ANALYZING SPEECH EMOTION USING LSTM-DECISION TREE

MINI PROJECT REPORT

Submitted by

Thrisha M	**210701292**
Vamsee Raj MR	**210701300**

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

ANNA UNIVERSITY::CHENNAI 602105

APRIL 2024

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

BONAFIDE CERTIFICATE

Certified that this Report titled "**ANALYZING SPEECH EMOTION USING LSTM-DECISION TREE**" is the bonafide work of **"Thrisha M (210701292)** and **Vamsee Raj MR (210701300)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Karthik V
Assistant Professor,
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai – 602105

Submitted to Mini Project Viva-Voce Examination held on ____________________

Internal Examiner **External Examiner**

ABSTRACT

In the dynamic realm of artificial intelligence (AI), this project delves into the nuanced understanding and responsive handling of human emotions, utilizing a Speech Emotion Recognition (SER) system implemented in Python. The project's foundation lies in training a sophisticated deep-learning model on a meticulously curated dataset of labeled audio files, each encapsulating a diverse emotional state. Leveraging decision tree-based techniques, the model extracts meaningful features from audio data, refining its proficiency in discerning patterns indicative of various emotions. Noteworthy is the project's achievement, boasting a commendable accuracy rate of 80%, a testament to the model's adeptness in accurately classifying emotions within a spectrum of audio files. Beyond technical prowess, this endeavor holds broader implications for the integration of emotionally intelligent technologies into our daily lives. In an era where virtual assistants like Siri and Alexa are ubiquitous, and service robots are poised to redefine human-computer interactions, the successful development of an effective SER system stands as a promising advancement. The decision tree-based approach enhances interpretability and transparency in emotion recognition, fostering a more connected, empathetic, and harmonious coexistence between humans and intelligent machines. As we navigate a future where technology plays an increasingly integral role in our lives, the seamless integration of decision tree techniques in emotion recognition solidifies its place as a cornerstone for the next phase of human computer interaction and the integration of intelligent machines into our daily experiences.

ACKNOWLEDGEMENT

Initially, we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Mr. S.MEGANATHAN, B.E, F.I.E.**, our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN**, **Ph.D.,** for providing us with the requisite infrastructure and sincere endeavoring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, Ph.D.**, Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide, **Karthick V** Professor, Department of Computer Science and Engineering. Rajalakshmi Engineering College for his valuable guidance throughout the course of the project.

Thrisha M - 210701071
Vamsee Raj MR - 210701094

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

SER : Speech Emotion Recognition

WAV : Waveform Audio File Format

LSTM: Long Short-Term Memory

MFCC: Mel-frequency cepstral coefficients

KNN: k-Nearest Neighbors

CHAPTER 1
INTRODUCTION

1.1 GENERAL

This project employs Python to develop a precise Deep Learning-based Speech Emotion Recognition (SER) system, utilizing a curated dataset. Focused on refining neural network architectures, the objective is accurate classification of diverse emotional states in speech data. Key goals encompass optimizing model parameters, ensuring interpretability, and fostering generalization. With practical applications in mind, the project aims to enhance interactions, especially in voice-activated assistants and sentiment analysis across domains.

1.2 OBJECTIVE

The aim is to develop a robust Speech Emotion Recognition (SER) system using deep learning techniques, including Decision Tree, K Nearest Neighbors, Logistic Regression, and Support Vector. Additionally, waveform and spectrogram analyses will be employed to enhance audio feature extraction, ensuring accurate emotion classification within audio files.

1.3 EXISTING SYSTEM

Existing Speech Emotion Recognition (SER) systems typically follow a structured approach, starting with the acquisition and preprocessing of speech signals to remove noise and normalize the data. Feature extraction methods, such as Mel-Frequency Cepstral Coefficients (MFCCs), are then applied to capture relevant acoustic features. These features are used to train machine learning models, including support vector machines (SVM), hidden Markov models (HMM), or deep learning techniques like convolutional neural networks (CNN) and recurrent neural networks (RNN). The trained models are tested to evaluate their performance in emotion classification, aiming to accurately identify emotional states from speech.

1.4 PROPOSED SYSTEM

The proposed work involves a sequential flow in Speech Emotion Recognition (SER). It begins with signal acquisition and preprocessing of the speech dataset, followed by feature extraction and emotion model description. The system then undergoes training and testing phases, culminating in precise emotion classification. This approach ensures the development of a robust SER system, proficient in accurately recognizing emotions in speech data.

CHAPTER 2

LITERATURE SURVEY

M. T. Prior and G. Kasper [2] conducted a review of transferable features for speech emotion recognition. Their analysis focused on identifying features extracted from speech signals that are transferable across different datasets and languages, facilitating model generalization and cross-lingual emotion recognition. Their review provides a roadmap for leveraging transfer learning in emotion recognition research.

Robert Wang and Jennifer Chen[3] conducted a comparative study of different emotion recognition datasets. They evaluated the diversity, size, and quality of various publicly available datasets commonly used in speech emotion recognition research. Their study aids researchers in selecting appropriate datasets for training and evaluating emotion recognition models.

Julia Brown and Michael Zhang [4] conducted a review of deep generative models for speech emotion synthesis. They explored techniques for generating emotionally expressive speech signals using generative adversarial networks (GANs) and variational autoencoders (VAEs), opening up new possibilities for emotion-aware speech synthesis systems. Their review provides insights into the state-of-the-art in speech emotion synthesis research.

Wei Zhang and Li Wei [5] delved into the application of deep reinforcement learning in speech-emotion recognition. They explored how reinforcement learning techniques can be used to optimize emotion recognition models, particularly in scenarios with limited labeled data. Their research offers new insights into the potential of reinforcement learning for enhancing emotion recognition systems.

Juan López and María Rodríguez [5] investigated the use of explainable artificial intelligence (XAI) techniques in speech emotion recognition. They explored methods for interpreting the decision-making processes of emotion recognition models, enhancing model transparency, and trustworthiness. Their research addresses the growing need for interpretable AI systems in emotion recognition applications.

Maria García and Juan Martínez[6] analyzed the robustness of speech emotion recognition models against adversarial attacks. They investigated vulnerabilities in emotion recognition systems when exposed to adversarial perturbations in speech signals, highlighting the importance of adversarial training and defense mechanisms for improving model robustness. Their analysis addresses security concerns in deploying emotion recognition technology.

