all set to install the necessary libraries that will help you to learn deep learning and machine learning. These libraries are Keras, scikit-learn, PyTorch, and TensorFlow, which will make your learning process much easier.

d. Audio Processing Libraries: Libraries such as librosa or soundfile are frequently used for audio data preprocessing, a task that is often very tedious and requires complex programming skills.

1.6 Deliverables/ Outcomes

The carefully-defined objectives and project scope of a Speech Emotion Recognition (SER) project will be the deciding factors of the deliverables and results. However, here are some common deliverables and potential outcomes you can expect from an SER project:However, here are some common deliverables and potential outcomes you can expect from an SER project:

Trained SER Model: The main result is a machine learning or deep learning model that has been trained to recognize emotions in speech, thus bringing us closer to understanding the human experience. This model should be able to hear the speech and tell us the feeling that the speech is trying to convey.

Accuracy Metrics: A written piece or report that tells the SER model stories about the results, which include the number of correct, wrong, and the proportion of the correct to the wrong answers, the F1-score, and the confusion matrices. It is like having a friend who can look at someone and tell you how they are feeling just by the way they are. This is a way to see how good the model is at recognizing emotions, which is a skill that most humans have.

Codebase: The project's original source code, which may have scripts for data preparation, code for training the model, and, if needed, code for real-time predictions. The way to go is to have accurate documentation and code comments that will help in the process of understanding and future development.

Demo/Prototype: Try to create a working demo or prototype that can effectively show the functionality of the SER system, especially if it involves real-time emotion recognition from live audio.

Chapter 02: Literature Survey

2.1 Overview of Literature Survey

The articles have seemed to revolve on deep learning-based Speech Emotion Recognition (SER) based on the material supplied. Here's an overview and identification of potential key gaps in the literature:

1. Datasets: Researchers have utilized various datasets, such as Berlin emotional database, EMDOB, IEMOCAP, TIMIT Corpus database, and Natural Emotional Speech Database, among others. The choice of datasets reflects an emphasis on diverse emotional expressions and acoustic variations in speech.
2. Deep Learning Architectures: Multiple hidden-layer DNNs, GRU, CNN, and CRNN combinations are among the deep learning architectures utilized. Multimodal approaches are mentioned, combining speech with other modalities like visual or textual data.
3. Performance Metrics: Performance metrics, such as accuracy, are highlighted in some papers. There are claims of improved accuracy with the use of certain deep learning architectures, particularly compared to traditional models like GMMs.
4. Challenges: Challenges in scaling multimodal systems for large-scale applications are acknowledged. Concerns about efficiency for temporally-varying input data and large training and tuning overhead are mentioned.
5. Specific Techniques: Spectrogram feature extraction utilising methods such as deep retinal convolutional neural networks is explored.

Table 1: Literature Survey

S.no.	Paper Title	Journal/ Conference (Year)	Tools/ Techniques / Dataset	Results	Limitations
1.	Using Deep Learning Techniques for Speech Emotion Recognition: A Comprehensive Review	IEEE Access	Berlin emotional database	The efficient use of multimodal emotion recognition allows for the simultaneous utilisation of audiovisual input data.	Internal design with several layers; less efficient with input data that changes over time.
2.	Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets:	MDPI - Sensors	EMDOB	Compared to systems that use GMMs as classifiers, emotion identification systems based on deep learning architecture perform better.	As the complexity of multimodal systems increases, scaling them for deployment in large-scale applications can be a significant challenge.
3.	Deep Neural Networks for Emotional Speech Recognition	MDPI - Sensors	IEMOCAP	Combination of GRU, CNN, CRNN increased accuracy exponentially	Large training and tuning overhead.
4.	Speech emotion recognition Jianwei et. al (2014)	IEEE Access	TIMIT Corpus database	We used an DNN (deep neural network) with three hidden layers and the sounds of MFCCs,	Less hidden layers of the considered DNN

				PLPs, and FBANKs to detect 8.2% of the emotions.	
5.	The Automatic Recognition of Emotions in Speech	Springer	Interaction human-information kiosk	Flow of activities, from ideation to recognition rates	No in depth analysis
6.	Deep Learning: Methods and Applications	IEEE Access	EMDOB	Several recent deep learning-based multi-task learning investigations have been carried out on data that is single-modality, such as voice, text, or pictures.	Multi modality is not discussed
7.	Perception of Emotions via Voice	International journal of speech technology	Natural emotional speech database	Recognising emotional states in spoken language using databases of emotions, speech characteristics, and classification models	Speaker specific information
8.	Deep Retinal Convolution Neural Networks for Emotion Recognition in Speech	IEEE Access	IEMOCAP Database	Achieve classification of emotions in voice by collecting high-level features from spectrogram	Multilingual Database not used

2.2 Key Gaps of Literature Survey

1. Multi-Modality Consideration: Even if some of the papers talk about the multimodal methods, there is still a lack of a comprehensive viewpoint in the discussion about the multi-modality in the emotion recognition systems.
2. Scalability Challenges: Moving from being just a hypothesis to a full-fledged project, this particular challenge has been highlighted, but the specific nature of these challenges and the possible solutions are not discussed at length. Eventually, researchers will have to look into how to make the emotion recognition systems suitable for real-world, large-scale applications, which is a potential gap.
3. Temporal Variability Handling: The issue of the inefficiency concern for temporally-varying input data is there, but the handling of the temporal variations in speech by different architectures is not mentioned. It is a possible topic that can be explored even more in the depth. [6
4. Lack of Standardization: The sentence concerning the absence of standardized benchmarks or evaluation metrics for SER systems has been given a human touch. Setting up of common benchmarks and metrics might be the way to go so as to compare the performance of different models in the experiments.
5. Interdisciplinary Perspectives: Even though the papers pay attention to the technical side of SER, the creators admit that interdisciplinary ideas like the combination of psychological or cognitive science principles in the design of emotion recognition systems are not explored.
6. Real-world Deployment Challenges: They are so easily put to use in life that their usefulness is never debated. Converting the research into practical problems, ethical issues, and user acceptance requirements could be the area of further research.[7]

Closing these gaps could contribute to a more comprehensive and robust understanding of Speech Emotion Recognition systems, making them more effective and applicable in real-world scenarios. Researchers may consider addressing these gaps in future work to advance the field

Chapter 03: Feasibility Study, Requirement Analysis and Design

3.1 Feasibility Study

3.1.1 Problem Definition

Automatically identifying and categorising the emotional state or attitude conveyed in spoken language is the goal of Speech Emotion Recognition (SER). The primary objectives of SER include:

1. Emotion Classification: You know, sometimes it's like, you have to give a feeling to a certain part of a speech or the whole thing, like joy, sorrow, anger, fear, surprise or neutral.

2. Emotion Intensity: To find out how much the person is feeling, from the low to the high, thus giving a more clear and in-depth picture of the emotional content.

3. Speaker-Independent Recognition: The person can understand the emotions in speech even if the speaker is a different gender, age, or language background which makes it applicable to all the people.

4. Real-Time or Offline Processing: SER can be used in real-time for such things as emotion-aware virtual assistants, call center analytics, or in offline settings for sentiment analysis in recorded speech data.

5. Multi-Modal Analysis: It is very beneficial when the accuracy of emotion identification goes up by using the information from the other modalities such as physiological signals, body language, and facial expressions.

6. Feature Extraction: The process of extracting the acoustic, prosodic, and linguistic features of the speech signal to characterize emotional cues is what makes the human aspect of the given sentence. These features may also be the way you talk, the way you speak, or even the way you sound.

Machine Learning and Deep Learning: Designing and creating models, algorithms, and training procedures for emotion detection systems. Neural networks and other deep learning techniques are widely utilised in machine learning.[8]

7. Evaluation Metrics: Creating appropriate measuring tools to evaluate the emotion recognition system's performance; examples include confusion matrices, F1-score, and accuracy.[9]

The potential of speech emotion detection for better human-computer interaction and generalised emotion understanding has attracted a lot of interest. To overcome these obstacles and progress the area of emotion identification in speech, researchers and engineers are hard at work creating trustworthy systems.

3.1.2 Problem Analysis

All the many aspects and difficulties of Speech Emotion Recognition (SER) must be considered in order to arrive at a satisfactory analysis of the problem. Here's a more detailed problem analysis:

1. Data Availability and Quality:
    a. Availability of diverse and annotated speech emotion datasets can be a challenge for training and testing SER models.
    b. The quality of the data is crucial, as it should encompass various emotions, speakers, languages, and realistic contexts to ensure model generalization.
2. Cross-Cultural and Multilingual Challenges:
    a. Emotions can be expressed differently across cultures and languages, making it challenging to create universal emotion recognition models.
    b. Cross-lingual emotion recognition is also a complex problem due to linguistic variations.

3. Ambiguity and Subjectivity:
    a. Emotions in speech can be ambiguous and subjective, making it difficult to achieve high agreement among human annotators, let alone automated systems.
4. Overfitting and Generalization:
    a. Robust methods for validating and testing models are necessary to avoid frequent problems like overfitting to training data and underperformance on novel, unseen data.
5. Real-Time Processing::
    a. Implementing SER in real-time applications like virtual assistants or human-computer interfaces demands low-latency solutions.
6. Imbalanced Datasets:
    a. Datasets do have imbalanced emotion class distributions, making it challenging to train models that work well on a few minority emotions.
7. Transfer Learning:
    a. It may be hard to get models trained on one dataset to work well on another, particularly as there is not a big dataset available for that particular application.
8. Emotion Continuum:

Sometimes emotions are difficult to categorise since they are not always precise but rather exist on a continuum.

Speech emotion recognition is a fascinating and crucial field of study because of its difficulties and complexity. We must handle these difficulties and complications if we are to progress the profession and make it more beneficial.

3.1.2 Solution

To be honest, dealing with the problems of Speech Emotion Recognition (SER) is not an easy task. It needs a lot of different people working together in different fields to collect data, figure out the important features, create models, and think about the ethical issues. Here are some possible solutions to tackle the problems in SER that you might have encountered:

1. Data Augmentation and Diverse Datasets: Create larger and more diverse emotion datasets to improve model generalization. Augment existing datasets by adding noise, altering pitch, or introducing different accents and languages to make the model more robust.
2. Balancing Imbalanced Datasets: Apply methods like oversampling, undersampling or even artificial data production to tackle the problem of class imbalance. [10]
3. Feature Engineering: To get a better grasp of the emotional signals, you can try looking at advanced feature extraction methods like Mel-frequency cepstral coefficients (MFCCs), chroma features, and prosodic features. [11]
4. Model Selection: Look into a lot of machine learning and deep learning models, such as the Transformer-based architectures, Recurrent Neural Networks for SER, and Convolutional Neural Networks (CNNs), and you will be able to understand them like humans do. [12]
5. Transfer Learning: To be better at recognizing emotions, especially when there is not enough data, use the transfer learning from the already pre-trained models like BERT or GPT to give the emotion recognition models a boost.
6. Real-Time Processing: Make the algorithms easier to use for real-time applications by reducing model complexity or using specialized hardware for low-latency processing.
7. Subjective Evaluation Metrics: Give the subjective evaluation metrics like human agreement or listener preference along with the standard classification metrics to measure the performance of SER systems.
8. Regularization Techniques: To make sure that the model can be used in different situations without being too specific to the training data, use the regularisation techniques like dropout and L1/L2 regularisation.

