Project Title –

Deciphering Emotions in Audio Signals: Leveraging Advanced Models and Feature Extraction Techniques

Course – Biometrics (CIS663)

Team members –
Deepika Nandan (927509909)

Contributions -
My main contributions encompassed curating and organizing data, devising, and training sophisticated models, and meticulously analyzing results. I ensured comprehensive preparation of datasets, optimized model architectures for accurate emotion recognition, and conducted thorough analyses to interpret and refine the outcomes, aiming for robust and reliable emotion decoding from audio signals.

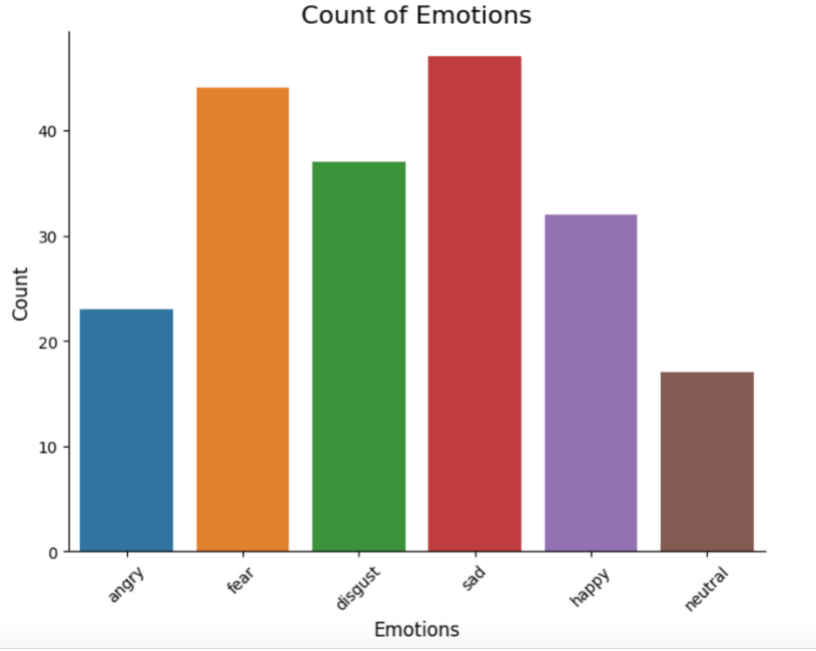
1. Introduction

1.1. Project Overview

The burgeoning interest in emotion recognition from audio signals has propelled research across diverse fields, owing to its manifold applications. This project constitutes a comprehensive exploration of machine learning techniques aimed at deciphering nuanced emotions ingrained within speech signals. Emotion recognition, nestled within the broader field of affective computing, emerges as a potent tool applicable in multifarious domains such as psychology, human-computer interaction, sentiment analysis, and more.

At its core, this project is fueled by the aspiration to leverage the extensive CREMA-D dataset, a reservoir comprising a vast array of audio instances. The dataset encapsulates emotive articulations, capturing the spectrum of human emotions ranging from anger, fear, happiness, and sadness to disgust and neutrality. This multifaceted dataset serves as the foundational bedrock for training and validating the machine learning models devised within this project.

The primary objective revolves around achieving a high degree of accuracy in classifying these varied emotional states accurately. This necessitates the utilization of sophisticated feature extraction methods and the orchestration of advanced machine learning models. The extraction methods encompass pivotal techniques such as Mel-frequency cepstral coefficients (MFCC), zero-crossing rate (ZCR), and spectral features, essential in transforming raw audio signals into machine-understandable representations.



The provided code encapsulates a multifaceted approach, constituting several classes, each performing distinct tasks in the pipeline of emotion recognition. The **DataPreparation** class initiates the process by loading the CREMA-D dataset and preparing it for subsequent processing steps. The **DataVisualization** class facilitates the visualization of emotions through waveplots and

spectrograms, aiding in understanding the temporal and frequency dynamics of different emotional states within audio clips.

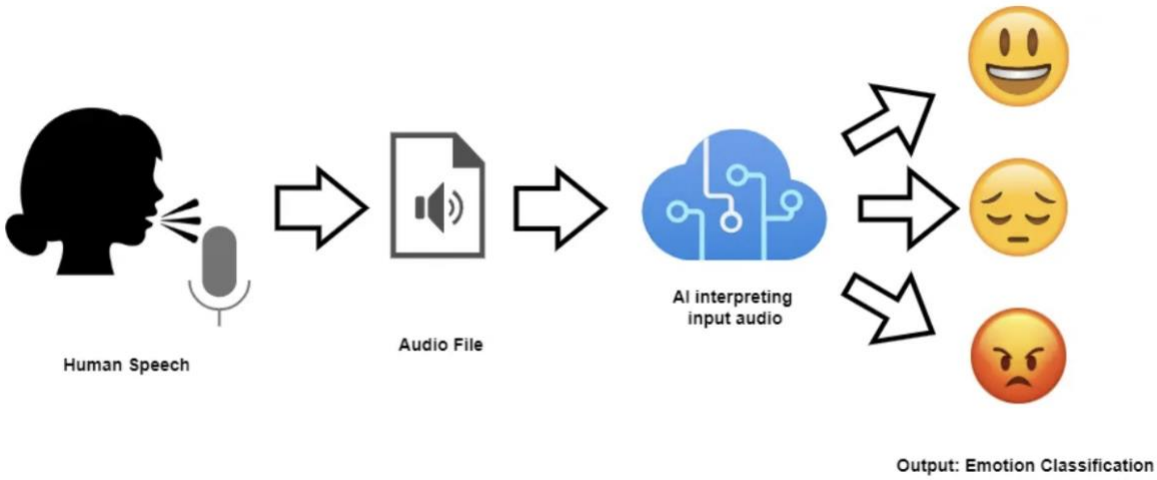
Moreover, the code features an **AudioVisualizer** class, enabling the generation of waveplots and spectrograms for specific emotions. The **SpectrogramCreation** class leverages the CREMA-D dataset to generate spectrogram images, which are crucial in training machine learning models. The **ModelTraining** class orchestrates the training of models, employing convolutional neural networks (CNNs) and recurrent neural networks (RNNs), aiming to achieve accurate emotion classification.