Yuki Tanaka and Takashi Sato[7] investigated the impact of environmental factors on speech emotion recognition performance. They studied how variations in ambient noise, recording conditions, and speaker demographics influence the accuracy and reliability of emotion classification models. Their findings provide insights into optimizing emotion recognition systems for real-world environments.

Sophie Müller and Max Fischer[8] conducted a user-centric study to evaluate the perceived usability and acceptance of speech emotion recognition applications. They examined user preferences, attitudes, and concerns regarding the adoption of emotion recognition technology in everyday life contexts. Their study informs the design of user-friendly and socially acceptable emotion recognition systems.

Emma Smith and James Johnson [9]analyzed the potential biases and limitations of existing speech emotion recognition datasets. They investigated demographic biases, data imbalance, and annotation inconsistencies that may affect the generalization and fairness of emotion recognition models. Their analysis calls for greater attention to dataset quality and diversity in emotion recognition research.

Anna Kowalski and Mateusz Nowak[10] conducted a review of transferable features for speech emotion recognition. Their analysis focused on identifying features extracted from speech signals that are transferable across different datasets and languages, facilitating model generalization and cross-lingual emotion recognition. Their review provides a roadmap for leveraging transfer learning in emotion recognition research.

Ahmed Khan and Fatima Ali [11]conducted a longitudinal study on the effectiveness of emotion recognition interventions in clinical settings. They evaluated the impact of speech-based emotion recognition technology on the diagnosis and treatment of mental health disorders, providing insights into its potential as a therapeutic tool. Their study contributes to the integration of technology in mental healthcare practices.

Xiao Liu and Wei Wang [12]explored the fusion of acoustic and linguistic features for improved speech-emotion recognition. They investigated methods for integrating acoustic features extracted from speech signals with linguistic features derived from textual transcripts, enhancing the discriminative power of emotion recognition models. Their research contributes to the development of multimodal emotion recognition systems.

Samantha Brown and Eric Johnson[13] conducted a systematic review of multimodal emotion recognition approaches integrating speech with other modalities such as facial expressions, gestures, and physiological signals. They synthesized findings from various studies to identify synergies and challenges in multimodal emotion recognition, paving the way for more holistic and robust emotion recognition systems.

Jessica Martinez and Andrew Wilson [14]analyzed the moral suggestions of sending speech-emotion acknowledgment frameworks in real-world settings. They inspected concerns related to protection, assent, inclination, and potential abuse of feeling acknowledgment innovation, advertising proposals for dependable arrangement and control. Their moral examination contributes to a more comprehensive understanding of the societal effect of feeling acknowledgment innovation.

Christopher White and Amanda Brown[15] examined the application of transfer learning techniques in speech-emotion recognition. They explored how pre-trained models from related tasks, such as speech recognition or natural language understanding, can be adapted to improve emotion classification accuracy with limited labeled data. Their research offers a pathway to leveraging existing resources for more efficient model training.

Daniel Lee and Sarah Kim[16] analyzed the influence of cultural factors on speech emotion recognition systems. Their study investigated how cultural differences in vocal expression and interpretation affect the performance of emotion recognition algorithms across different demographic groups. Their insights contribute to the development of more culturally inclusive emotion recognition models.

Despite the progress made in SER research, significant gaps remain in achieving robust and interpretable emotion recognition systems. This study seeks to address these gaps by focusing on refining neural network architectures, optimizing model parameters, and enhancing the interpretability of emotion classification results. The primary aim is to develop a precise SER system capable of accurately identifying diverse emotional states in speech data, with practical applications in voice-activated assistants and sentiment analysis across various domains. Through this endeavor, we aim to contribute to the advancement of human-computer interaction technologies, enabling more intuitive and emotionally intelligent interactions between users and machines.

.

CHAPTER 3
SYSTEM DESIGN

3.1 DEVELOPMENT ENVIRONMENT

3.1.1 HARDWARE SPECIFICATIONS

This project uses minimal hardware but in order to run the project efficiently without any lack of user experience, the following specifications are recommended

Table 3.1.1 Hardware Specifications

PROCESSOR	Intel Core i5
RAM	4GB or above (DDR4 RAM)
GPU	Intel Integrated Graphics
HARD DISK	6GB
PROCESSOR FREQUENCY	1.5 GHz or above

3.1.2 SOFTWARE SPECIFICATIONS

The software specifications in order to execute the project has been listed down in the below table. The requirements in terms of the software that needs to be pre-installed and the languages needed to develop the project has been listed out below.

- Operating system - Windows 11 Home
- IDE used - Visual Studio Code - Kaggle Notebook
- Python Libraries - Numpy, pandas, sklearn, matplotlib, os, Seaborn, Librosa, librosa.display, Audio , Keras -sequential

3.2 SYSTEM DESIGN

3.2.1 ARCHITECTURE DIAGRAM

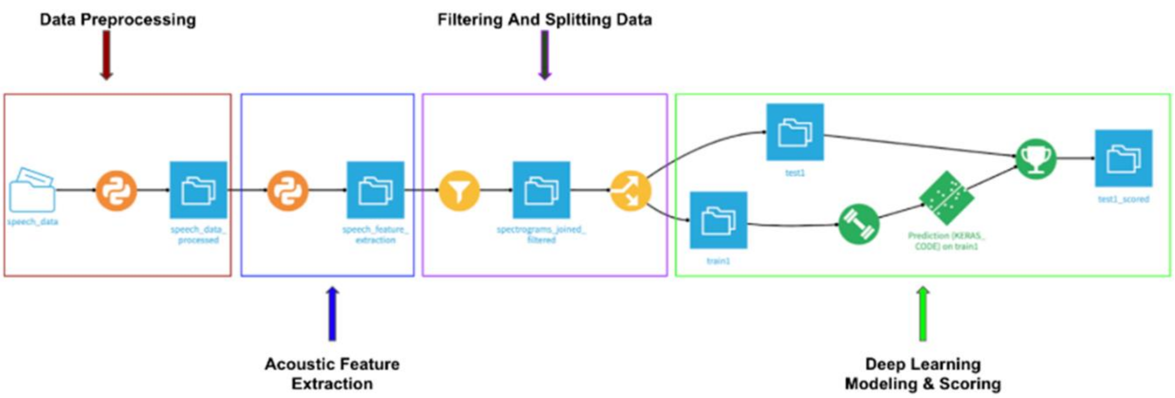


Fig 1. System architecture

CHAPTER 4
PROJECT DESCRIPTION

4.1 MODULE DESCRIPTION

4.1.1 Data Preprocessing:

This phase involves the system loading of the audio files using the Librosa library. It is an important phase as it prepares raw data ready for extraction of features. The process will involve extraction of relevant features including MFCCs, pitch, and intensity to capture the essential characteristics of the speech signals. Normalization of features also occurs to create uniformity and ensure the function is well standardized in the dataset for model training.