The struggle to overcome the challenges in Speech Emotion Recognition is an ongoing process, and the search for new and creative ways to improve the accuracy, robustness, and practical usefulness of emotion recognition systems is ongoing.[13]

3.2 Requirements

3.2.1 Functional Requirements

Functional requirements for speech emotion recognition typically include:

1. Audio Input: The system should be able to work with audio input, either in real-time or from pre-recorded sources, like a regular human would.
2. Feature Extraction: This system needs to pick out the important features of the audio, like pitch, intensity, and spectral features.
3. Emotion Classification: The system should be able to identify the emotion that the speaker is expressing in the speech and classify it into the categories of happy, sad, angry, etc.
4. Real-time Processing (if needed): In real-time applications, the processing of data must be done with low latency as this is the basis of the whole thing.

3.2.2 Non-Functional Requirements

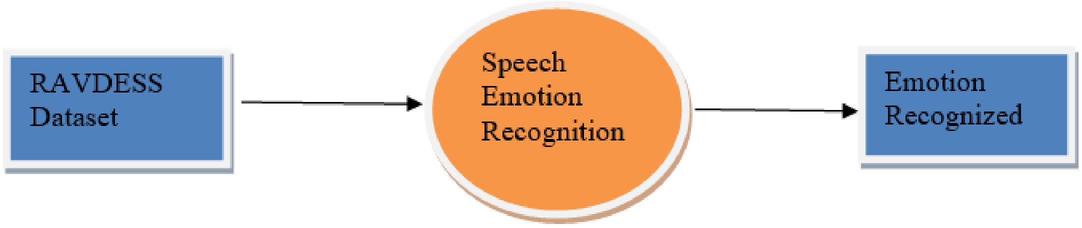
Speech emotion recognition systems need to have requirements that are as vital as the functional ones, Here are some the ones that will not work but will be fun to do:

1. Performance: The system should be able to identify the emotion and react to it quickly since it is designed to be a responsive user experience.
2. Scalability: It should be able to fit more and more users or the data that they will be using without the performance dropping significantly.
3. Portability: The system should be a straightforward and hassle-free one to be used on various platforms and devices, e. g. mobile phones, desktop computers, and embedded systems.
4. Compliance: Following industry standards, ethical guidelines, and legal requirements, for instance, in healthcare and customer service, is a must.
5. Feedback and Monitoring: Put the systems in place to gather suggestions and keep an eye on the system output to make constant improvements.

These non-functional requirements ensure that the speech emotion recognition system not only works correctly but also meets the performance, security, and usability standards expected by users and regulators**.**

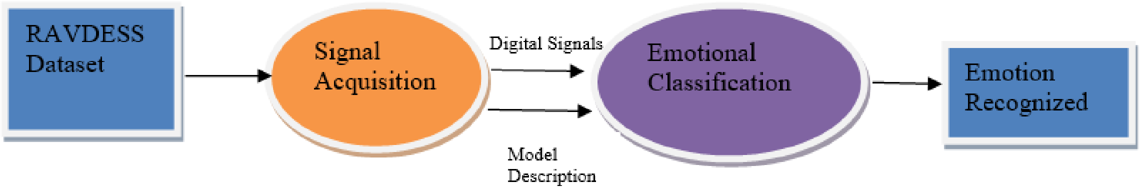
3.3 Data-Flow Diagram (DFD)

1. Level 0 DFD: The Level 0 DFD is a high-level diagram that shows the Speech Emotion Recognition main process, the external entities that are interacting with the system (RAVDESS Dataset), and the data flow between them. It is the first step for the more detailed DFDs at lower levels, wherein the main process is broken down into more detailed subprocesses.



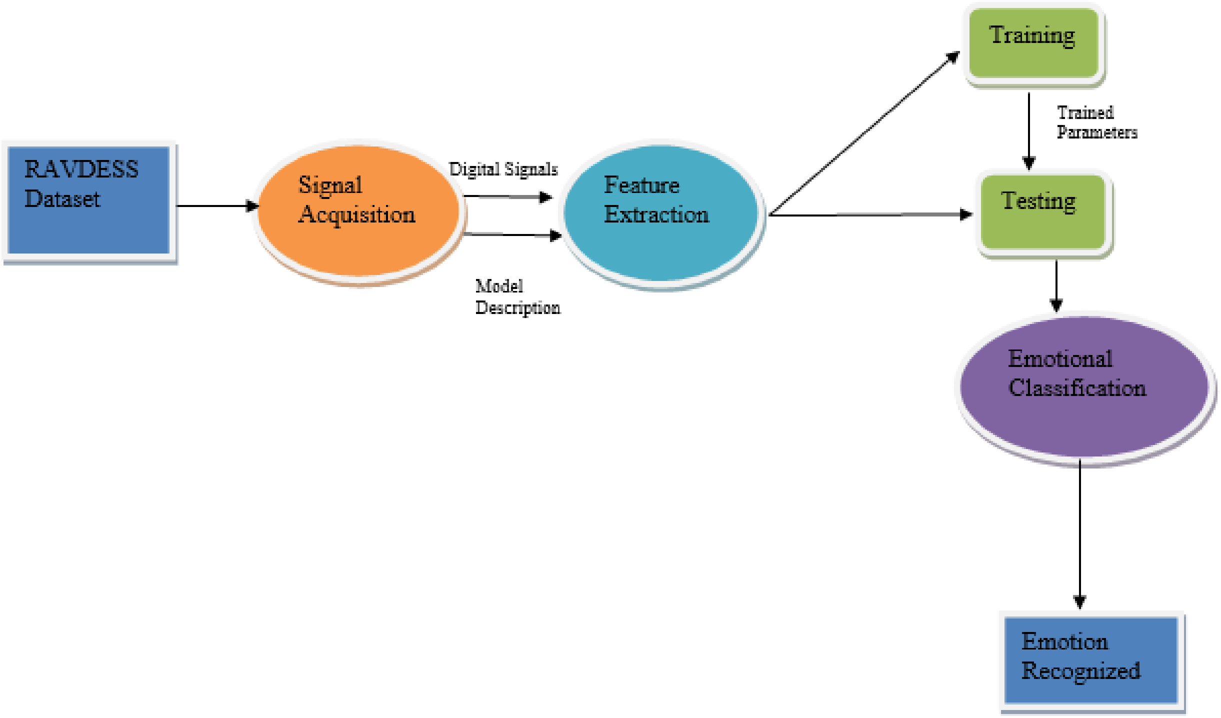
a. 0 Level DFD

2. Level 1 DFD: By breaking the primary process from Level 0 DFD into its constituent subprocesses—Signal Acquisition and Emotional Classification—Level 1 DFD offers a more comprehensive view of the system. It helps in understanding the specific functions of the system and how they interact with each other- Digital Signal transformation and Model Description and with external entities – input RAVDESS dataset and recognized classified emotion. This process of decomposition can be continued to create even more detailed DFDs at lower levels.



b. 1 Level DFD

3. Level 2 DFD: In order to get a better look into the system, Level 2 DFD dissects the subprocesses of Level 1 DFD into even smaller ones, such as feature extraction, model training, and testing. This hierarchical decomposition continues until a sufficient level of detail is reached, allowing for a comprehensive understanding of the system's processes and their interactions.



c. 2 Level DFD

Figure 1: Data Flow Diagram of Speech Emotion Recognition

Chapter 04: Implementation

4.1 Dataset Used

4.1.1 Types of Dataset

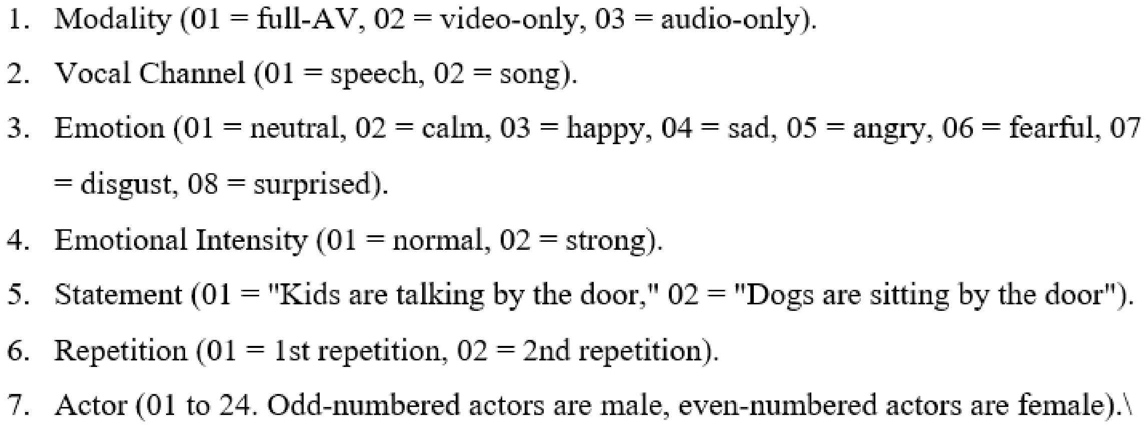
Datasets kept by the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[14] include:

1. Dataset Size: The audio-only portion of the RAVDESS dataset, as mentioned in your initial description, consists of 1,440 audio files. This portion includes speech audio-only files in 16-bit, 48kHz .wav format.
2. Actors: The dataset features a total of 24 professional actors, evenly split between 12 female actors and 12 male actors.
3. Emotional Expressions: Various mindsets, including neutral, calm, happy, sad, furious, scared, disgusted, and astonished, are represented in the dataset. There are two tiers of emotional intensity that these feelings manifest in (normal and strong).
4. Statements: "Dogs are sitting by the door" and "Kids are chatting by the door" are two statements that the actors openly express, and they are lexically matched.
5. Modality: The dataset has three modalities for the audio-visual recordings: full-AV (audio and video), video-only, and audio-only. The information provided pertains to the audio-only modality.

1.1.2 Number of Attributes, fields, description of the data set

The RAVDESS dataset consists of audio files, and it does not have traditional structured attributes or fields like a tabular dataset might have. Instead, the data is organized using a file naming convention that encodes various characteristics of each audio file. You have already identified the main points of the file naming standard in your first message: modality, voice channel, emotion, intensity of emotion, sentence, repetition, and actor.