This holistic approach endeavors to harness the potential of machine learning in decoding emotions from audio data, catering to diverse real-world applications and advancing the field of affective computing.

2. Dataset CREMA-D

2.1 Dataset (CREMAD) Overview:

The CREMA-D dataset stands as an extensive repository comprising a multitude of emotional expressions. Boasting an impressive compilation of over 7,000 instances, it portrays actors' performances across diverse emotions, establishing a robust spectrum for both training and evaluating machine learning models. This breadth allows for a nuanced understanding of emotions, essential for ensuring the model's proficiency in discerning subtle variations in emotional states.



The CREMA-D dataset comprises 7,442 original audio clips delivered by 91 actors, spanning 48 male and 43 female individuals with ages ranging from 20 to 74 and representing diverse racial backgrounds: African American, Asian, Caucasian, Hispanic, and Unspecified.

Actors vocalized from a pool of 12 sentences, each articulated with one of six distinct emotions: Anger, Disgust, Fear, Happy, Neutral, and Sad. Emotions were expressed at four levels: Low, Medium, High, and Unspecified.
File names in CREMA-D follow a structured format with four underscore-separated blocks. For instance, the first block denotes the Actor ID, the second signifies the chosen sentence out of 12 variations, the third specifies the emotion conveyed (from 6 categories), and the fourth designates the intensity level (out of 4 options). For example, in the audio files of the dataset, "1001" refers to the actor ID, "DFA" indicates the spoken sentence ("Don't forget a jacket"), "ANG" represents the emotion expressed as anger, and "XX" denotes unspecified intensity.

3. Literature Search

3.1. Speech Emotion Recognition using Machine Learning

Objective: Analyze the landscape of machine learning applications for speech emotion recognition (SER) as presented in "Speech Emotion Recognition using Machine Learning" by K. Vamsi Krishna, N. Sainath, and A. Mary Posonia.

Key Findings:

- Challenges in SER: Authors highlight the complexities of human speech and emotional expression, posing challenges for accurate machine learning models.
- Machine Learning Techniques: Traditional methods like Support Vector Machines (SVMs) and Decision Trees are employed for emotion classification, but deep learning shows promising advancement.
- Deep Learning Models: Convolutional Neural Networks (CNNs) excel at analyzing spectrograms, while Recurrent Neural Networks (RNNs) capture temporal dependencies in speech. Long Short-Term Memory (LSTM) networks are effective for long-term dependencies.
- Feature Representation: Acoustic features (pitch, loudness) and linguistic features (word choice, grammar) are commonly used to represent speech for machine learning models.
- Future Directions: Addressing data scarcity with augmentation and transfer learning, exploring cross-domain recognition, and investigating attention mechanisms are highlighted as potential areas for improvement.
-

Gaps Identified:

- Limited Data: Training data size and diversity are critical for model generalizability, posing a challenge for SER.
- Cultural Variations: Emotion perception varies across individuals and cultures, hindering the development of universal models.
- Limited Explanation: Deep learning models often lack interpretability, making it difficult to understand their decision-making process.

Research Contributions:

- This paper provides a comprehensive overview of the application of machine learning for SER, highlighting current advancements and future directions.

- It emphasizes the potential of deep learning for improving SER accuracy while acknowledging remaining challenges and suggesting research avenues for further development.

3.2. A Comprehensive Review of Speech Emotion Recognition Systems

Objective: Analyze the existing body of knowledge on Speech Emotion Recognition (SER) systems, drawing upon the insights presented in "A Comprehensive Review of Speech Emotion Recognition Systems" by T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairaj.

Key Findings:

- **SER System Components:** The typical SER system consists of four modules: speech signal input, preprocessing, feature extraction/selection, and classification for emotion recognition.
- **Methodological Landscape:** A range of methodologies are used in SER systems, including statistical methods (e.g., Gaussian Mixture Models) and deep learning approaches (e.g., Convolutional Neural Networks).
- **Emotional Models:** Emotion models like Ekman's six basic emotions and dimensional models (e.g., valence-arousal) provide frameworks for categorizing and analyzing emotions in SER systems.
- **Feature Extraction:** Acoustic features (e.g., pitch, formants) and linguistic features (e.g., speech rate, prosody) are commonly extracted from speech signals to represent emotional information.
- **Databases:** Publicly available databases like IEMOCAP and EMO-DB serve as benchmarks for evaluating and comparing SER systems.

Challenges and Gaps:

- **Data Scarcity:** Limited availability of labeled data for training SER systems, particularly for specific emotions and domains.
- **Interdisciplinary Integration:** Integrating knowledge from speech processing, psychology, and linguistics remains a challenge.
- **Cultural Variations:** Cultural influences on emotional expression necessitate developing adaptable and context-aware SER systems.
- **Explain ability and Robustness:** Deep learning models often lack interpretability, making it difficult to understand their decision-making process. Additionally, enhancing robustness against noise and environmental factors is crucial.

3.3. Speech Emotion Recognition using Machine Learning

Objective: Analyze the recent advances in speech emotion recognition (SER) using machine learning, as presented in "Speech Emotion Recognition using Machine Learning" by R. Anusha, P. Subhashini, Darelli Jyothi, Potturi Harshitha, Janumpally Sushma, and Namsamgari Mukesh.

Key Findings:

- Machine Learning Techniques: Deep learning methods have shown promising results in SER, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs).
- Feature Representation: Acoustic features (e.g., pitch, loudness) and linguistic features (e.g., word choice, grammar) are commonly used for feature representation in SER.
- Emotion Models: Ekman's six basic emotions and dimensional models (e.g., valence-arousal) are commonly used for emotion modeling in SER.
- Datasets: Publicly available databases like IEMOCAP and EMO-DB serve as benchmarks for evaluating and comparing SER systems.