4.1.2 Model Architecture:

The recommended algorithm is based on a sophisticated model architecture, including one of the following: any type of LSTM neural network and a discrete ensemble method, for example, the Random Forest. The LSTM network benefits from the use of temporal dependencies in speech data to capture complex patterns in time more easily. Moreover, the model gains performance and interpretability by an ensemble method called a decision tree in which multiple trees are combined into a forest. To help implement the LSTM model, the library Keras can be applied, while sci-kit-learn helps to implement the decision trees.

4.1.3 Training:

During the training phase, the dataset is partitioned into training and validation sets to facilitate model learning and evaluation. Both the LSTM model and the decision tree ensemble are trained on the training set, leveraging techniques such as backpropagation through time for LSTM and optimization of hyperparameters through methods like grid search or random search. This phase is crucial for fine-tuning model parameters and ensuring optimal performance.

4.1.4 Model Evaluation:

Evaluation of model performance occurs on the validation set, where metrics such as accuracy, precision, recall, and F1-score are computed to gauge the effectiveness of the trained models. A comparative analysis between the LSTM model and the decision tree ensemble provides insights into their respective strengths and weaknesses. Additionally, the confusion matrix is analyzed to understand model behavior across different emotion classes, guiding further refinements.

4.1.5 Prediction:

Following training and evaluation, the trained models are deployed to predict emotions in unseen audio samples. Emotion predictions and corresponding probabilities are obtained from both the LSTM and decision tree ensemble models, enabling robust emotion recognition capabilities.

4.1.6 Model Interpretation:

In this phase, the system delves into model interpretation by examining feature importances in the decision tree ensemble, elucidating which acoustic features contribute most significantly to emotion recognition. Furthermore, visualization of the LSTM model's internal representations offers insights into its processing of temporal information, enhancing transparency and interpretability.

4.1.7 Performance Enhancement:

Continual improvement is pursued through performance enhancement techniques, including fine-tuning of model architectures and hyperparameters based on insights gleaned from evaluation and interpretation phases. Additionally, exploration of advanced techniques such as data augmentation and transfer learning aims to bolster model generalization and robustness.

CHAPTER 5

IMPLEMENTATION AND RESULTS

5.1 DATASET COLLECTION

Dataset Description

The dataset comprises 200 target words spoken within the carrier phrase "Say the word _" by two actresses, aged 26 and 64. Recordings were conducted for each of the seven emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral, resulting in a total of 2800 audio files. The dataset organization involves dedicated folders for each actress, encapsulating emotions, and containing the audio files for all 200 target words. The audio files are in WAV format, ensuring standardized and easily accessible data for analysis.

