Mind Your Mind: Real-Time Emotional Insights from Voice

Harshith Manoharan[1]; Keerthana R E[2]; N. Selvaganesh[3]; Logeswari P[4]

[1,2,3,4] Department of Information Technology, Sri Venkateswara College of Engineering, Chennai, India

Publication Date: 2025/06/03

Abstract: The growing need for accessible mental health support highlights the importance of innovative digital solutions. Many individuals struggle to manage emotions, leading to heightened stress, anxiety, and a decline in well-being. Traditional methods like journaling or therapy, while beneficial, can often feel time-consuming, intimidating, or inaccessible. Current mental health apps frequently fall short, lacking emotional analysis. There is a rising demand for a non-intrusive, user-friendly solution can monitor emotions and provide meaningful insights outside conventional therapy. Mind Your Mind addresses the gap with a voice-based journaling system powered by emotional analysis. Using advanced speech processing, the platform evaluates tone, pitch, and sentiment to assess emotional states as users speak naturally by employing an AI-driven emotion recognition model, integrating Mel- Frequency Cepstral Coefficients (MFCC), Mel-Spectrograms, and Convolutional Neural Networks (CNNs) for accurate pattern recognition. The model achieves an accuracy of 92.3%, enabling reliable emotional detection. Users interact through an intuitive web interface, recording their thoughts and receiving immediate, actionable mood insights in textual format.

Keywords: *Emotion Recognition, Mental Health, Speech Analysis, Voice Journaling.*

How to cite: Harshith Manoharan; Keerthana R E; N. Selvaganesh; Logeswari P; (2025) Mind Your Mind: Real-Time Emotional Insights from Voice. International Journal of Innovative Science and Research Technology, 10(5), 2982-2996. https://doi.org/10.38124/ijisrt/25may894

I. INTRODUCTION

The Voice-Based Mood Recognition System operates within the domain of AI-driven mental wellness solutions, focusing on the subdomain of voice-based emotional analysis. Mental well-being refers to a state of mental health where individuals can cope with everyday stress, work productively, and contribute to their communities while maintaining a sense of inner balance and purpose. The primary objective is to enable users to gain real-time insights into their emotional state using vocal inputs, leveraging deep learning models for mood prediction [1]. As mental wellness platforms evolve, integrating non-intrusive, AI-powered mood detection can help bridge the gap between traditional self-reporting methods and intelligent, real-time emotional support.

In the context of voice-based mood analysis, advanced speech processing techniques such as MFCC extraction, spectrogram analysis, and deep learning models play a crucial role [24]. These technologies enable accurate emotion classification, ensuring that mood predictions are reliable even with varying vocal expressions.

Additionally, AI-powered mental wellness assistants can enhance user experience by offering personalized recommendations, affirmations, and guided interventions based on detected moods [2]. By incorporating real-time processing through Flask, the system ensures seamless interaction, making emotional self-awareness more accessible, engaging, and intuitive.

Voice-based emotion recognition has emerged as a transformative approach to mental wellness, enabling users to express their emotions naturally without relying on manual mood tracking [15]. Traditional methods, such as journaling and questionnaire-based assessments, often require active user participation and may not accurately capture subtle emotional shifts. By contrast, AI-driven vocal analysis offers a passive and real-time alternative, allowing for automated emotion detection and personalized well-being insights. To tackle the challenges associated with accurate emotion classification, the proposed system utilizes a deep learning approach, combining:

- ➢ *Cnn for Pattern Recognition in Audio Signals [16].*

Supporting technologies such as Flask for backend processing and a web-based frontend for user interaction ensure low-latency, real-time analysis, enhancing the overall experience [12].

AI integration is significantly transforming mental health and emotional well-being by providing intelligent, data-driven insights into user emotions [6]. Machine Learning (ML) and Natural Language Processing (NLP) techniques enable real-time mood analysis, sentiment detection, and personalized intervention strategies [14]. These advancements ensure that users receive contextual emotional support without requiring extensive self-reporting.

➢ *The Proposed System Integrates AI to Enhance User Experience in Multiple ways:*

- Emotion-based content recommendations: Personalized affirmations, mood-boosting activities, and guided meditation suggestions [8].
- Voice-driven journaling and sentiment tracking: AI-assisted journaling with contextual mood insights based on vocal tone [2].