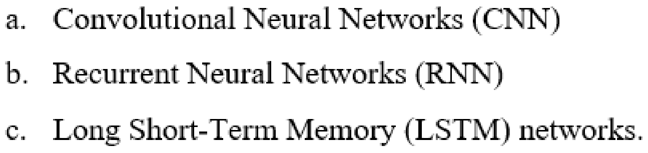
Here is a brief summary of the components of the file naming convention:



1. Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
2. Vocal Channel (01 = speech, 02 = song).
3. Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
4. Emotional Intensity (01 = normal, 02 = strong).
5. Statement (01 = "Kids are talking by the door," 02 = "Dogs are sitting by the door").
6. Repetition (01 = 1st repetition, 02 = 2nd repetition).
7. Actor (01 to 24. Odd-numbered actors are male, even-numbered actors are female).\

1.2 Algorithm / Pseudo code of the Project Problem

1. Gather a dataset of audio recordings that includes a variety of emotional states (e.g., happy, sad, angry, neutral).
2. Convert the audio signals into a suitable format (e.g., waveform).
3. Extract relevant features from each audio frame. Common features include:
    a. A metric of spectral properties is the mell-frequency cepstral coefficient (MFCC).
    b. Energy and zero-crossing rate: Indicate loudness and speech rate.
    c. Statistical features (mean, variance, skewness, kurtosis): Provide statistical information about the frame.
4. Normalize the feature values to have a similar scale to ensure that no feature dominates the others during classification.
5. Choose a machine learning model or deep learning architecture for emotion recognition. Common choices include:Decision Tree



    a. Convolutional Neural Networks (CNN)
    b. Recurrent Neural Networks (RNN)
    c. Long Short-Term Memory (LSTM) networks.

6. Train the selected model on the training data with corresponding emotional labels.
7. To get the most out of your model, tweak its hyperparameters using the validation set.
8. Analyze the model's efficacy on the validation set by calculating its confusion matrix, F1 score, or accuracy.
9. Use the trained model in a practical setting to identify and react to the speaker's emotional state. This might be done through a chatbot, virtual assistant, or customer support system.
10. Be sure to monitor on the model's performance and add fresh data as needed so it can adjust to different speech patterns and emotional expressions.

4.3 Flow graph of the Major Project Problem

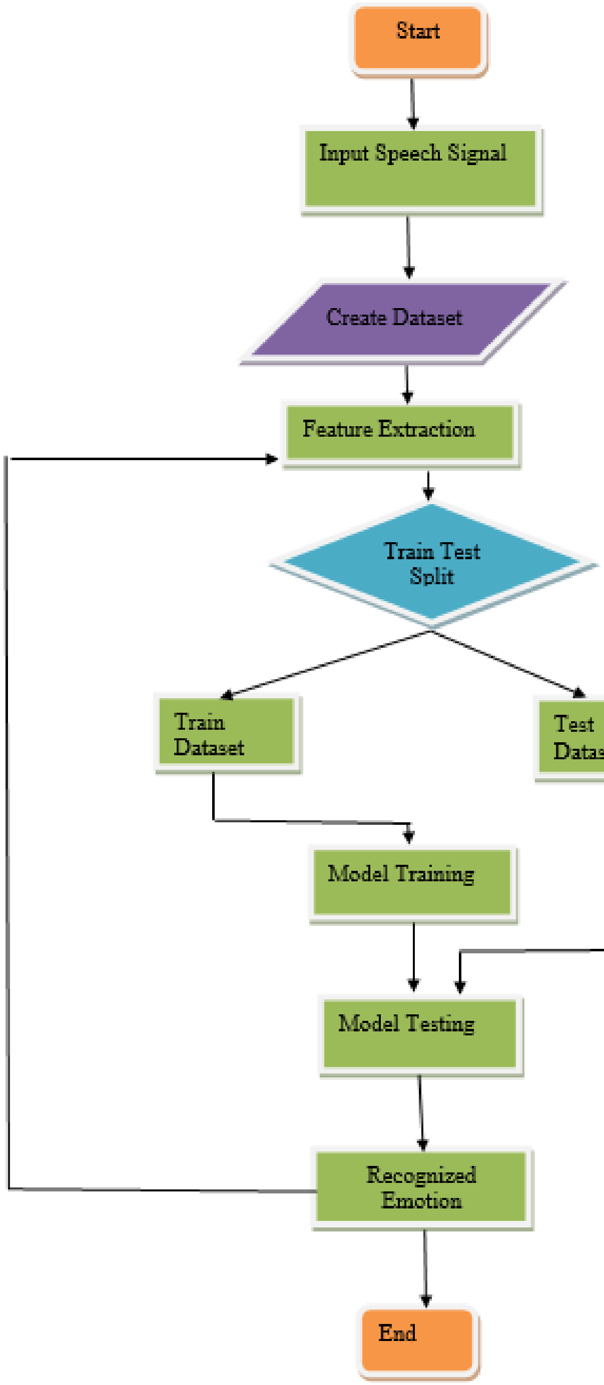


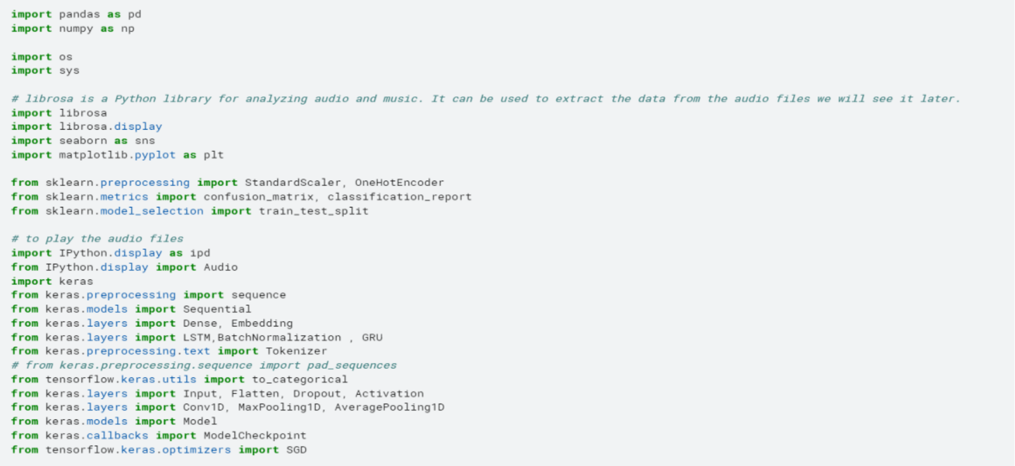
Figure 2 : Flow Graph of Major Project Problem

The flow graph of the Major Project Problem as described in Figure 2 lays down the framework of processing input, extracting features, training the model and finally evaluating its performance. Detailed description of entire project implementation with different stages of implementation is given in the subsequent section.

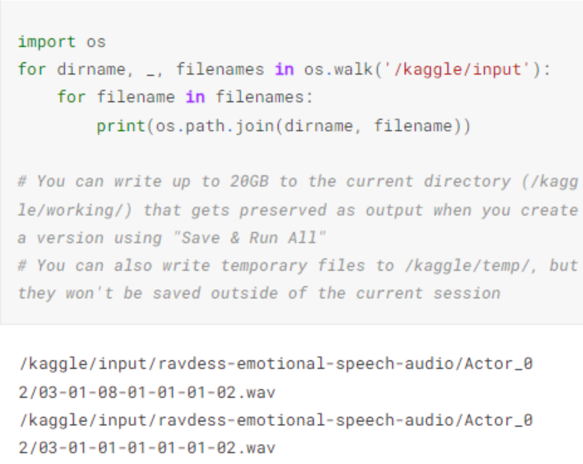
4.4 Screen shots of the various stages of the Project

Step 1: Importing necessary libraries and data collection

The necessary libraries are imported and then the dataset is downloaded from the Kaggle[6].



```
import pandas as pd
import numpy as np

import os
import sys

# librosa is a Python library for analyzing audio and music. It can be used to extract the data from the audio files we will see it later.
import librosa
import librosa.display
import seaborn as sns
import matplotlib.pyplot as plt

from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.metrics import confusion_matrix, classification_report
from sklearn.model_selection import train_test_split

# to play the audio files
import IPython.display as ipd
from IPython.display import Audio
import keras
from keras.preprocessing import sequence
from keras.models import Sequential
from keras.layers import Dense, Embedding
from keras.layers import LSTM,BatchNormalization , GRU
from keras.preprocessing.text import Tokenizer
# from keras.preprocessing.sequence import pad_sequences
from tensorflow.keras.utils import to_categorical
from keras.layers import Input, Flatten, Dropout, Activation
from keras.layers import Conv1D, MaxPooling1D, AveragePooling1D
from keras.models import Model
from keras.callbacks import ModelCheckpoint
from tensorflow.keras.optimizers import SGD
```



```
import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kagg
le/working/) that gets preserved as output when you create
a version using "Save & Run All"
# You can also write temporary files to /kaggle/temp/, but
they won't be saved outside of the current session
```

```
/kaggle/input/ravdess-emotional-speech-audio/Actor_0
2/03-01-08-01-01-01-02.wav
/kaggle/input/ravdess-emotional-speech-audio/Actor_0
2/03-01-01-01-01-01-02.wav
```

Code Snippet 1: Downloaded dataset.

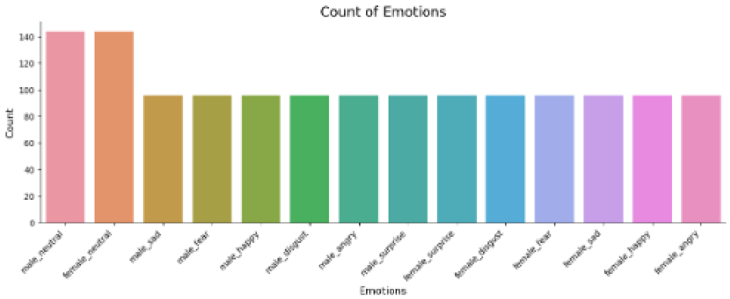


Figure 3: The distribution of recordings across different emotions in the RAVD_df dataset.

The dataset contains 7 emotions namely neutral, happy, sad, fear, disgust, surprise and angry with 2 different classes namely male and female. The distribution of these emotions is displayed in Figure 3.

Step 2: Data Augmentation

In data augmentation, we generate additional polymerized data samples by inserting tiny disturbances into our original training set. Audio polymerization may be achieved by the use of noise injection, time shifting, pitch and speed manipulation, and other similar techniques. The Figure 4 given below shows and compares Mel spectrogram of clean speech as well as augmented noise speech based on MFCC coefficients. Our goal is to improve our model's generalizability and make it resistant to such disturbances. The original training sample's label must be preserved when adding the disturbance for this to function.