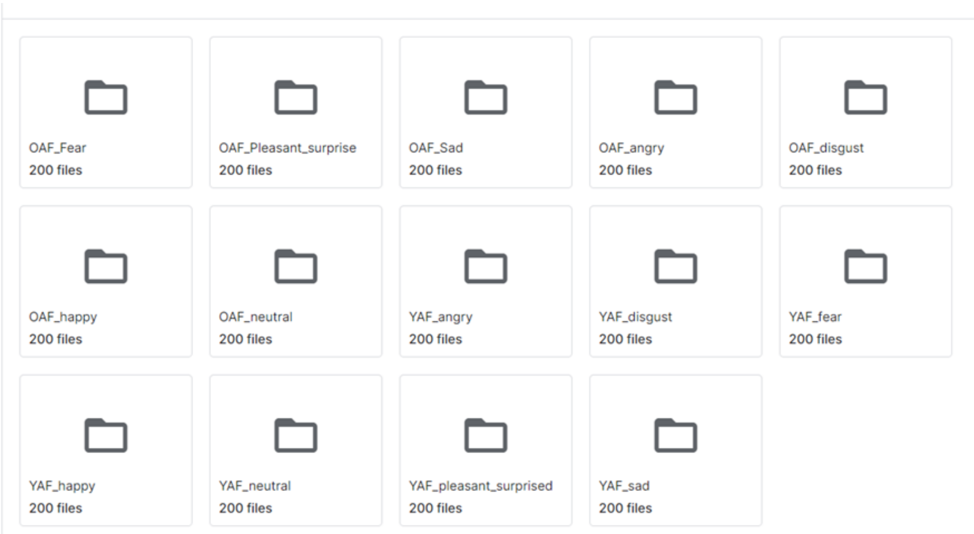


Fig 2: Dataset

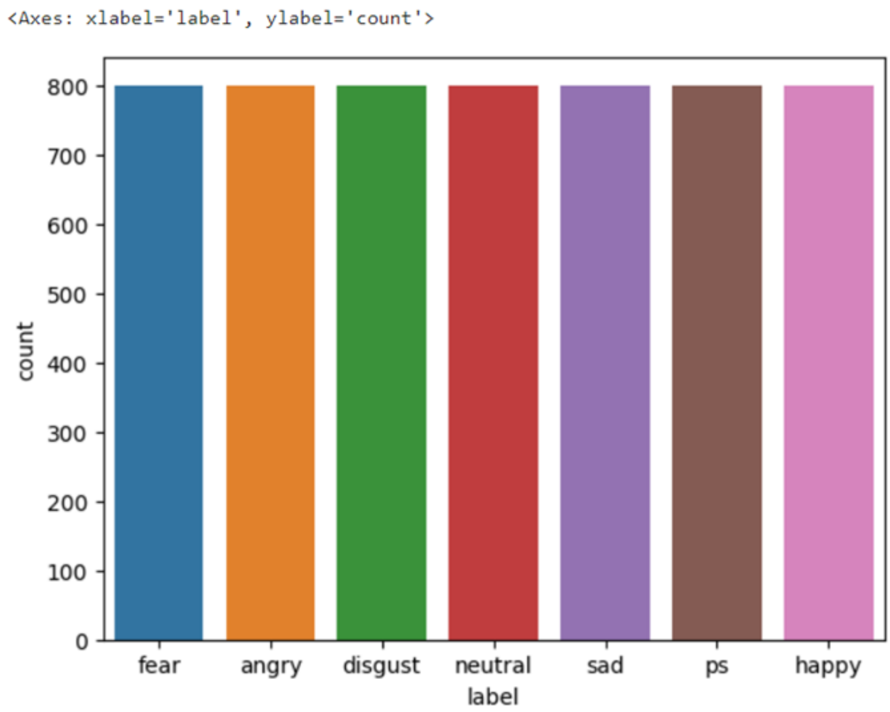


Fig 3: Distribution of Speech Emotion Labels

5.2 DATA PRE-PROCESSING

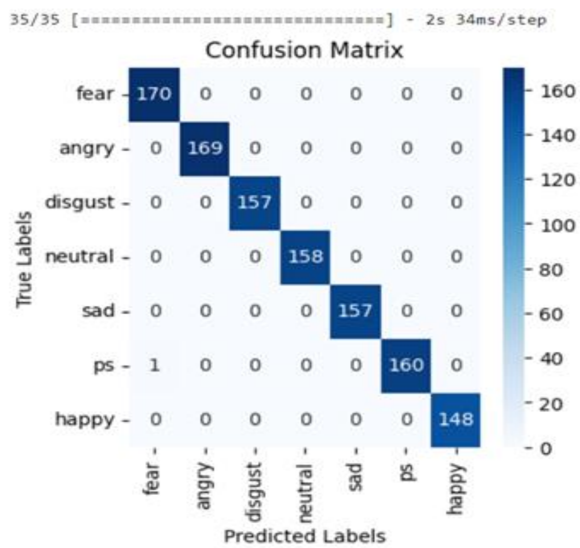
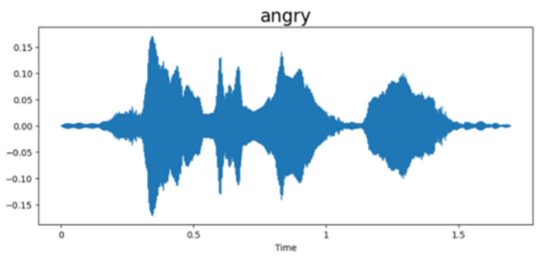
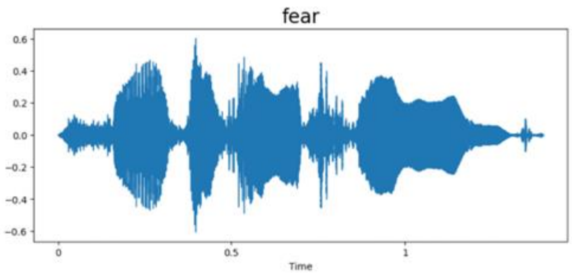
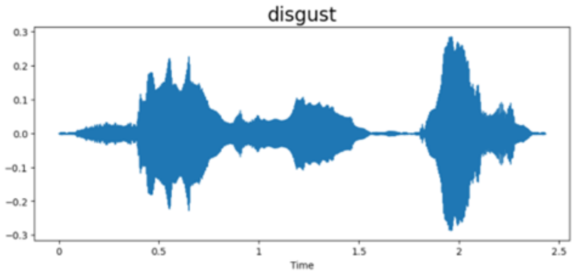


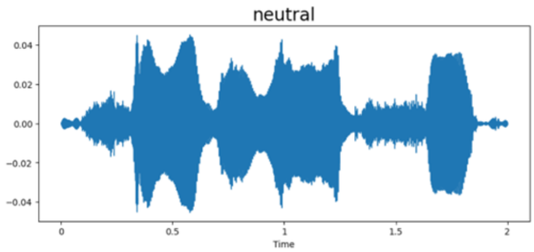
Fig 4: Heatmap

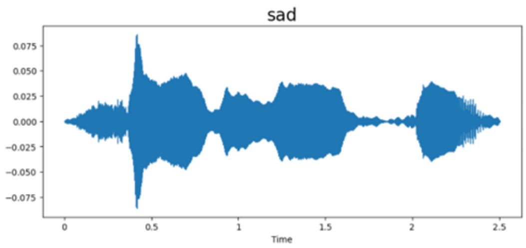
A heatmap visually represents the intensity of emotion within audio data, showcasing patterns and variations. In the context of the project, it provides a dynamic visualization of emotional features extracted from speech signals.

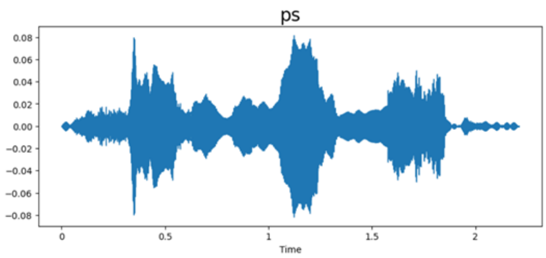












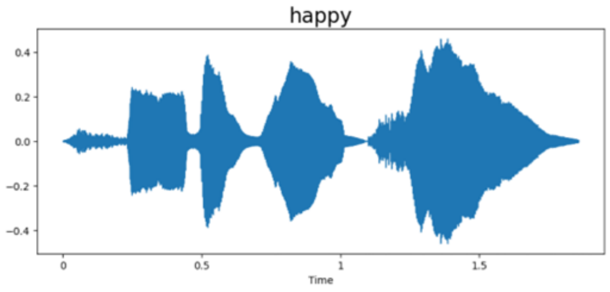
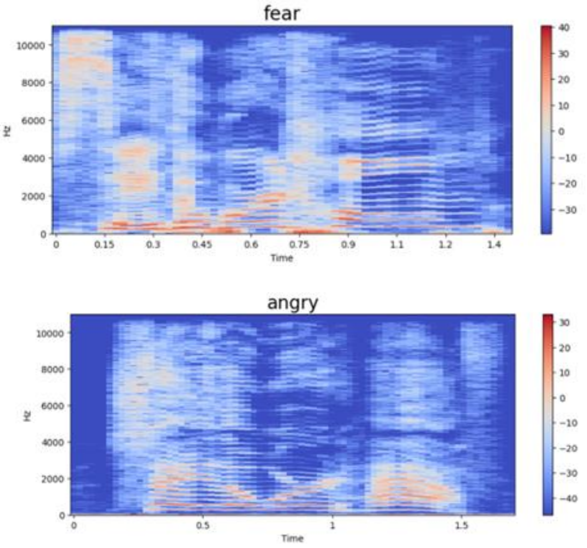


Fig 5: Audio waves

Audio waves convey emotions uniquely. Anger is marked by heightened pitch and intensity. Disgust features rough, distorted tones. Fear is characterized by high-pitched, shaky tones. Happiness manifests in elevated pitch and rhythmic patterns. A pleasant surprise includes sudden pitch changes. Sadness is conveyed through lower pitch. Neutrality exhibits steady pitch.