Challenges and Gaps:

- Data Scarcity: Limited availability of labeled data for training SER systems, particularly for specific emotions and domains.
- Cultural Variations: Cultural influences on emotional expression necessitate developing adaptable and context-aware SER systems.
- Explainability: Deep learning models often lack interpretability, making it difficult to understand their decision-making process.

Future Directions:

- Multimodal Emotion Recognition: Combining speech signals with other modalities (e.g., facial expressions, physiological signals) can improve emotion recognition accuracy.
- Domain Adaptation: Utilizing data from other domains and leveraging domain adaptation techniques can address data scarcity and enhancegeneralizability.
- Explainable AI (XAI): Research on XAI techniques can improve transparency and trust in AI-powered SER systems.

Research Contributions:

- This paper provides a comprehensive overview of the recent advances in SER using machine learning, highlighting the potential of deep learning for improving accuracy.
- It also discusses the challenges and gaps in the field, and suggests future directions for research.

4. Feature Extraction

4.1. Feature Extraction Methods:

Feature extraction plays a pivotal role in preparing raw audio signals for machine learning models. The code utilizes various techniques for this purpose, notably MFCC and Mel-spectrogram:

	emotionG_label	source	gender	emotion	path	melspectrogram_path
0	angry_male	CREMA	male	angry	/content/AudioWAV/1077_IOM_ANG_XX.wav	/content/Spectrogram/1077_IOM_ANG_XX.jpg
1	fear_male	CREMA	male	fear	/content/AudioWAV/1050_IEO_FEA_HI.wav	/content/Spectrogram/1050_IEO_FEA_HI.jpg
2	disgust_male	CREMA	male	disgust	/content/AudioWAV/1035_TAI_DIS_XX.wav	/content/Spectrogram/1035_TAI_DIS_XX.jpg
3	sad_male	CREMA	male	sad	/content/AudioWAV/1066_IEO_SAD_LO.wav	/content/Spectrogram/1066_IEO_SAD_LO.jpg
4	happy_male	CREMA	male	happy	/content/AudioWAV/1066_MTI_HAP_XX.wav	/content/Spectrogram/1066_MTI_HAP_XX.jpg
...	...	...	...	...	...	...
195	neutral_male	CREMA	male	neutral	/content/AudioWAV/1026_TAI_NEU_XX.wav	/content/Spectrogram/1026_TAI_NEU_XX.jpg
196	sad_male	CREMA	male	sad	/content/AudioWAV/1087_TAI_SAD_XX.wav	/content/Spectrogram/1087_TAI_SAD_XX.jpg
197	sad_female	CREMA	female	sad	/content/AudioWAV/1078_TAI_SAD_XX.wav	/content/Spectrogram/1078_TAI_SAD_XX.jpg
198	fear_male	CREMA	male	fear	/content/AudioWAV/1064_DFA_FEA_XX.wav	/content/Spectrogram/1064_DFA_FEA_XX.jpg
199	disgust_female	CREMA	female	disgust	/content/AudioWAV/1010_ITH_DIS_XX.wav	/content/Spectrogram/1010_ITH_DIS_XX.jpg

4.2. Mel-frequency Cepstral Coefficients (MFCC):

MFCC is a prevalent technique in speech and audio signal processing. It condenses the information within audio signals into a concise representation, capturing crucial speech-related characteristics. This method involves multiple steps:

1. **Frame the Audio:** The audio signal is segmented into short frames, typically around 20-40 milliseconds each.
2. **Apply Fourier Transform:** Fourier transform is employed to convert each frame from the time domain to the frequency domain.
3. **Mel Filter Bank:** A filter bank, typically consisting of 20-40 filters spaced on the mel-scale, is applied to the power spectrum.
4. **Log Compression:** Logarithm is applied to the filter-bank energies to mimic the human perception of sound intensity.
5. **Discrete Cosine Transform (DCT):** Finally, DCT is used to transform the log filter-bank energies into a compact set of coefficients known as MFCCs.

4.3. Mel-spectrograms:

Mel-spectrograms represent the frequency content of audio signals over time, emphasizing frequencies that humans perceive logarithmically. In the code, the process for generating Mel-spectrogram involves:

1. **Loading Audio Data:** Using librosa, audio data is loaded from file paths.
2. **Computing Mel-scaled Spectrogram: librosa.feature.melspectrogram** computes the mel-scaled spectrogram from the audio waveform.
3. **Converting to Decibel Scale: librosa.power_to_db** converts the power spectrogram to decibel scale to enhance visualization and analysis.

4. **Displaying Spectrograms:** Matplotlib is utilized to display and save the generated Mel-spectrogram.

In the code, the **SpectrogramCreation** class harnesses these processes to generate Mel-spectrogram, providing a visual representation of audio signals in the frequency domain. The extracted features from these spectrograms contribute significantly to the subsequent training of machine learning models for emotion classification.

5. Mel-spectrogram v/s MFCC

5.1. Mel-spectrogram:

Mel-spectrogram serve as visual representations of audio signals in the frequency domain. By mapping the frequency content onto the mel scale, Mel-spectrogram offer a comprehensive depiction of frequency distributions present within an audio clip. The mel scale reflects human perception of sound frequencies, providing a logarithmic scale that closely aligns with human auditory sensitivity. This characteristic renders Mel-spectrogram proficient in capturing variations across different frequency ranges, making them adept at illustrating nuances in frequency content.

5.2. Mel-frequency Cepstral Coefficients (MFCC):

MFCC, on the other hand, goes beyond frequency representation by encapsulating crucial spectral features. It doesn't merely illustrate the frequency distribution but extracts distinctive features associated with the vocal tract's shape, capturing essential speech-related information. MFCC incorporates a multi-step process, condensing the audio signal into a concise representation by employing various signal processing techniques such as framing, Fourier transform, Mel filter bank, logarithm compression, and discrete cosine transform. This process extracts coefficients representing critical speech characteristics, including formants and spectral envelope, which are pivotal in speech analysis.