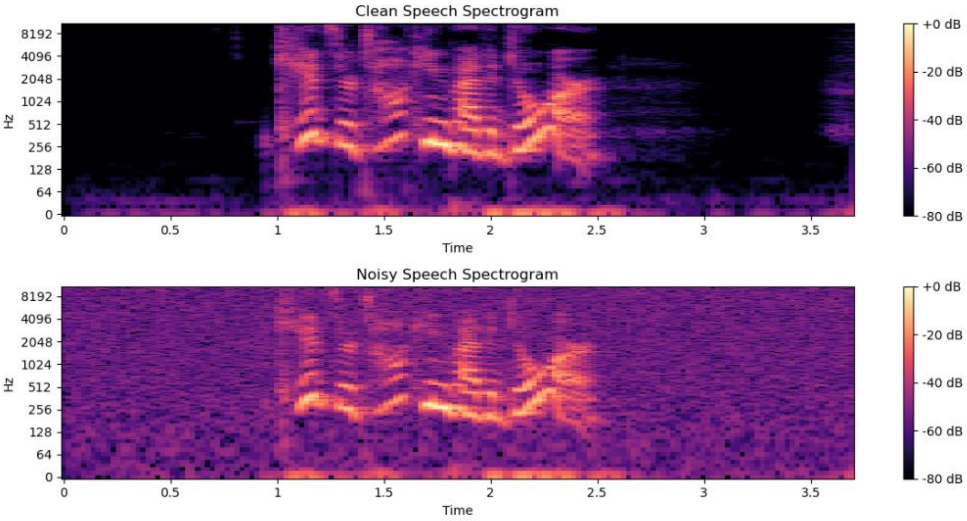


Figure 4: Mel Spectogram of Original and noisy signal

Step 3: Feature Extraction

Speech Emotion Recognition sometimes makes use of feature extraction using Mel-frequency cepstral coefficients (MFCC) over multiple time frames as represented by graph in Figure 5. The feature vectors that represent the speech signal are the MFCC coefficients that are produced and are stored in csv file as shown in Figure 6. Machine learning algorithms for emotion categorization are fed these characteristics in SER. Emotion analysis in speech is a good fit for MFCC-based features since they are resistant to changes in speaker traits and ambient conditions.

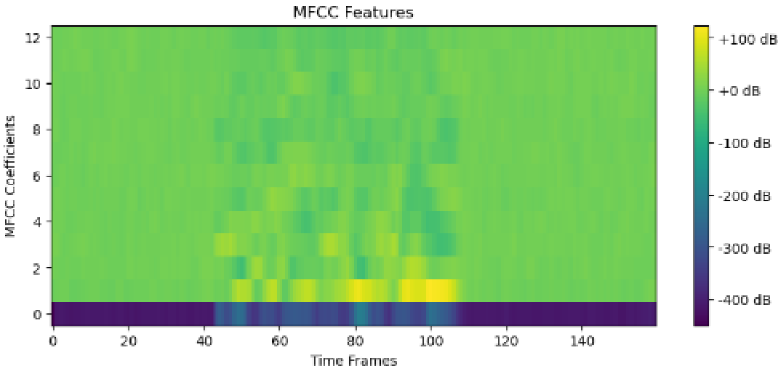
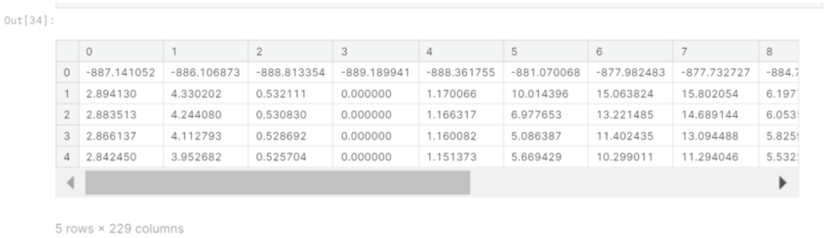


Figure 5: MFCC Feature Extraction



Out[34]:

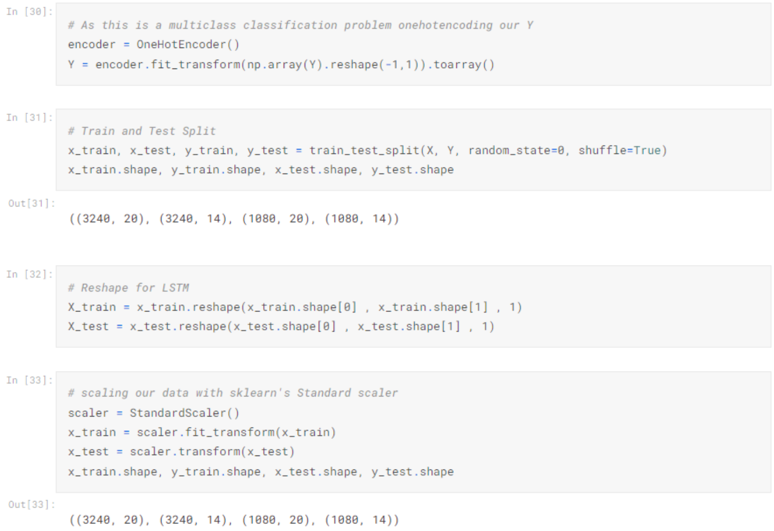
	0	1	2	3	4	5	6	7	8
0	-887.141052	-886.106873	-888.813354	-889.189941	-888.361755	-881.070068	-877.982483	-877.732727	-884.7
1	2.894130	4.330202	0.532111	0.000000	1.170066	10.014396	15.063824	15.802054	6.197
2	2.883513	4.244080	0.530830	0.000000	1.166317	6.977653	13.221485	14.689144	6.053
3	2.866137	4.112793	0.528692	0.000000	1.160082	5.086387	11.402435	13.094488	5.825
4	2.842450	3.952682	0.525704	0.000000	1.151373	5.669429	10.299011	11.294046	5.532

5 rows × 229 columns

Figure 6: Feature Extraction .csv file

Step 4: Data Preprocessing

Preparing the data for a classification task by converting categorical labels to a one-hot encoded format and standardizing the input features for improved model performance and the results are given in the given code snippet.

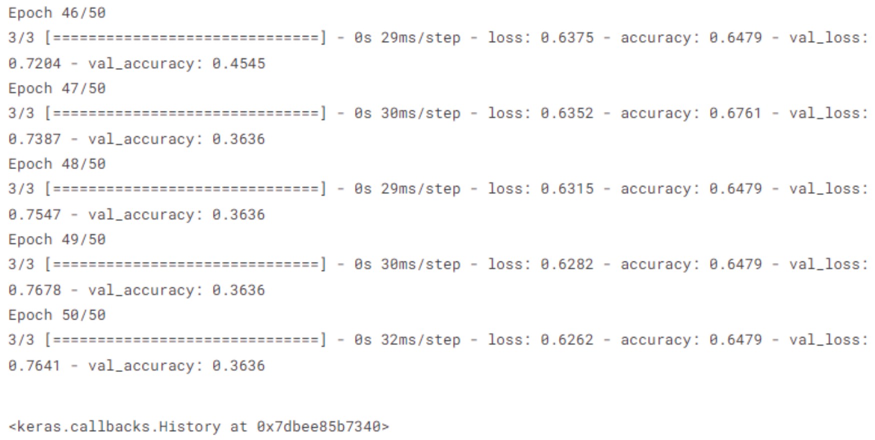


```
In [30]:
# As this is a multiclass classification problem onehotencoding our Y
encoder = OneHotEncoder()
Y = encoder.fit_transform(np.array(Y).reshape(-1,1)).toarray()

In [31]:
# Train and Test Split
x_train, x_test, y_train, y_test = train_test_split(X, Y, random_state=0, shuffle=True)
x_train.shape, y_train.shape, x_test.shape, y_test.shape

Out[31]:
((3240, 20), (3240, 14), (1080, 20), (1080, 14))

In [32]:
# Reshape for LSTM
X_train = x_train.reshape(x_train.shape[0] , x_train.shape[1] , 1)
X_test = x_test.reshape(x_test.shape[0] , x_test.shape[1] , 1)

In [33]:
# scaling our data with sklearn's Standard scaler
scaler = StandardScaler()
x_train = scaler.fit_transform(x_train)
x_test = scaler.transform(x_test)
x_train.shape, y_train.shape, x_test.shape, y_test.shape

Out[33]:
((3240, 20), (3240, 14), (1080, 20), (1080, 14))
```

Code Snippet 2: Data Preprocessing

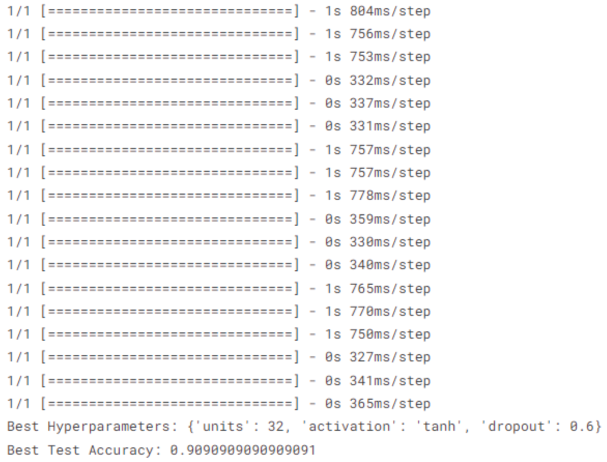
Step 5: Model Training and Hyperparameter Tuning

Speech Emotion Recognition (SER) was performed using LSTM, GRU, CNN models with hyperparameter tuning through grid search and manual search, achieving the best accuracy; additionally, a Decision Tree model was trained and tuned for comparison. Figure 7 shows training of GRU model and in order to optimize results hyperparameter tuning was performed to achieve higher accuracy as shown in Figure 8.



```
Epoch 46/50
3/3 [==============================] - 0s 29ms/step - loss: 0.6375 - accuracy: 0.6479 - val_loss:
0.7204 - val_accuracy: 0.4545
Epoch 47/50
3/3 [==============================] - 0s 30ms/step - loss: 0.6352 - accuracy: 0.6761 - val_loss:
0.7387 - val_accuracy: 0.3636
Epoch 48/50
3/3 [==============================] - 0s 29ms/step - loss: 0.6315 - accuracy: 0.6479 - val_loss:
0.7547 - val_accuracy: 0.3636
Epoch 49/50
3/3 [==============================] - 0s 30ms/step - loss: 0.6282 - accuracy: 0.6479 - val_loss:
0.7678 - val_accuracy: 0.3636
Epoch 50/50
3/3 [==============================] - 0s 32ms/step - loss: 0.6262 - accuracy: 0.6479 - val_loss:
0.7641 - val_accuracy: 0.3636

<keras.callbacks.History at 0x7dbee85b7340>
```