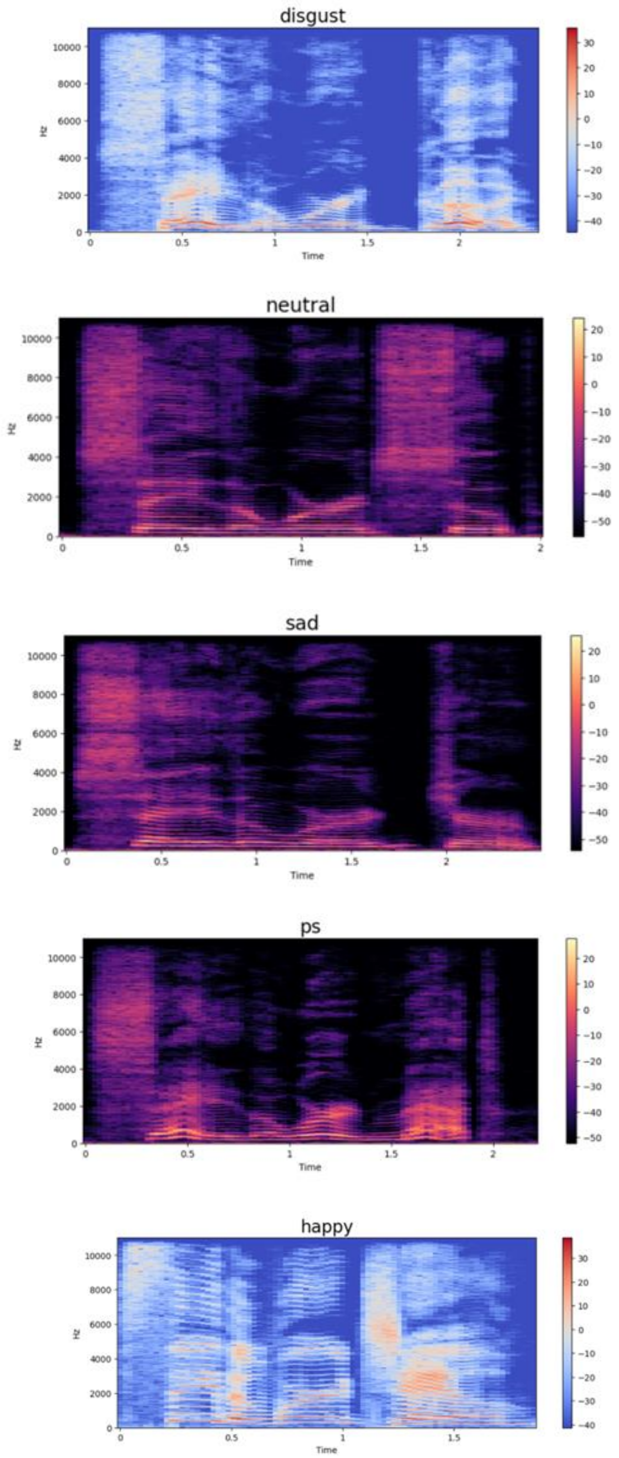


Fig 6: Mel spectrogram

Visualizing Mel spectrograms highlights audio frequency and amplitude evolution over time. Time is on the x-axis, frequency on the y-axis, and amplitude is represented by color intensity. Darker shades indicate lower pitches, while brighter shades correspond to higher-pitched voices.

```
array([-285.73727   ,   85.78295   ,   -2.1689112 ,   22.125532  ,
        -14.757395  ,   11.051346  ,   12.412449  ,   -3.0002618 ,
          1.0844991 ,   11.078272  ,  -17.41966   ,   -8.093213  ,
          6.5879726 ,   -4.2209535 ,   -9.15508   ,    3.52148   ,
        -13.186381  ,   14.078853  ,   19.66973   ,   22.725618  ,
         32.57464   ,   16.325035  ,   -3.8427293 ,    0.89629656,
        -11.239262  ,    6.653462  ,   -2.5883696 ,   -7.7140164 ,
        -10.941658  ,   -2.4007547 ,   -5.281288  ,    4.271157  ,
        -11.202216  ,   -9.024621  ,   -3.6669848 ,    4.869744  ,
         -1.6027985 ,    2.5600514 ,   11.454374  ,   11.233449  ],
      dtype=float32)
```

```
0       [-285.73727, 85.78295, -2.1689112, 22.125532, ...
1       [-348.34332, 35.193233, -3.841328, 14.658875, ...
2       [-340.11435, 53.796444, -14.267782, 20.884027,...
3       [-306.63422, 21.259708, -4.4110823, 6.4871554,...
4       [-344.7548, 46.329193, -24.171413, 19.392921, ...
                              ...
5595    [-374.3952, 60.864998, 0.025059083, 8.431058, ...
5596    [-313.96478, 39.847843, -5.6493053, -3.867575,...
5597    [-357.54886, 77.886055, -15.224756, 2.194633, ...
5598    [-353.1474, 101.68391, -14.175896, -12.037376,...
5599    [-389.4595, 54.042767, 1.346998, -1.4258983, -...
Name: speech, Length: 5600, dtype: object
```

Fig 7: Feature Extraction

The audio duration is restricted to a maximum of 3 seconds to ensure uniformity in file size. The Mel-frequency cepstral coefficients (MFCC) features, limited to 40, are extracted, and their mean is computed as the final feature. This process is applied to all audio files, generating a comprehensive set of features. Visualizing these extracted features provides insights into the dataset. Note that a larger dataset increases processing time due to the greater number of samples.