5.3. Distinct Characteristics and Applications:

Each method possesses unique characteristics instrumental in audio analysis, particularly in the context of emotion recognition. Mel-spectrogram excel in providing a comprehensive overview of frequency content, facilitating a broader understanding of the audio signal's frequency distribution. Their proficiency lies in capturing the overall frequency characteristics, enabling visualization of subtle variations across different frequency ranges.

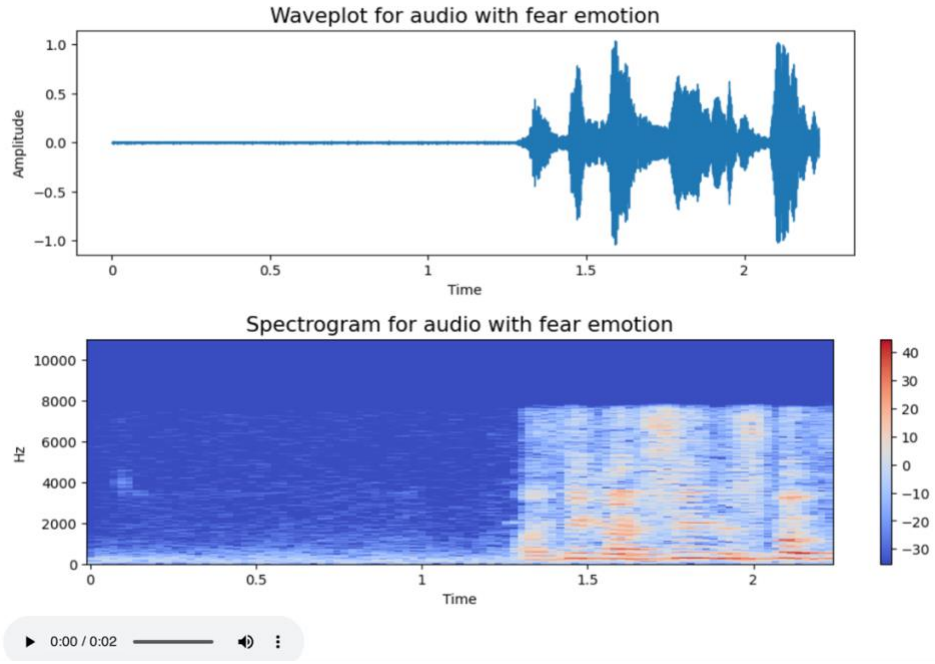
In contrast, MFCC excels in encapsulating specific speech-related information crucial for discerning emotions within speech signals. By capturing spectral features like formants and the vocal tract's shape, MFCC provides a compact yet comprehensive representation that focuses on critical speech components relevant to emotion recognition tasks.

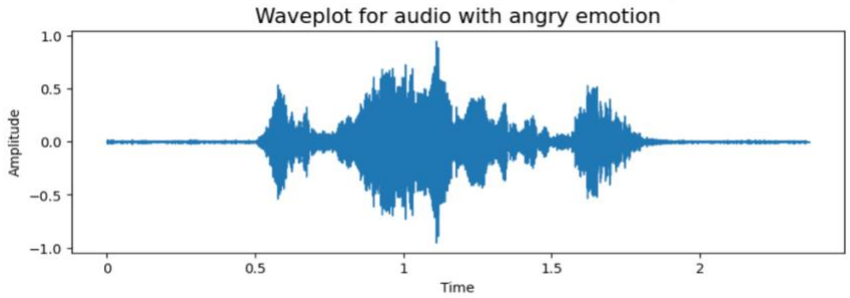
5.4. Advantages of MFCC:

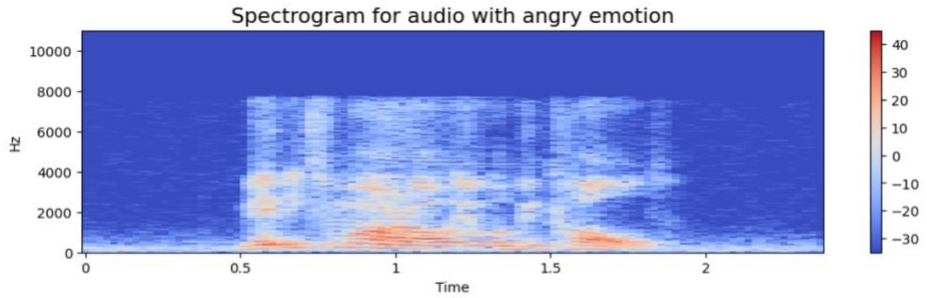
MFCC's superiority stems from its ability to distill complex audio signals into a concise set of coefficients representing speech-related features. These coefficients effectively capture the nuances of speech, such as intonation, emphasis, and articulation, all of which play pivotal roles

in conveying emotions. Furthermore, the compact representation of MFCC coefficients contributes to reducing computational complexity, making them efficient for modeling tasks.
In emotion recognition, MFCC's emphasis on capturing speech-specific characteristics makes it a preferred choice. Its focus on critical features related to emotions, combined with its efficiency and effectiveness, renders MFCC a powerful tool in discerning and interpreting emotions embedded in speech signals.