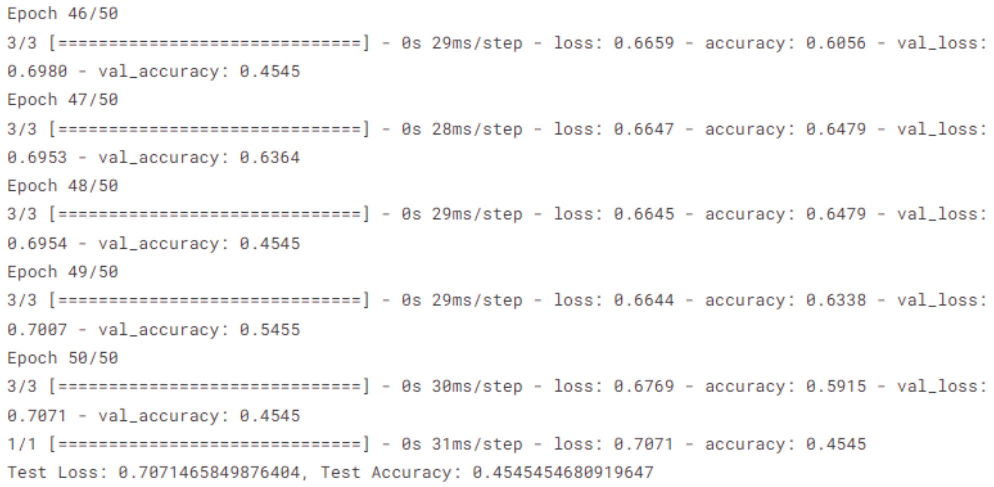
Figure 7: GRU Model Training



```
1/1 [==============================] - 1s 804ms/step
1/1 [==============================] - 1s 756ms/step
1/1 [==============================] - 1s 753ms/step
1/1 [==============================] - 0s 332ms/step
1/1 [==============================] - 0s 337ms/step
1/1 [==============================] - 0s 331ms/step
1/1 [==============================] - 1s 757ms/step
1/1 [==============================] - 1s 757ms/step
1/1 [==============================] - 1s 778ms/step
1/1 [==============================] - 0s 359ms/step
1/1 [==============================] - 0s 330ms/step
1/1 [==============================] - 0s 340ms/step
1/1 [==============================] - 1s 765ms/step
1/1 [==============================] - 1s 770ms/step
1/1 [==============================] - 1s 750ms/step
1/1 [==============================] - 0s 327ms/step
1/1 [==============================] - 0s 341ms/step
1/1 [==============================] - 0s 365ms/step
Best Hyperparameters: {'units': 32, 'activation': 'tanh', 'dropout': 0.6}
Best Test Accuracy: 0.9090909090909091
```

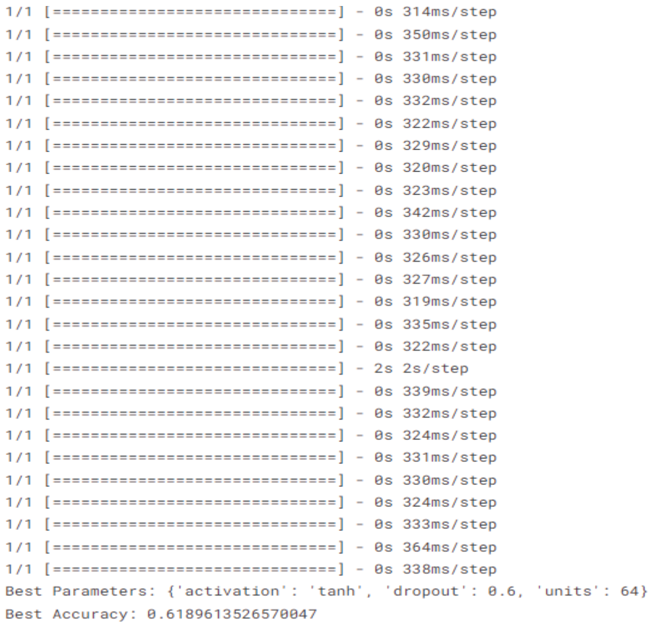
Figure 8: GRU Model Hyperparameter Tuning

Similarly, Figure 9 shows training of LSTM model and in order to optimize results hyperparameter tuning was performed to achieve higher accuracy as shown in Figure 10.



```
Epoch 46/50
3/3 [==============================] - 0s 29ms/step - loss: 0.6659 - accuracy: 0.6056 - val_loss:
0.6980 - val_accuracy: 0.4545
Epoch 47/50
3/3 [==============================] - 0s 28ms/step - loss: 0.6647 - accuracy: 0.6479 - val_loss:
0.6953 - val_accuracy: 0.6364
Epoch 48/50
3/3 [==============================] - 0s 29ms/step - loss: 0.6645 - accuracy: 0.6479 - val_loss:
0.6954 - val_accuracy: 0.4545
Epoch 49/50
3/3 [==============================] - 0s 29ms/step - loss: 0.6644 - accuracy: 0.6338 - val_loss:
0.7007 - val_accuracy: 0.5455
Epoch 50/50
3/3 [==============================] - 0s 30ms/step - loss: 0.6769 - accuracy: 0.5915 - val_loss:
0.7071 - val_accuracy: 0.4545
1/1 [==============================] - 0s 31ms/step - loss: 0.7071 - accuracy: 0.4545
Test Loss: 0.7071465849876404, Test Accuracy: 0.4545454680919647
```

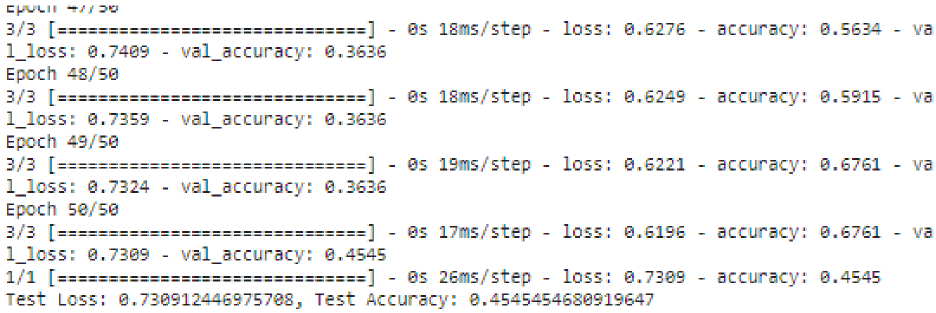
Figure 9: LSTM Model Training



```
1/1 [==============================] - 0s 314ms/step
1/1 [==============================] - 0s 350ms/step
1/1 [==============================] - 0s 331ms/step
1/1 [==============================] - 0s 330ms/step
1/1 [==============================] - 0s 332ms/step
1/1 [==============================] - 0s 322ms/step
1/1 [==============================] - 0s 329ms/step
1/1 [==============================] - 0s 320ms/step
1/1 [==============================] - 0s 323ms/step
1/1 [==============================] - 0s 342ms/step
1/1 [==============================] - 0s 330ms/step
1/1 [==============================] - 0s 326ms/step
1/1 [==============================] - 0s 327ms/step
1/1 [==============================] - 0s 319ms/step
1/1 [==============================] - 0s 335ms/step
1/1 [==============================] - 0s 322ms/step
1/1 [==============================] - 2s 2s/step
1/1 [==============================] - 0s 339ms/step
1/1 [==============================] - 0s 332ms/step
1/1 [==============================] - 0s 324ms/step
1/1 [==============================] - 0s 331ms/step
1/1 [==============================] - 0s 330ms/step
1/1 [==============================] - 0s 324ms/step
1/1 [==============================] - 0s 333ms/step
1/1 [==============================] - 0s 364ms/step
1/1 [==============================] - 0s 338ms/step
Best Parameters: {'activation': 'tanh', 'dropout': 0.6, 'units': 64}
Best Accuracy: 0.6189613526570047
```

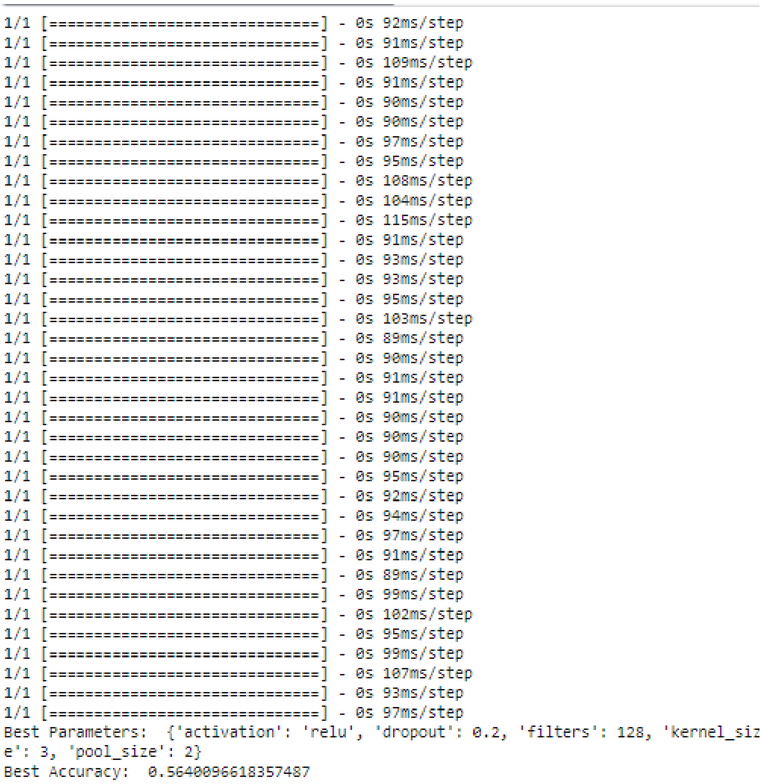
Figure 10: LSTM Model Hyperparamter Tuning

Similarly, Figure 11 shows training of CNN model and in order to optimize results hyperparameter tuning was performed to achieve higher accuracy as shown in Figure 12.



```
Epoch 47/50
3/3 [==============================] - 0s 18ms/step - loss: 0.6276 - accuracy: 0.5634 - va
l_loss: 0.7409 - val_accuracy: 0.3636
Epoch 48/50
3/3 [==============================] - 0s 18ms/step - loss: 0.6249 - accuracy: 0.5915 - va
l_loss: 0.7359 - val_accuracy: 0.3636
Epoch 49/50
3/3 [==============================] - 0s 19ms/step - loss: 0.6221 - accuracy: 0.6761 - va
l_loss: 0.7324 - val_accuracy: 0.3636
Epoch 50/50
3/3 [==============================] - 0s 17ms/step - loss: 0.6196 - accuracy: 0.6761 - va
l_loss: 0.7309 - val_accuracy: 0.4545
1/1 [==============================] - 0s 26ms/step - loss: 0.7309 - accuracy: 0.4545
Test Loss: 0.730912446975708, Test Accuracy: 0.4545454680919647
```