The proposed system not only enhances emotional awareness but also provides proactive mental wellness support, fostering a more empathetic, intuitive, and responsive user experience.

II. RELATED WORK

Barhoumi and BenAyed (2024) [1] proposed a SER system utilizing deep learning methodologies and data augmentation techniques. The study employed various feature extraction methods, including MFCC, ZCR, Mel spectrograms, Root Mean Square Value (RMS), and chroma features. Three deep learning models were explored: Multi-Layer Perceptron (MLP), CNN, and a hybrid CNN+Bidirectional Long Short-Term Memory (BiLSTM) architecture. The hybrid model demonstrated superior performance by capturing both spatial and temporal dependencies in speech data.

Prosodic features such as pitch and intonation have been identified as critical for emotion classification. Spectral features like MFCCs, which reflect energy distribution across frequencies, were widely used in SER systems. It has been shown that MFCCs and their derivatives improved classification accuracy. Spectrogram-based methods have also been employed for SER classification, achieving promising results.

Classical machine learning algorithms such as SVM were applied to SER tasks, with some studies achieving notable accuracy using SVM with MFCC features. However, deep learning approaches have shown significant advancements. LSTM networks have been used with theRyerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset to classify emotions, achieving a certain level of accuracy. Multiple classifiers, including MLP, CNN, and Random Forest, have been compared using EmoDB data, with MLP achieving a high accuracy rate.

A DnCNN-CNN architecture for denoising and classification was introduced, achieving high accuracy on Korean speech datasets. A hierarchical ConvLSTM-based SER system has also been tested on IEMOCAP and RAVDESS datasets, achieving a good recognition rate. Hybrid CNN+BiLSTM model achieved high classification accuracy when evaluated on TESS, EmoDB, and RAVDESS datasets. Data augmentation techniques such as noise addition and spectrogram shifting were employed to enhance model generalization and robustness. Noise addition prevented overfitting by introducing random noise into training data, while spectrogram shifting improved robustness by varying the timing of sound events.

The findings demonstrated that deep learning models outperformed traditional machine learning techniques in SER tasks. Future research directions include incorporating additional acoustic features, expanding datasets, and developing cross-lingual SER systems.

Pan et al. (2024) [10] developed a deep learning-based multimodal emotion recognition (MER) framework called Deep-Emotion, which integrated features from facial expressions, speech, and Electroencephalogram (EEG) signals to enhance emotion recognition performance. By combining these three modalities, the system aimed to capture a more comprehensive understanding of emotional states, improving the system's ability to recognize emotions in varied real-world conditions. The system incorporated three specialized neural network models for feature extraction from each modality, followed by decision-level fusion for improved accuracy and robustness, ensuring the model could handle complex and noisy input data effectively.

For facial expression recognition, an improved GhostNet architecture was introduced, which alleviated overfitting during training and enhanced classification accuracy compared to the original GhostNet model. A lightweight fully convolutional neural network (LFCNN) was implemented for speech emotion recognition, ensuring efficient feature extraction with minimal computational resources. In EEG emotion recognition, a tree-like LSTM (tLSTM) model was designed to fuse multi-stage features, capturing both shallow and deep emotional characteristics.

Decision-level fusion was performed using an optimal weight distribution algorithm, which dynamically assigned weights to each modality based on reliability, improving overall recognition accuracy. The Deep-Emotion framework was evaluated on CK+, EMO-DB, and MAHNOB-HCI datasets, achieving superior performance compared to existing unimodal and multimodal approaches.

Experimental results demonstrated that the multimodal approach outperformed traditional unimodal methods by leveraging complementary emotional cues across different modalities. The fusion of external signals (facial expressions and speech) with internal signals (EEG) ensured a more robust and accurate emotion classification, particularly in scenarios where individual modalities exhibited inconsistencies or lower reliability. The integration of EEG helped mitigate cases where facial expressions and speech did not accurately reflect a person's true emotions, as EEG signals provided a direct measure of internal emotional states. Furthermore, the optimal weight distribution strategy dynamically adjusted the contribution of each modality based on its reliability, allowing the system to maintain high performance even when one signal was weak or noisy. Future research suggested optimizing weight allocation methods for more adaptive fusion strategies, incorporating larger and more diverse datasets, and enhancing real-time applicability for deployment in mobile and embedded systems to make emotion recognition more accessible and practical in real-world scenarios.