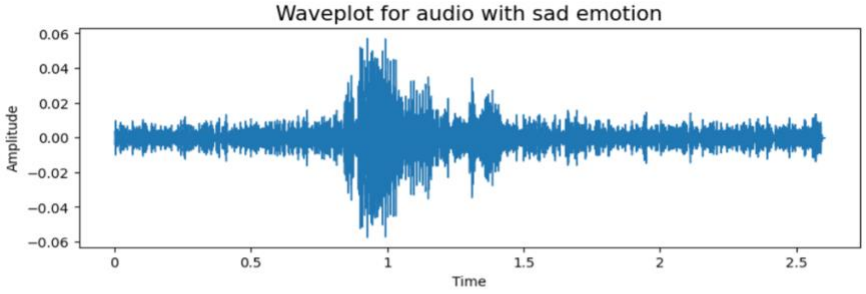
6. Data Visualization of audio wave-plots and spectrograms

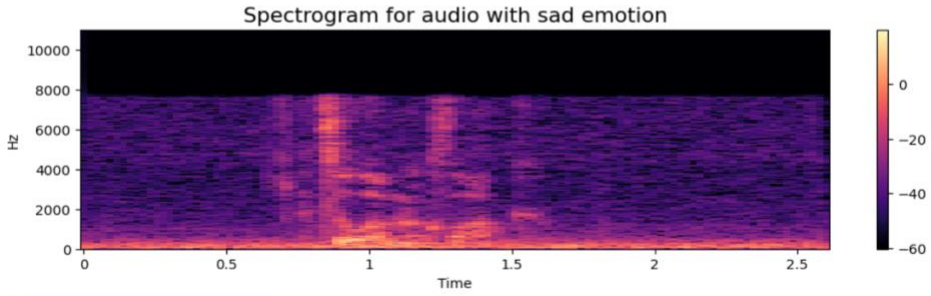




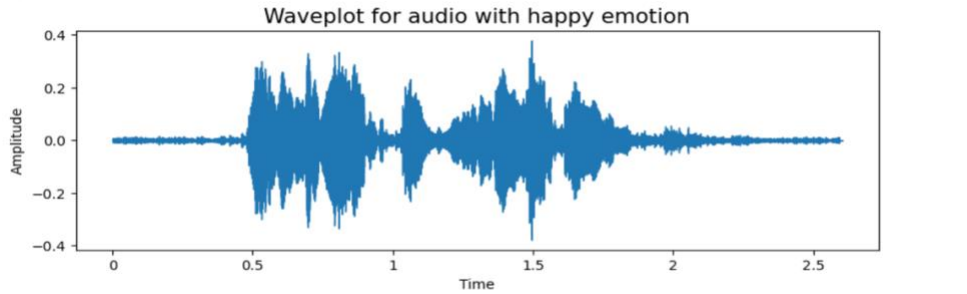


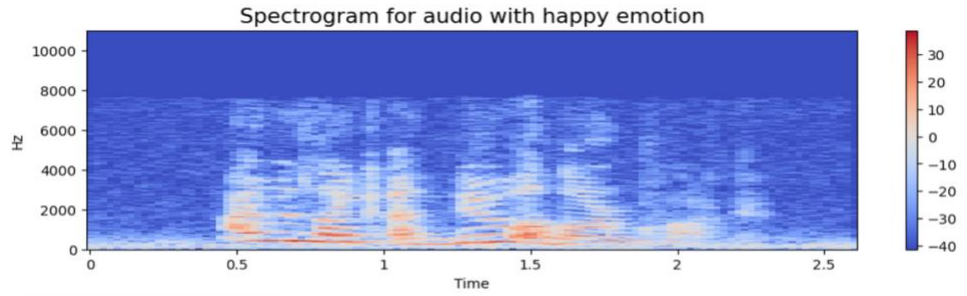
0:00 / 0:02



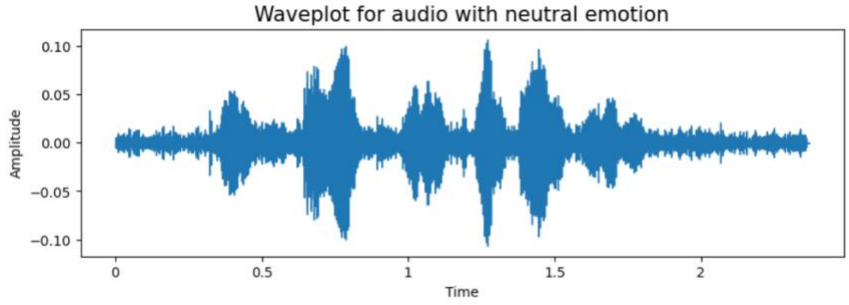


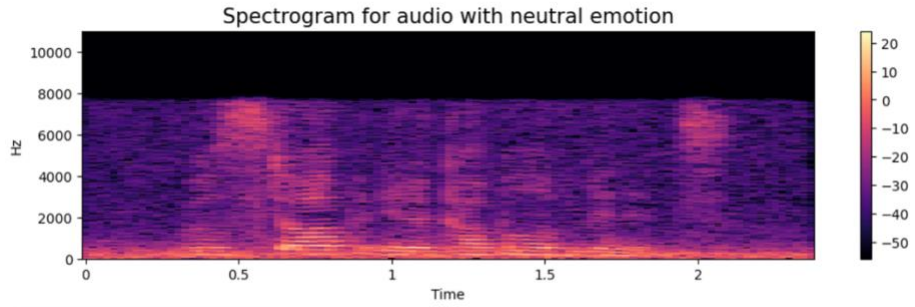
0:00 / 0:02



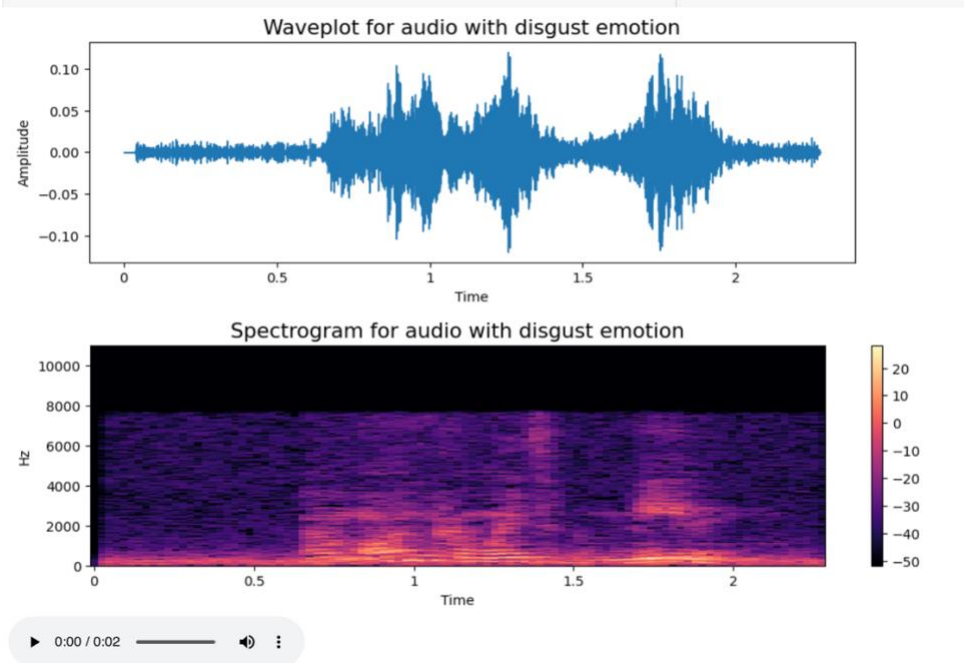


0:00 / 0:02





0:00 / 0:02



7. Models Used and Execution

7.1. Convolutional Neural Networks (CNNs):

CNNs are adept at extracting spatial patterns and hierarchical representations from input data. In the context of audio analysis, especially with spectrogram images generated from audio signals, CNNs excel in capturing distinctive spatial features present within these spectrograms. By employing convolutional layers, pooling operations, and feature mapping, CNNs effectively discern patterns and spectral characteristics crucial for emotion classification.

7.2. Recurrent Neural Networks (RNNs) - LSTM Networks:

RNNs, specifically Long Short-Term Memory (LSTM) networks, are tailored to capture temporal dependencies within sequential data. In the realm of audio analysis, LSTM networks are particularly adept at learning long-range dependencies and patterns existing across different time steps in audio sequences. By leveraging memory cells and gating mechanisms, LSTMs retain essential information over extended periods, ensuring a comprehensive understanding of temporal dynamics embedded in audio signals.

7.3. Hybrid Architectures:

The fusion of CNNs and RNNs in hybrid architectures amalgamates their strengths, exploiting both spatial and temporal information present in audio data. This hybrid approach aims to harness the spatial patterns extracted by CNNs from spectrogram images and the temporal dependencies captured by RNNs. By leveraging the synergy between these architectures, the model achieves a holistic comprehension of both spatial and temporal features within audio signals, thus enhancing accuracy and efficacy in discerning emotions embedded in the data.

This strategic amalgamation of neural network architectures optimizes their complementary strengths, enhancing the model's ability to decipher and interpret complex emotional cues inherent in audio signals. The hybrid approach ensures a more comprehensive and robust understanding of audio features, consequently improving the accuracy and reliability of emotion recognition tasks.

This project stands as a pioneering endeavor in the domain of emotion recognition from audio signals, carrying substantial significance across various fields including mental health diagnostics, human-computer interaction, and sentiment analysis. By amalgamating state-of-the-art feature extraction techniques and leveraging hybrid neural network architectures, it aims to propel the accuracy and efficacy of recognizing emotions embedded within audio signals to unprecedented levels.

7.4. Execution

7.4.1. Mel-spectrogram Processing with CNN:

In the implementation for Mel-spectrogram processing using Convolutional Neural Networks (CNN), the workflow involves several key steps:

- **Data Preparation:** Audio files are processed to extract Mel-spectrogram representations through the **librosa** library, ensuring standardized input size.
- **Model Building:** A CNN architecture is constructed using TensorFlow/Kera's, comprising convolutional layers to extract spatial features, pooling layers for down sampling, and dense layers for classification.
- **Training and Validation:** The compiled model is trained on the Mel-spectrogram data, leveraging appropriate loss functions and optimizers. Validation occurs on separate datasets to assess model performance.
- **Evaluation:** The trained CNN model is evaluated using various performance metrics to gauge its accuracy, precision, recall, and F1-score on unseen Mel-spectrogram data.

7.4.2. MFCC Data Processing with LSTM:

For Mel-frequency cepstral coefficients (MFCC) processing using Long Short-Term Memory (LSTM) networks, the implementation follows a distinct workflow:

- **Feature Extraction:** Audio files are processed to extract MFCCs using the **librosa** library. These features are organized into sequential data suitable for LSTM input.
- **LSTM Model Setup:** An LSTM architecture is constructed using TensorFlow/Keras, specifically designed to handle sequential MFCC data. Stacked LSTM layers capture temporal dependencies, leading to better emotion recognition.
- **Training and Validation:** The compiled LSTM model is trained on the prepared MFCC sequences. Like the CNN, validation occurs to assess the model's generalization capabilities.
- **Comparison and Analysis:** Both the CNN (Mel-spectrogram) and LSTM (MFCC) models' performances are evaluated and compared, analyzing their accuracies and overall efficacy in recognizing emotions from audio signals.

8. Significance and Novelty

8.1. Significance in Various Domains:

- **Mental Health Diagnostics:**
  Accurate emotion recognition from audio signals holds immense promise in mental health diagnostics. It opens avenues for monitoring and assessing emotional states, aiding mental health professionals in understanding patients' emotional well-being and identifying potential indicators of psychological conditions.
- **Human-Computer Interaction:**
  Advancements in emotion recognition technology enrich human-computer interaction, enabling systems to comprehend users' emotional cues. This facilitates more personalized and responsive interactions, leading to enhanced user experiences across various applications, from virtual assistants to gaming and entertainment platforms.
- **Sentiment Analysis:**
  In the realm of sentiment analysis, precise emotion recognition from audio data provides a deeper understanding of public sentiment in large-scale data, such as social media content, customer reviews, and market sentiments. This can be invaluable for businesses in gauging customer satisfaction, refining marketing strategies, and understanding consumer sentiments.