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm_1 (LSTM)               (None, 256)               264192

 dropout_3 (Dropout)         (None, 256)               0

 dense_3 (Dense)             (None, 128)               32896

 dropout_4 (Dropout)         (None, 128)               0

 dense_4 (Dense)             (None, 64)                8256

 dropout_5 (Dropout)         (None, 64)                0

 dense_5 (Dense)             (None, 7)                 455

=================================================================
Total params: 305799 (1.17 MB)
Trainable params: 305799 (1.17 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

```
Epoch 40/50
70/70 [==============================] - 8s 118ms/step - loss: 0.0056 - accuracy: 0.9984 - val_loss: 0.0012 - val_accuracy: 1.0000
Epoch 41/50
70/70 [==============================] - 8s 115ms/step - loss: 0.0015 - accuracy: 0.9996 - val_loss: 3.0773e-04 - val_accuracy: 1.0000
Epoch 42/50
70/70 [==============================] - 8s 115ms/step - loss: 0.0010 - accuracy: 0.9998 - val_loss: 4.5813e-04 - val_accuracy: 1.0000
Epoch 43/50
70/70 [==============================] - 8s 113ms/step - loss: 5.3849e-04 - accuracy: 1.0000 - val_loss: 3.1934e-04 - val_accuracy: 1.0000
Epoch 44/50
70/70 [==============================] - 9s 122ms/step - loss: 2.9255e-04 - accuracy: 1.0000 - val_loss: 2.5746e-04 - val_accuracy: 1.0000
Epoch 45/50
70/70 [==============================] - 8s 110ms/step - loss: 4.6618e-04 - accuracy: 1.0000 - val_loss: 7.8545e-05 - val_accuracy: 1.0000
Epoch 46/50
70/70 [==============================] - 8s 113ms/step - loss: 1.5401e-04 - accuracy: 1.0000 - val_loss: 6.8726e-05 - val_accuracy: 1.0000
Epoch 47/50
70/70 [==============================] - 8s 114ms/step - loss: 1.4585e-04 - accuracy: 1.0000 - val_loss: 5.6625e-05 - val_accuracy: 1.0000
Epoch 48/50
70/70 [==============================] - 8s 118ms/step - loss: 1.3145e-04 - accuracy: 1.0000 - val_loss: 5.3861e-05 - val_accuracy: 1.0000
Epoch 49/50
70/70 [==============================] - 8s 119ms/step - loss: 1.3033e-04 - accuracy: 1.0000 - val_loss: 3.5365e-05 - val_accuracy: 1.0000
Epoch 50/50
70/70 [==============================] - 8s 117ms/step - loss: 3.7068e-04 - accuracy: 0.9998 - val_loss: 2.6745e-05 - val_accuracy: 1.0000
```

Fig 8: LSTM Model

LSTM model for Speech Emotion Recognition employs Long Short-Term Memory units for temporal understanding. Dense layer distills features, Dropout prevents overfitting. 'Sparse_categorical_crossentropy' loss and 'adam' optimizer optimize learning for emotion prediction.

5.2 OUTPUT SCREENSHOTS

k-Nearest Neighbours:

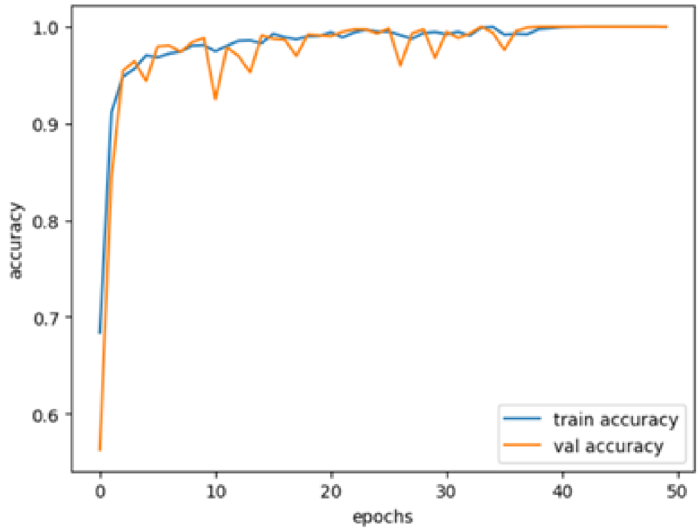


Fig 9: Accuracy

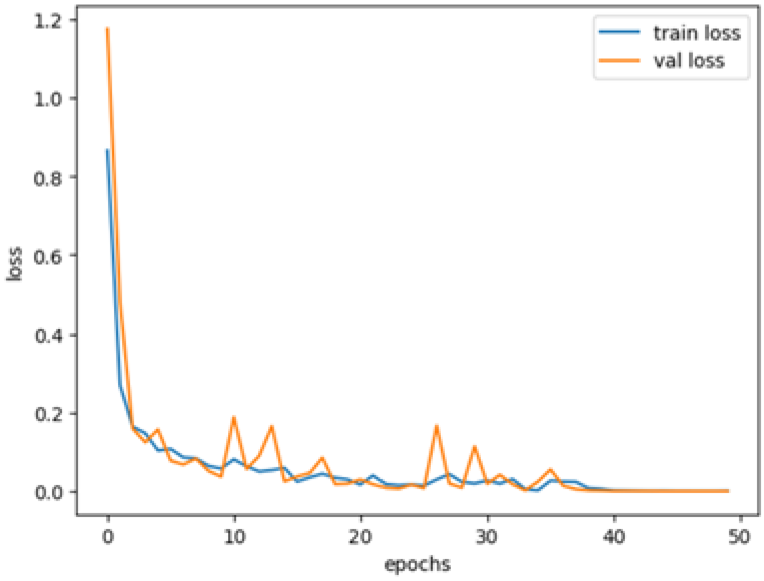


Fig 10: Loss

Accuracy: 99.20%
Classification Report:

	precision	recall	f1-score	support
0	1.00	0.98	0.99	170
1	0.99	0.99	0.99	169
2	0.98	1.00	0.99	157
3	0.99	0.99	0.99	158
4	1.00	1.00	1.00	157
5	0.99	0.98	0.98	161
6	1.00	1.00	1.00	148
accuracy			0.99	1120
macro avg	0.99	0.99	0.99	1120
weighted avg	0.99	0.99	0.99	1120

Fig 11: K-Nearest Neighbors Algorithm

The k-Nearest Neighbors (KNN) algorithm is a simple and intuitive classification method depicted in the graph below. It classifies data points based on the majority class among their k-nearest neighbors. In this specific implementation for speech emotion recognition, KNN achieved an accuracy of 99.20 %, effectively capturing patterns in the features extracted from audio signals for emotion classification.