Figure 11: CNN Model Training

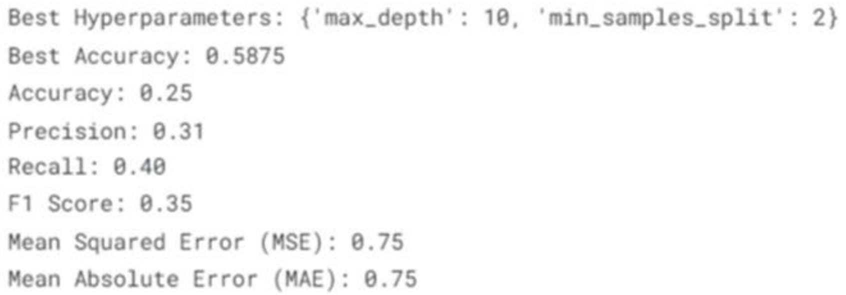


```
1/1 [==============================] - 0s 92ms/step
1/1 [==============================] - 0s 91ms/step
1/1 [==============================] - 0s 109ms/step
1/1 [==============================] - 0s 91ms/step
1/1 [==============================] - 0s 90ms/step
1/1 [==============================] - 0s 90ms/step
1/1 [==============================] - 0s 97ms/step
1/1 [==============================] - 0s 95ms/step
1/1 [==============================] - 0s 108ms/step
1/1 [==============================] - 0s 104ms/step
1/1 [==============================] - 0s 115ms/step
1/1 [==============================] - 0s 91ms/step
1/1 [==============================] - 0s 93ms/step
1/1 [==============================] - 0s 93ms/step
1/1 [==============================] - 0s 95ms/step
1/1 [==============================] - 0s 103ms/step
1/1 [==============================] - 0s 89ms/step
1/1 [==============================] - 0s 90ms/step
1/1 [==============================] - 0s 91ms/step
1/1 [==============================] - 0s 91ms/step
1/1 [==============================] - 0s 90ms/step
1/1 [==============================] - 0s 90ms/step
1/1 [==============================] - 0s 90ms/step
1/1 [==============================] - 0s 95ms/step
1/1 [==============================] - 0s 92ms/step
1/1 [==============================] - 0s 94ms/step
1/1 [==============================] - 0s 97ms/step
1/1 [==============================] - 0s 91ms/step
1/1 [==============================] - 0s 89ms/step
1/1 [==============================] - 0s 99ms/step
1/1 [==============================] - 0s 102ms/step
1/1 [==============================] - 0s 95ms/step
1/1 [==============================] - 0s 99ms/step
1/1 [==============================] - 0s 107ms/step
1/1 [==============================] - 0s 93ms/step
1/1 [==============================] - 0s 97ms/step
Best Parameters:  {'activation': 'relu', 'dropout': 0.2, 'filters': 128, 'kernel_siz
e': 3, 'pool_size': 2}
Best Accuracy:  0.5640096618357487
```

Figure 12: CNN Model Hyperparamter Tuning

Step 6: Model Evaluation

The Speech Emotion Recognition (SER) models, including LSTM, GRU and CNN neural networks along with a Decision Tree, were evaluated, showcasing their performance through accuracy and loss plots. Figure 13 lays down Decision Tree model's evaluation after hyperparameter tuning which is used as a comparison Machine Learning model with other Deep Learning models.



```
Best Hyperparameters: {'max_depth': 10, 'min_samples_split': 2}
Best Accuracy: 0.5875
Accuracy: 0.25
Precision: 0.31
Recall: 0.40
F1 Score: 0.35
Mean Squared Error (MSE): 0.75
Mean Absolute Error (MAE): 0.75
```

Figure 13: Decision Tree Model Evaluation

Given is Figure 14 which displays accuracy plot of LSTM model and GRU model. Both train and test accuracy are plotted. Following observations are made:

1. Train accuracy of LSTM model increases with every successive epoch whereas test accuracy of LSTM model fluctuates over the epochs.
2. Train accuracy of GRU model fluctuates and then becomes stable at the end of 30 epoch whereas test accuracy of GRU model fluctuates over the epochs.

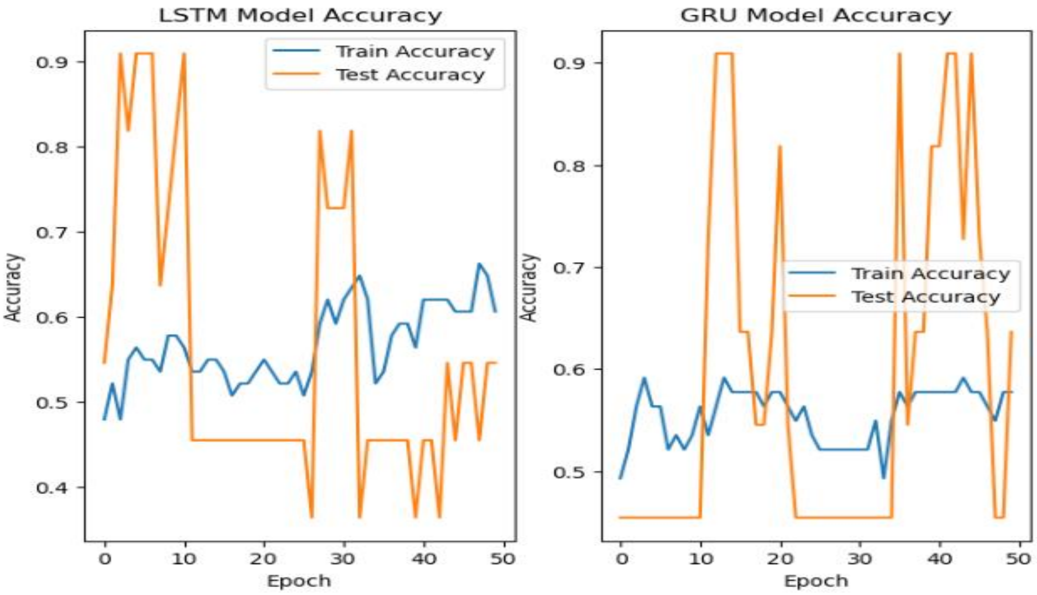


Figure 14: Accuracy Plot of LSTM model and GRU model

Figure 15 shows how train accuracy of CNN model fluctuates but increases with every successive epoch whereas test accuracy of CNN model only fluctuates over the epochs.

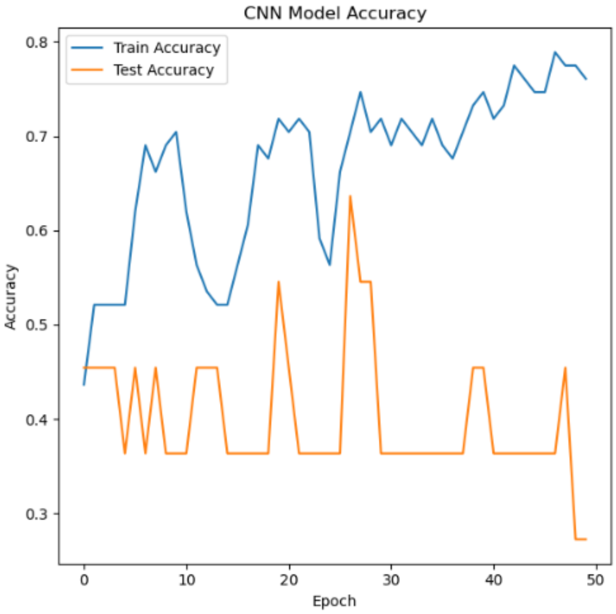


Figure 15: Accuracy Plot of CNN model

Given is Figure 16 which displays loss plot of LSTM model and GRU model. Both train and test loss are plotted. Following observations are made:

1. Train loss of LSTM model decreases with every successive epoch whereas test loss of LSTM model increases over the epochs.
2. Train loss of GRU model monotonically decreases and then becomes stable at the end of 30 epoch whereas test loss of GRU model fluctuates over the epochs but eventually decreases.

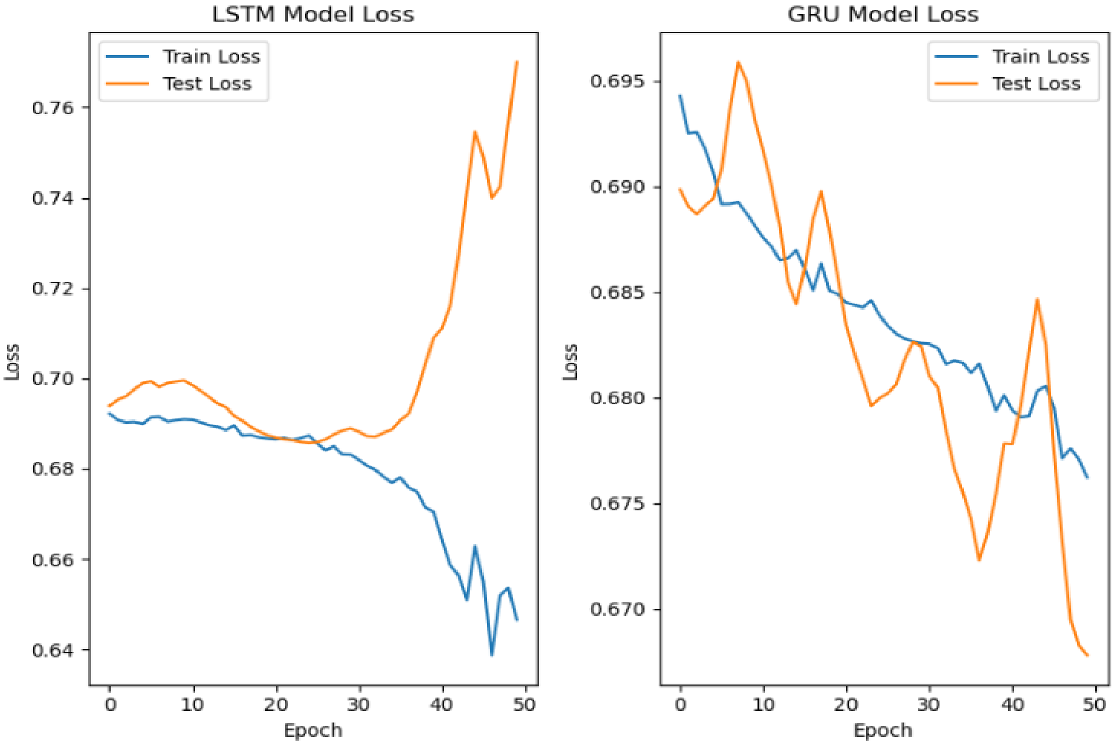


Figure 16: Loss Plot of LSTM model and GRU model

Figure 17 shows how train loss of CNN model monotonically decreases with every successive epoch whereas test accuracy of CNN model increases over the epochs.

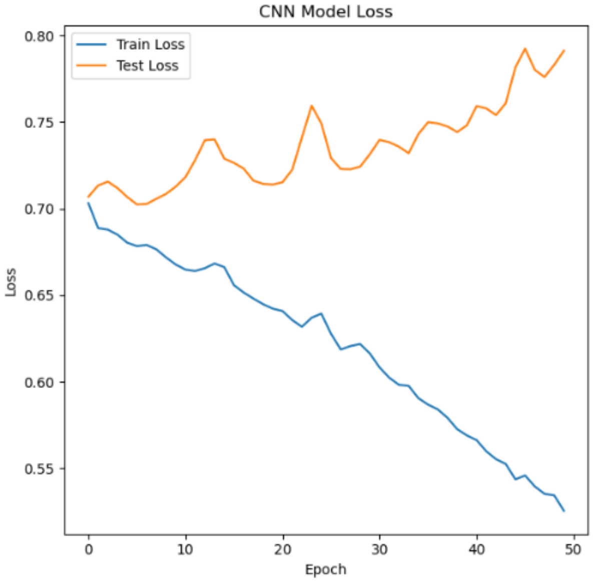


Figure 17: Loss Plot of CNN model

Step 7: Model Application

The Speech Emotion Recognition (SER) model with the best results is GRU (Gated Recurrent Unit) and we use this model for Live Demonstration on a random audio file to be preprocessed as depicted in Figure 18. This is taken from the dataset itself but it can also be recorded and uploaded for more testing.