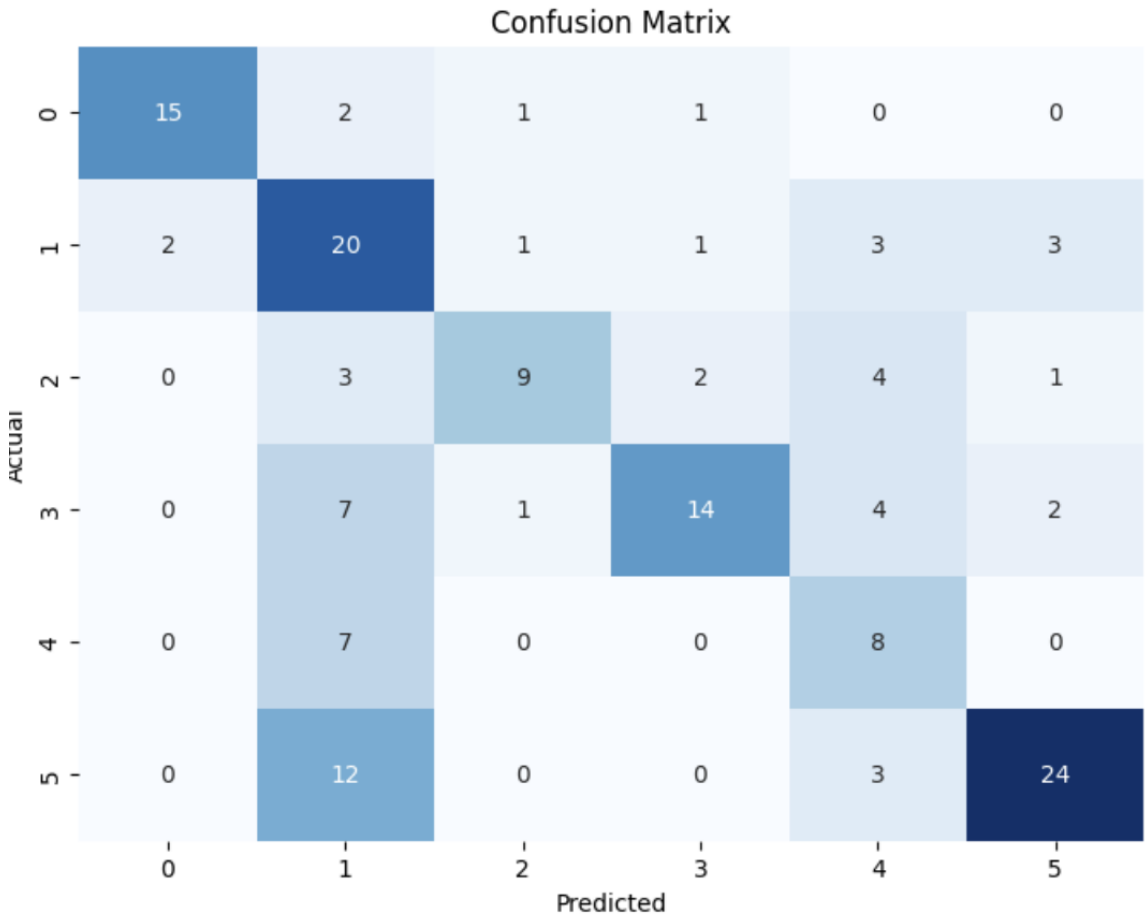
8.2. Novelty and Holistic Approach:

The novelty of this project lies in its holistic approach, integrating advanced signal processing techniques with cutting-edge machine learning models. By synergizing diverse feature extraction methods like MFCC and Mel-spectrogram with hybrid neural network architectures encompassing CNNs and RNNs, it endeavors to decipher nuanced emotional information embedded within audio data more accurately than conventional approaches.

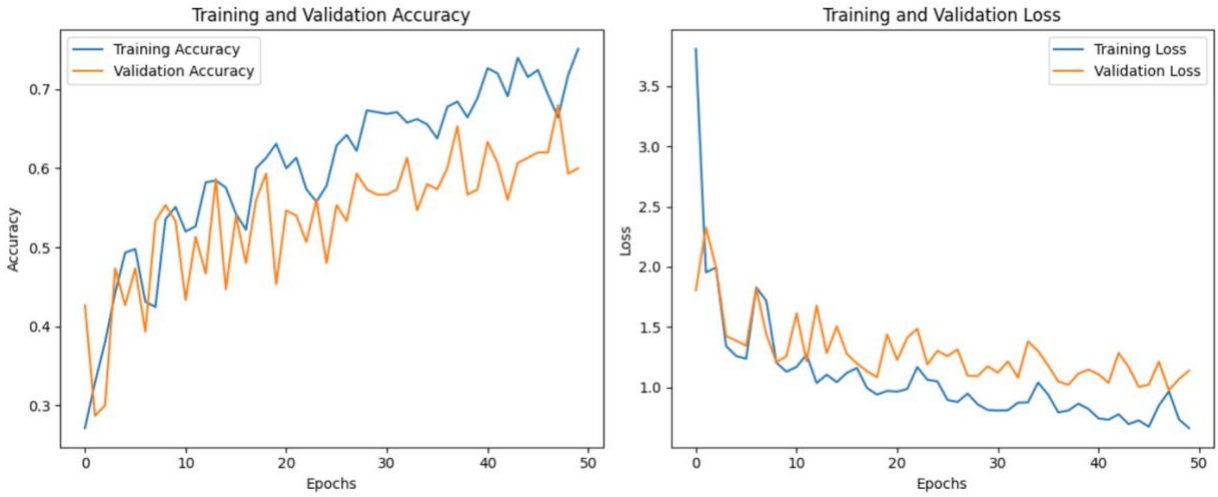
This holistic amalgamation of signal processing techniques and sophisticated machine learning models represents a unique advancement. It not only promises superior accuracy in emotion recognition but also signifies a paradigm shift towards more comprehensive and nuanced understanding of emotional cues within audio signals, paving the way for more robust and refined emotion recognition systems with multifaceted applications across diverse domains.