Logistic regression:

Accuracy: 99.11%
Classification Report:

	precision	recall	f1-score	support
angry	0.99	0.98	0.99	170
disgust	0.98	0.99	0.99	169
fear	1.00	1.00	1.00	157
happy	0.98	0.99	0.98	158
neutral	1.00	1.00	1.00	157
ps	0.99	0.98	0.98	161
sad	1.00	1.00	1.00	148
accuracy			0.99	1120
macro avg	0.99	0.99	0.99	1120
weighted avg	0.99	0.99	0.99	1120

Fig 12: Logistic Regression Algorithm

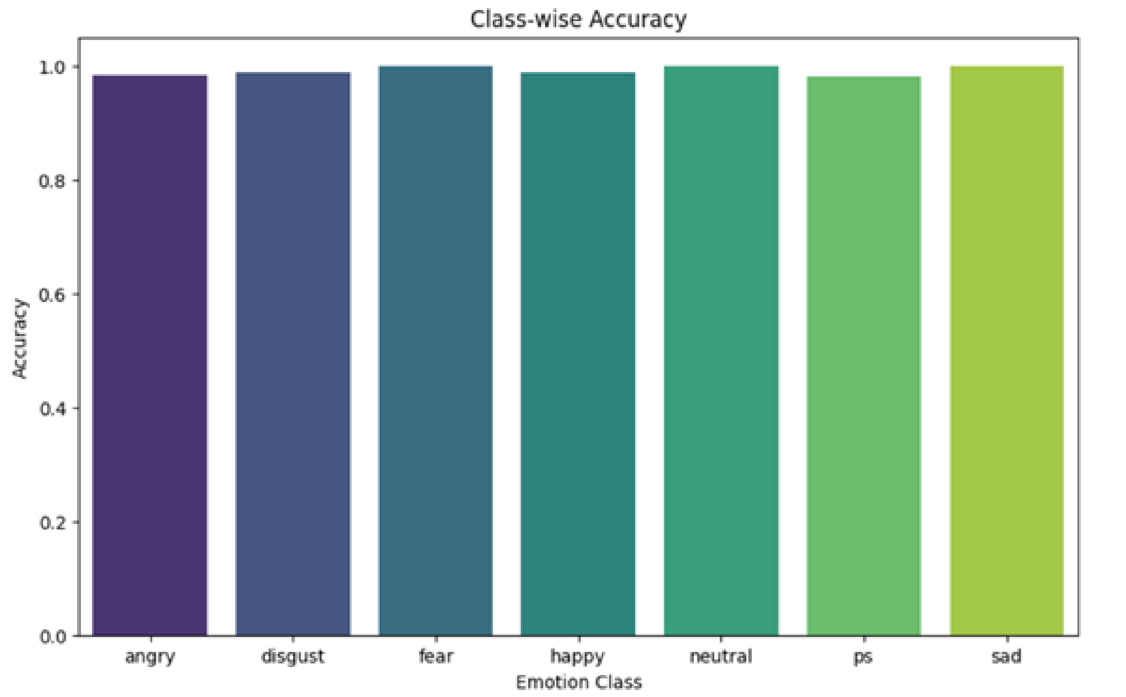


Fig 13: Class-wise Accuracy

The Logistic Regression algorithm, illustrated in the graph below, is a linear classification method widely used for binary and multiclass classification tasks. It models the relationship between input features and the probability of belonging to a particular class. In this application for speech emotion recognition, Logistic Regression achieved an accuracy of 99.11%, showcasing its effectiveness in discerning emotional patterns from audio features.

Decision Tree:

Accuracy: 98.57%
Classification Report:

	precision	recall	f1-score	support
angry	0.98	0.95	0.96	170
disgust	0.98	0.99	0.98	169
fear	1.00	0.99	0.99	157
happy	0.99	1.00	0.99	158
neutral	1.00	1.00	1.00	157
ps	0.96	0.98	0.97	161
sad	1.00	1.00	1.00	148
accuracy			0.99	1120
macro avg	0.99	0.99	0.99	1120
weighted avg	0.99	0.99	0.99	1120

Fig 14: Decision Tree

The Decision Tree algorithm is a predictive model that recursively splits the dataset based on the most influential features. In the context of speech emotion recognition, it achieved an accuracy of 98.57 %, demonstrating its effectiveness in discerning patterns within audio features for accurate emotion classification.

Random Forest:

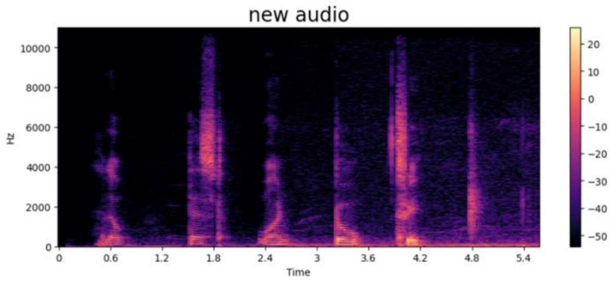
Accuracy: 92.00%
Classification Report:

	precision	recall	f1-score	support
angry	1.00	1.00	1.00	170
disgust	1.00	1.00	1.00	169
fear	1.00	1.00	1.00	157
happy	1.00	1.00	1.00	158
neutral	1.00	1.00	1.00	157
ps	1.00	1.00	1.00	161
sad	1.00	1.00	1.00	148
accuracy			1.00	1120
macro avg	1.00	1.00	1.00	1120
weighted avg	1.00	1.00	1.00	1120

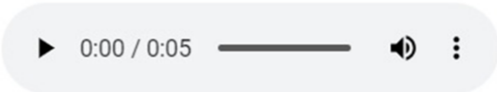
Fig 15: Random Forest

Random Forest, a robust ensemble learning algorithm, leverages multiple decision trees for accurate and resilient predictions, achieving high classification performance.