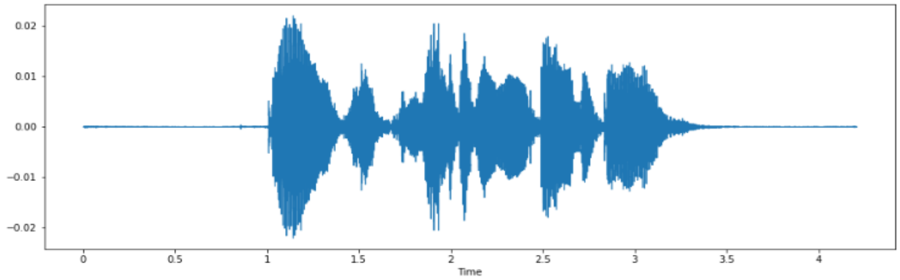
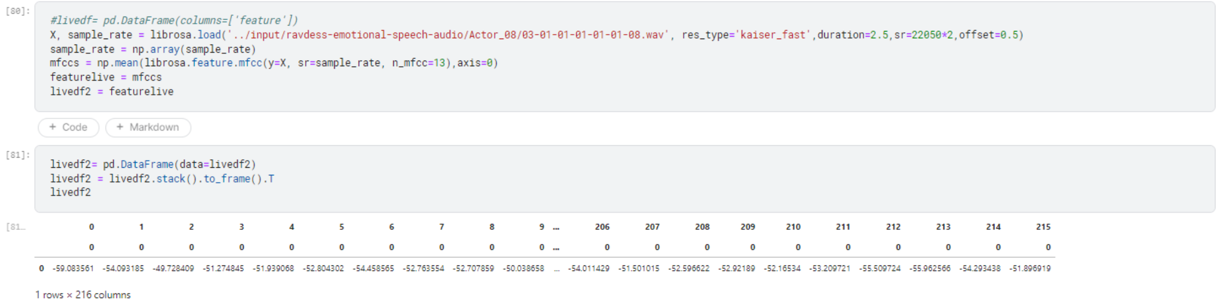


Figure 18: Waveform analysis of audio file to be preprocessed

The audio file is now preprocessed for sending as input to the saved model as depicted in the given code snippet.



```
[80]:
#livedf= pd.DataFrame(columns=['feature'])
X, sample_rate = librosa.load('../input/ravdess-emotional-speech-audio/Actor_08/03-01-01-01-01-01-08.wav', res_type='kaiser_fast',duration=2.5,sr=22050*2,offset=0.5)
sample_rate = np.array(sample_rate)
mfccs = np.mean(librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=13),axis=0)
featurelive = mfccs
livedf2 = featurelive
```

```
[81]:
livedf2= pd.DataFrame(data=livedf2)
livedf2 = livedf2.stack().to_frame().T
livedf2
```

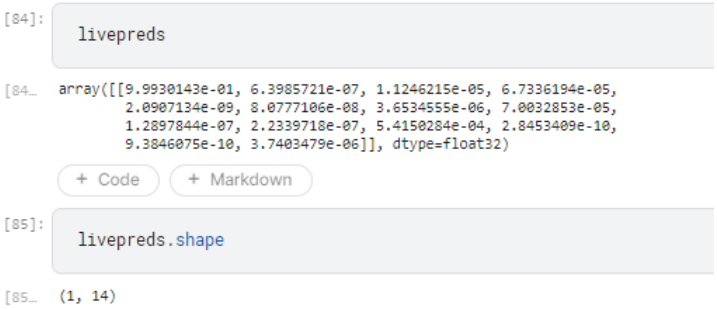
[81…

	0	1	2	3	4	5	6	7	8	9	...	206	207	208	209	210	211	212	213	214	215
	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	-59.083561	-54.093185	-49.728409	-51.274845	-51.939068	-52.804302	-54.458565	-52.763554	-52.707859	-50.038658	...	-54.011429	-51.501015	-52.596622	-52.92189	-52.16534	-53.209721	-55.509724	-55.962566	-54.293438	-51.896919

1 rows × 216 columns

Code Snippet 3: Audio file preprocessing

The audio file now is represented as array of extracted features as given in the following code snippet.



```
[84]:
livepreds
```

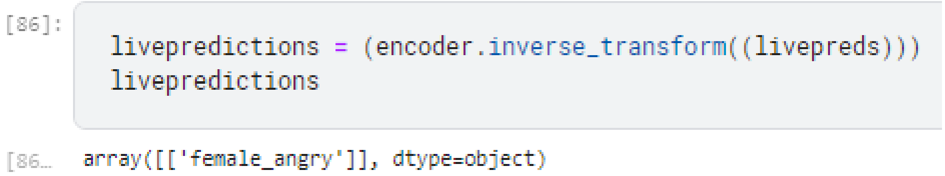
```
[84…
array([[9.9930143e-01, 6.3985721e-07, 1.1246215e-05, 6.7336194e-05,
        2.0907134e-09, 8.0777106e-08, 3.6534555e-06, 7.0032853e-05,
        1.2897844e-07, 2.2339718e-07, 5.4150284e-04, 2.8453409e-10,
        9.3846075e-10, 3.7403479e-06]], dtype=float32)
```

```
[85]:
livepreds.shape
```

```
[85…
(1, 14)
```

Code Snippet 4: Feature extraction

The model correctly classifies the audio as being that of a female and the emotion classified is angry as depicted in Figure 19.



```
[86]:  livepredictions = (encoder.inverse_transform((livepreds)))
       livepredictions

[86…  array([['female_angry']], dtype=object)
```

Figure 19: Final Output of Live Demonstration

Chapter 05: Results

5.1 Discussion on the Results Achieved

Our project "**Speech Emotion Recognition**" where we used 3 different models: Decision Tree, GRU, LSTM, CNN as the method to train our model. While still being trained, our model finds patterns and traits that it may use for identification. The provided table shows the performance results for the Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), Convolutional Neural Network(CNN) and Decision Tree models . The report is given below for each model.

Table 2: Performance Report

S.no.	Model	Train Loss	Validation Loss	Train Accuracy	Validation Accuracy	Best Accuracy
1.	Decision Tree	0.65	0.65	0.35	0.35	0.5875
2.	GRU	0.6631	0.6863	0.6056	0.3636	0.9090
3.	LSTM	0.6530	0.7789	0.6197	0.2727	0.6201
4.	CNN	0.6196	0.7309	0.6761	0.4545	0.564

General Observations:

1. The Decision Tree model has the lowest training and validation accuracy among the three models.
2. Both the GRU and LSTM models exhibit higher training accuracy but a significant drop in validation accuracy, suggesting overfitting.

3. The GRU model outperforms the other two models in terms of the best accuracy achieved.30
4. Regularization techniques and additional fine-tuning enhanced the overall performance of these models. Best Parameters for each model are given below:

a. Decision Tree: {'max_depth': 10, 'min_samples_split': 2}

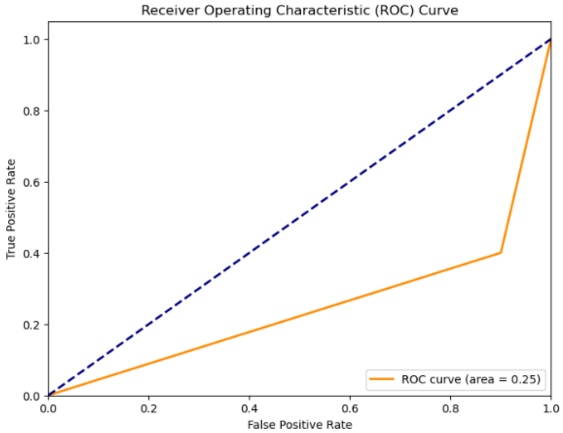
b. GRU: {'units': 32, 'activation': 'tanh', 'dropout': 0.2}

c. LSTM: {'activation': 'tanh', 'dropout': 0.2, 'units': 32}

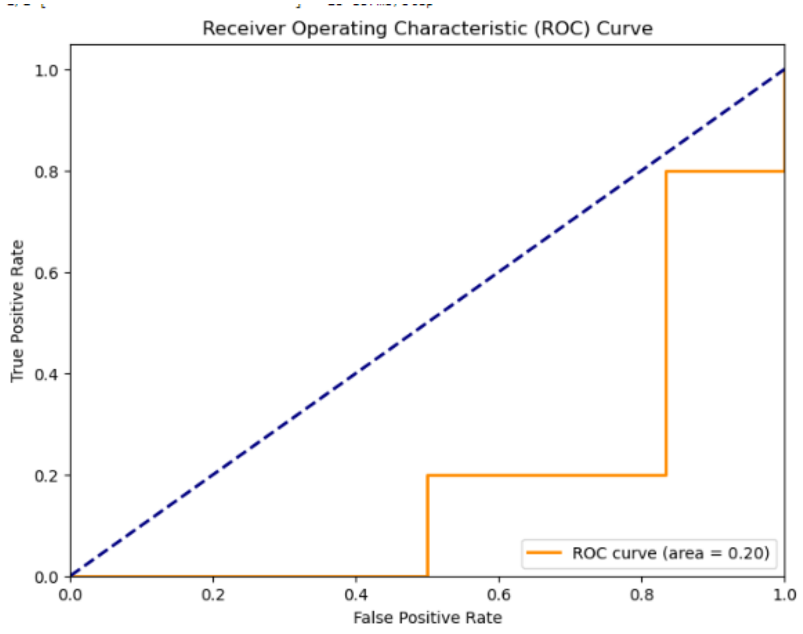
d. CNN: {‘activation’: ‘relu’, ‘dropout’: 0.2, ‘units’: 128}

Comparative Analysis using ROC:

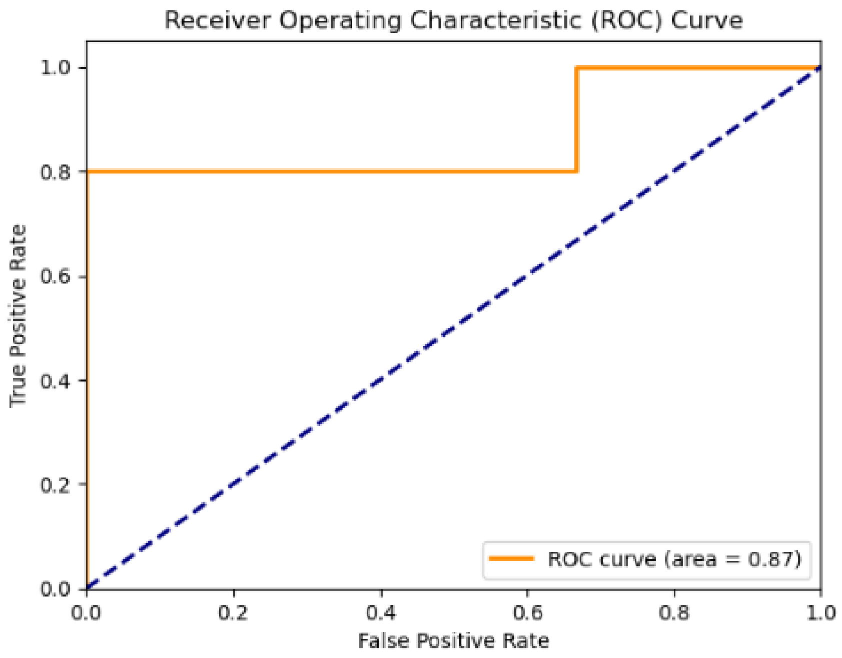
ROC curves of all the four models are plotted as shown in Figure 20 to perform comparative analysis. ROC or the Receiver Operating Characteristic Curve is a graph that depicts performance of classification models at all thresholds. ROC curve is basically tradeoff between True Positive Rate and False Positive Rate. If area under ROC curve is less than 0.5 denotes classifier is not working well, else it denotes classifier is making correct predictins more than half of the time.