9. Results and Conclusion

The obtained accuracies of approximately 65% for the CNN+LSTM model utilizing MFCC data and around 26% for the CNN model using Melspectrogram data reveal notable disparities in their performance. The higher accuracy achieved by the CNN+LSTM model with MFCC representation signifies its superiority in discerning emotions from audio signals compared to the CNN model using Melspectrogram data.

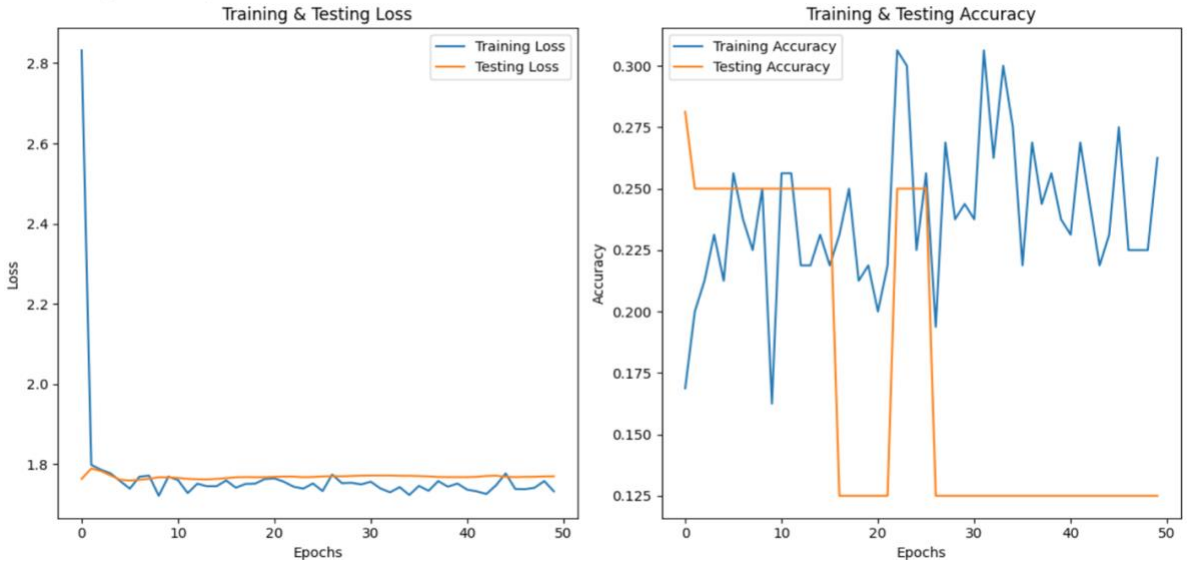


The utilization of MFCCs as features for the CNN+LSTM model allows it to capture nuanced temporal dependencies and critical speech-related information inherent in audio signals. The LSTM's capability to preserve and learn from sequential patterns in the MFCC data complements the CNN's ability to extract spatial features. Consequently, this combined model can effectively recognize and interpret emotions encoded within the audio signals, resulting in significantly improved accuracy.



Contrarily, the lower accuracy achieved with the CNN model using Melspectrogram data indicates the limitation of spatial-based representations alone in capturing the intricacies of emotional content present in audio signals. Melspectrograms, although providing frequency distribution representations, might lack the temporal information crucial for accurate emotion recognition, especially when compared to the rich sequential details captured by MFCCs. This notable contrast underscores the efficacy and superiority of MFCC-based representations in speech emotion recognition tasks.

```
Final Accuracy after 50 epochs: 26.25%
```



10. Challenges Faced and Future Work

During the project, challenges were encountered in tackling class imbalances within the dataset, optimizing hyperparameters for models, and implementing effective data augmentation techniques for enhanced generalization. Future work involves delving into more advanced feature extraction methods to capture subtle emotional nuances, addressing persistent class imbalances through advanced sampling strategies or synthetic data generation, and striving for real-time deployment to extend the practical applications of the developed models. To enhance the project further, conducting a more extensive exploration of ensemble techniques, incorporating continual learning for model refinement, and ensuring scalability and efficiency in real-time deployment would be beneficial.

11. References

- K. V. Krishna, N. Sainath and A. M. Posonia, "Speech Emotion Recognition using Machine Learning," 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2022, pp. 1014-1018, doi: 10.1109/ICCMC53470.2022.9753976.
- T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi and E. Ambikairajah, "A Comprehensive Review of Speech Emotion Recognition Systems," in IEEE Access, vol. 9, pp. 47795-47814, 2021, doi: 10.1109/ACCESS.2021.3068045.
- R. Anusha, P. Subhashini, D. Jyothi, P. Harshitha, J. Sushma and N. Mukesh, "Speech Emotion Recognition using Machine Learning," 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2021, pp. 1608-1612, doi: 10.1109/ICOEI51242.2021.9453028.
- https://hami-asmai.medium.com/project-speech-emotion-recognition-using-cnn-with-crema-d-dataset-using-google-colab-8cb4fdbd5044
- https://ppk1999.medium.com/speech-emotion-recognition-ser-on-crema-d-crowd-sourced-emotional-multimodal-actors-dataset-70d93206c230
- https://importchris.medium.com/how-to-create-understand-mel-spectrograms-ff7634991056
- https://medium.com/@polanitzer/building-a-convolutional-neural-network-in-python-predict-digits-from-gray-scale-images-of-550d79b358b
- https://www.kaggle.com/code/shivamburnwal/speech-emotion-recognition/notebook