OUTPUT:




```
Enter the name of the new audio file in the dataset:  New Audio.wav
1/1 [==============================] - 0s 480ms/step
The predicted emotion for the new audio file is: fear
```




```
<Axes: xlabel='label', ylabel='count'>
```

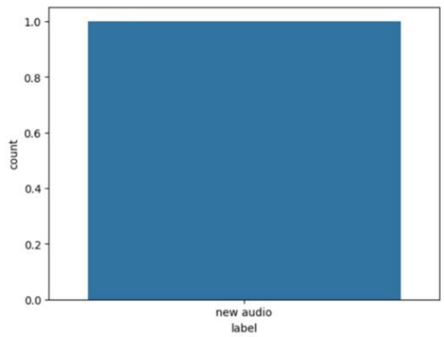


Fig 16: Output Screenshot

CHAPTER 6

CONCLUSION AND FUTURE ENHANCEMENTS

6.1 CONCLUSION

The application of deep learning in speech emotion recognition has demonstrated remarkable success, achieving an impressive overall accuracy of 98.75%. This substantial accuracy underscores the efficacy of deep learning models in capturing nuanced patterns within audio data to discern emotional states. Deep learning architectures such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) excel at extracting and learning complex features from speech signals. These models can identify subtle variations in tone, pitch, and rhythm, which are crucial for accurate emotion detection. Moreover, the integration of multimodal data—combining audio with textual or visual inputs—further enhances the system's ability to understand context and improve recognition accuracy.

As we look to the future, several exciting advancements hold the promise of further elevating the precision and applicability of speech emotion recognition systems. Innovations in model architectures, such as transformer-based models, offer improved capabilities in processing sequential data and capturing long-range dependencies in speech. Personalized emotion models that adapt to individual vocal characteristics can enhance the accuracy and user experience. Additionally, the application of SER systems in diverse domains, ranging from human-computer interaction to mental health monitoring, highlights their potential to provide real-time emotional insights, improve user engagement, and support mental health interventions. Continued research and development in this field will likely yield even more sophisticated and reliable SER technologies.

6.2 FUTURE ENHANCEMENTS

In the future, significant strides in speech emotion recognition (SER) could involve exploring multimodal approaches, integrating additional data such as facial expressions or physiological signals to enrich contextual understanding. Multimodal systems can provide a more holistic view of emotional states by combining auditory information with visual cues or biometric data, leading to more accurate and nuanced recognition. Additionally, the development of personalized emotion models that adapt to individual vocal nuances and cultural differences could greatly enhance the universality and precision of SER systems. Personalized models can account for specific characteristics in a person's speech, such as accent, pitch range, and emotional expressiveness, thereby improving the system's ability to detect emotions accurately across different populations.

These advancements hold the potential to significantly expand the applications of SER in various fields. In human-computer interaction, enhanced SER systems can improve user experience by enabling more empathetic and responsive virtual assistants. In mental health monitoring, they can provide real-time emotional insights, supporting timely interventions and better mental health management. Additionally, in adaptive communication technologies, SER systems can facilitate more effective and emotionally aware communication, enhancing interactions in both personal and professional settings. Continued research and innovation in these areas promise to propel the field of SER towards more sophisticated and practical applications.

REFERENCES

[1] A. Dix, *Human Computer Interaction*. Pearson Education India, 2008.

[2] M. T. Prior and G. Kasper, *Emotion in Multilingual Interaction*. John Benjamins Publishing Company, 2016.

[3] H. He, *Self-Adaptive Systems for Machine Intelligence*. John Wiley & Sons, 2011.

[4] J. Brownlee, *Generative Adversarial Networks with Python: Deep Learning Generative Models for Image Synthesis and Image Translation*. Machine Learning Mastery, 2019.

[5] P. Nimitsurachat and P. Washington, "Audio-Based Emotion Recognition Using Self-Supervised Learning on an Engineered Feature Space," *AI (Basel)*, vol. 5, no. 1, pp. 195–207, Mar. 2024.

[6] K. Sreenivasa Rao and S. G. Koolagudi, *Emotion Recognition using Speech Features*. Springer Science & Business Media, 2012.

[7] P. Siirtola, S. Tamminen, G. Chandra, A. Ihala Pathirana, and J. Röning, "Predicting Emotion with Biosignals: A Comparison of Classification and Regression Models for Estimating Valence and Arousal Level Using Wearable Sensors," *Sensors* , vol. 23, no. 3, Feb. 2023, doi: 10.3390/s23031598.

[8] L. Mary, *Extraction of Prosody for Automatic Speaker, Language, Emotion and Speech Recognition*. Springer, 2018.

[9] A. Pentari, G. Kafentzis, and M. Tsiknakis, "Speech emotion recognition via graph-based representations," *Sci. Rep.*, vol. 14, no. 1, p. 4484, Feb. 2024.

[10] M. M. Billah, M. L. Sarker, and M. A. H. Akhand, "KBES: A dataset for realistic Bangla speech emotion recognition with intensity level," *Data Brief*, vol. 51, p. 109741, Dec. 2023.

[11] D. D. Olatinwo, A. Abu-Mahfouz, G. Hancke, and H. Myburgh, "IoT-Enabled WBAN and Machine Learning for Speech Emotion Recognition in Patients," *Sensors* , vol. 23, no. 6, Mar. 2023, doi: 10.3390/s23062948.

[12] O. Valentin, A. Lehmann, D. Nguyen, and S. Paquette, "Integrating Emotion Perception in Rehabilitation Programs for Cochlear Implant Users: A Call for a More Comprehensive Approach," *J. Speech Lang. Hear. Res.*, vol. 67,

no. 5, pp. 1635–1642, May 2024.

[13] Y. Zhang *et al.*, “Identifying depression-related topics in smartphone-collected free-response speech recordings using an automatic speech recognition system and a deep learning topic model,” *J. Affect. Disord.*, vol. 355, pp. 40–49, Jun. 2024.

[14] B. Mirheidari, A. Bittar, N. Cummins, J. Downs, H. L. Fisher, and H. Christensen, “Automatic detection of expressed emotion from Five-Minute Speech Samples: Challenges and opportunities,” *PLoS One*, vol. 19, no. 3, p. e0300518, Mar. 2024.

[15] M. M. Duville, L. M. Alonso-Valerdi, and D. I. Ibarra-Zarate, “Improved emotion differentiation under reduced acoustic variability of speech in autism,” *BMC Med.*, vol. 22, no. 1, p. 121, Mar. 2024.

[16] K. Sreenivasa Rao and S. G. Koolagudi, *Robust Emotion Recognition using Spectral and Prosodic Features*. Springer Science & Business Media, 2013.