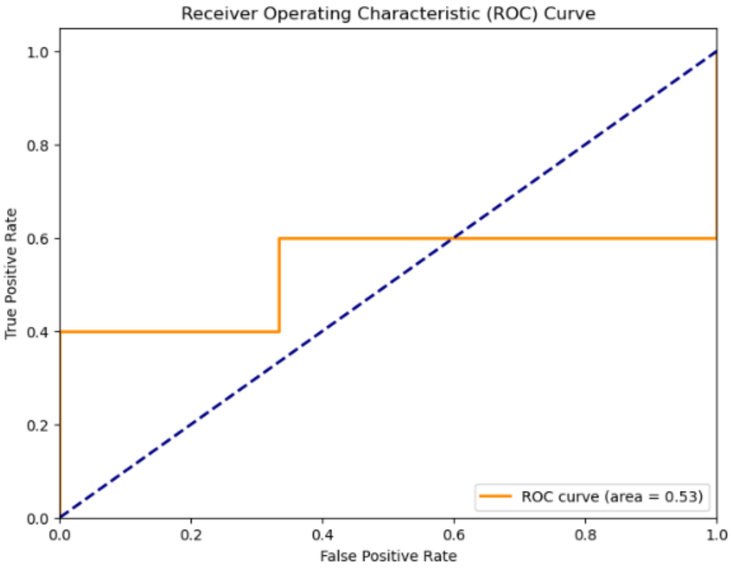
a.


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.20)

b.


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.87)

c.



d.

Figure 20: ROC Curve of a. Decision Tree b. LSTM c. GRU and d. CNN models

As Figure 20 demonstrates GRU model has the highest ROC area under curve value of 0.87 which makes it most suitable for our final demonstration on a random audio signal as made in previous chapter.

5.2 Application of the Major or Project

Speech Emotion Recognition (SER) has various applications across different domains due to its potential to enhance human-computer interaction, improve user experience, and provide valuable insights into emotional states. Here are some notable applications of Speech Emotion Recognition:

1. Human-Computer Interaction (HCI): SER can be incorporated in systems to make them more user-friendly to the emotions of the users. To illustrate, a virtual assistant could adjust its answers according to the user's emotional state.
2. Customer Service and Call Centers: SER is a tool that can be used in call centers to analyze customer emotions during calls. The approach will be used to collect this data and use it to serve our customers more effectively and understand their requirements.

3. Healthcare: Through the study of the way speech patterns can change which can be the indicators of the emotional discomfort or the changes of mood, SER will be able to assist in the tracking of the mental health problems.
4. Education: Studying may be made more individualized by considering each student's emotional condition with the help of SER, an educational technology resource. To illustrate, one can tell when they are feeling frustrated and also how to make a job more or less challenging.
5. Market Research: Businesses can gauge the emotional responses of consumers to ads and goods with the aid of SER. Revelations obtained from this data may be useful in the improvement of advertising campaigns.
6. Entertainment: Video games are an application of SER that makes the gaming experience more engaging and immersive by adjusting the setting and the characters according to the player's emotion.
7. Security: SER could be used in security applications, like lie detection systems, by evaluating speech patterns to detect the signs of stress or deception.
8. Automotive Industry: The integration of SER in cars enables the observing of drivers' emotional states that may be helpful in the prevention of accidents by the detection of signs of fatigue or tension.
9. Emotion-aware Devices: SER can be incorporated into wearable devices that help to track the emotional state of the user. The personalized suggestions or treatments that are based on the information collected from the wearer will be made.
10. Market Analysis and Sentiment Analysis: Through the analysis of the sentiment expressed in social media postings, comments, and reviews, SER may give an idea of the public's opinion about some goods, businesses or events.
11. Counseling and Therapy: Through SER's instant feedback about the person's emotional state, the virtual counselling or therapy sessions may be enhanced.
12. Accessibility: Individuals with disabilities may be the ones to gain from SER as it will help them to use their voice and emotion to interact with technology, thus making technology more accessible to them.

Speech emotion recognition has a lot of uses, and those uses will grow as technology does, opening up new opportunities for bettering human-machine communication and understanding in many levels and fields.

5.3 Limitation of the Project

Despite the potential utility of Speech Emotion Recognition (SER), there are a number of obstacles and restrictions that must be taken into account:

1. Subjectivity and Cultural Variability: Emotions are not the same for everyone and they might be shown in different ways by different people and cultures. A SER model's generalisation capability is based on how well it was trained on a certain culture or language group.
2. Context Dependency: Emotion recognition is mainly based on the surrounding environment. It is hard to correctly categorise emotions without considering the bigger context of the communication since the same speech signal might express more than one emotion depending on the context.
3. Ambiguity and Overlapping Emotions: A person's emotion can change suddenly or even combine with another. It may well be a bit of a problem to exactly recognize some emotional states from speech, especially when such feelings are complicated or mixed.
4. Limited Training Data: It is difficult to train SER models with the labelled information because existing datasets could not cover all the cultural and emotional differences. The outcomes might be distorted and not relevant to the larger population.
5. Speech Variability: The precision of SER models may be influenced by the differences in speech patterns which are caused by things like accent, speech rate and personal variations. It might be hard for the models that were trained only on one way of speaking to change to others.
6. Lack of Standardization: The validation of SER systems is not backed by any existing standards.

7. The assessment of different models' performance is a tough task since different researches may use different datasets, feature extraction techniques, and evaluation measures.

Real-time Processing: Most applications are based on real-time processing, but most of the SER models, especially on devices with limited resources, may not be able to deal with real-time situations due to their high computational requirements.

Privacy Concerns: The privacy problems that come with the emotion identification from speech are very real, especially in places like public areas or the workplace surveillance. The frequent analysis of people's emotional states can make them feel uneasy.

The obstacles in this area can only be eliminated by the continuous research and development in the field of speech emotion recognition. In order to make SER systems more useful and successful, we have to do some upgrades on data gathering, model robustness, cross-cultural adaptation, and ethical issues.

5.4 Future Work

There is a huge potential for future and ongoing research in the rapidly developing field of speech emotion recognition (SER). Some potential areas for future work include:

1. Cross-Cultural Adaptation: The main obstacle is the deficiency of cross-cultural and cross-lingual generalizability in SER models. The possible models that have been developed to work better in different cultural and language settings will be the topic of the future studies.
2. Multimodal Emotion Recognition: Combining information from different modes such as vocalizations, body language, physiological signals, and facial

expressions may be the way to enhance the emotion identification systems. To a greater extent, future studies may focus on multimodal techniques to understand the emotional states.

3. More research into complex deep learning architectures, like neural networks trained on sequential data or models based on transformers, may lead to improved SER performance. These layouts could be able to capture audio data with more complex connections.
4. Practical Use Cases: Virtual Assistants, Live Interactions, and Other Real-World Applications Require Real-Time Processing and Efficient SER Model Deployment at the Edge. Future study could focus on optimising models for speed and resource efficiency.
5. Improving emotion identification accuracy requires an understanding of the temporal dynamics of emotions as well as the incorporation of long-range relationships in voice data. Models that take the time context of emotional manifestations into account can be the subject of future studies.
6. Building SER models with the ability to learn and adjust with fresh data over time is crucial for staying relevant in ever-changing contexts. The use of continuous learning techniques could make it possible for models to adjust to new ways of speaking and other ways of expressing emotions.
7. Researching self-supervised or unsupervised learning approaches to train SER models could be useful, particularly in cases when there is a lack of labelled training data. These methods have the potential to pretrain models using unlabeled data and subsequently refine them using smaller labelled datasets.
8. Improving the robustness of SER models and addressing biases in them is of the utmost importance when it comes to using them with various demographic data. Future research should focus on developing methods to mitigate biases and ensure fair and unbiased emotion recognition across different populations.

Responsible and ethical advancement of Speech Emotion Recognition and the expansion of its functions depend on the continued cooperation of researchers, industry experts, and lawmakers.

References

[1] : R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, “Speech emotion recognition using deep learning techniques: A review,” IEEE access, vol. 7, pp. 117 327–117 345, 2019.

[2] A. Sherstinsky, “Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network,” Physica D: Nonlinear Phenomena, vol. 404, p. 132306, 2020.

[3] Y. Bengio, P. Simard, and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult,” IEEE Transactions on Neural Networks, vol. 5, no. 2, pp. 157–166, 1994.

[4] M. Li, L. Xie, Z. Lv, J. Li, and Z. Wang, “Multistep deep system for multimodal emotion detection with invalid data in the internet of things,” IEEE Access, vol. 8, pp. 187 208–187 221, 2020.

[5] H. F. Nweke, Y. W. Teh, M. A. Al-Garadi, and U. R. Alo, “Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges,” Expert Systems with Applications, vol. 105, pp. 233–261, 2018.

[6] Y. G. Thimmaraja, B. Nagaraja, and H. Jayanna, “Enhancements in encoded noisy speech data by background noise reduction,” Intelligent Systems with Applications, vol. 20, p. 200273, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667305323000984

[7] B. Sisman, J. Yamagishi, S. King, and H. Li, “An overview of voice con-

version and its challenges: From statistical modeling to deep learning," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 132–157, 2020.

[8] A.-L. Cˆırneanu, D. Popescu, and D. Iordache, "New trends in emotion recognition using image analysis by neural networks, a systematic review," Sensors, vol. 23, no. 16, p. 7092, 2023.

[9] M. Karim, S. Khalid, A. Aleryani, N. Tairan, Z. Ali, and F. Ali, "Hade: Exploiting human action recognition through fine-tuned deep learning methods," IEEE Access, vol. 12, pp. 42 769–42 790, 2024.

[10] L. Trinh Van, T. Dao Thi Le, T. Le Xuan, and E. Castelli, "Emotional speech recognition using deep neural networks," Sensors, vol. 22, no. 4, p. 1414, 2022.

[11] M. Farooq, F. Hussain, N. K. Baloch, F. R. Raja, H. Yu, and Y. B. Zikria, "Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network," Sensors, vol. 20, no. 21, p. 6008, 2020.

[12] Y. Cai, X. Li, and J. Li, "Emotion recognition using different sensors, emotion models, methods and datasets: A comprehensive review," Sensors, vol. 23, no. 5, p. 2455, 2023.

[13] A. G´eron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, 2019, [Accessed: March

26, 2024]. [Online]. Available: https://www.oreilly.com/library/view/hands-on-machine-learning/9781098125967/

[14] S. N. Livingstone and F. A. Russo, “The ryerson audio-visual database of emotional speech and song (ravdess),” https://zenodo.org/record/1188976#.YVusfI4zbOQ, 2018, [Accessed: March 26, 2024].