Kakuba et al. (2022) [15] developed a Deep Learning-based SER model utilizing multi-level fusion to enhance the concurrent learning of spatial, temporal, and semantic features. The proposed Concurrent Spatial-Temporal and Grammatical Analysis (CoSTGA) model integrated multiple deep learning techniques, including Dilated Causal Convolutions (DCC), BiLSTM, and transformer-based multi-head attention mechanisms, to capture intricate speech emotion patterns effectively.

The model incorporated a two-stage feature extraction approach. In the Local Feature Learning Block (LFLB), spatial and temporal features were extracted separately using DCC and BiLSTM networks, while semantic information was captured using multi-head attention. The extracted features were then fused at the Global Feature Learning Block (GFLB), where multi-level fusion was applied to enhance intra- and cross-modality interactions. The multi-level fusion approach outperformed traditional single-level fusion methods by progressively combining feature representations at different hierarchical levels.

The CoSTGA model was evaluated using the IEMOCAP dataset, where acoustic features such as MFCCs and lexical features obtained from pre-trained BERT embeddings were utilized. Experimental results demonstrated that the multi-level fusion strategy significantly improved classification performance.

The model achieved weighted and unweighted accuracy scores of 75.50% and 75.82%, respectively, surpassing existing SER models that relied on single-level fusion or sequential feature extraction methods. The integration of grammatical and semantic features further enhanced recognition robustness, particularly in distinguishing similar emotional states such as anger and excitement.

The findings highlighted the superiority of multi-level fusion in SER by leveraging concurrent feature learning. Unlike conventional models that process spatial and temporal features sequentially, leading to potential loss of fine-grained emotional details, the CoSTGA model concurrently learned these representations. Future advancements in SER research suggested refining weight allocation strategies in feature fusion to allow dynamic adjustments based on modality reliability. Expanding dataset diversity by incorporating data from different linguistic, cultural, and demographic backgrounds was recommended to enhance model robustness and mitigate biases. Additionally, extending the multimodal framework to integrate facial expressions, physiological signals (e.g., EEG, heart rate variability), and body language cues could provide a more holistic approach to emotion recognition. Such an extension would enable the development of comprehensive affective computing systems capable of understanding human emotions more accurately across various contexts, further bridging the gap between artificial intelligence and human emotional intelligence.

Chamishka et al. (2022) [11] developed a voice-based real-time emotion detection system using a RNN empowered with feature modeling techniques. The proposed approach integrated a novel Bag-of-Audio-Words (BoAW) feature extraction method and an RNN-based classification model to enhance the accuracy of emotion recognition in conversational audio data.

The BoAW technique was implemented to generate compact yet information-rich feature embeddings, enabling better emotion classification. Unlike traditional handcrafted acoustic features, the method represented audio data as a collection of discrete audio-word patterns, capturing intricate variations in speech signals. Additionally, an attention-enhanced recurrent model was introduced, incorporating separate Gated Recurrent Units (GRUs) for modeling global conversational context, speaker states, and emotion states. Multi-stage processing approach improved the system’s ability to preserve long-term dependencies in conversations while dynamically adapting to emotional variations.

The emotion detection system was evaluated on the IEMOCAP dataset, achieving a weighted accuracy of 60.87% and an unweighted accuracy of 60.97%, surpassing existing audio-based emotion recognition models. The experimental setup demonstrated the effectiveness of the BoAW representation in reducing the performance gap between audio and text-based emotion recognition techniques. By integrating conversational context modeling, the system enhanced its ability to recognize complex emotional patterns in speech.

The findings emphasized the advantages of BoAW-based feature extraction in emotion detection, particularly in preserving essential speech characteristics while minimizing computational overhead. The combination of sequential modeling with attention mechanisms contributed to more precise and interpretable emotion predictions. Unlike conventional models that process utterances independently, the proposed approach leveraged contextual information, allowing for a more natural representation of emotional dynamics in dialogue. Future research suggested optimizing feature encoding methods, expanding datasets to diverse linguistic and cultural settings, and integrating multimodal cues such as facial expressions and physiological signals for comprehensive emotion analysis.

Mirsamadi et al. (2017) [24] developed an automatic SER system using RNNs with a local attention mechanism. The proposed model addressed the challenge of extracting emotionally relevant speech features by employing deep learning techniques for both short-term and long-term feature modeling. Unlike conventional SER approaches that relied on handcrafted statistical features, the method utilized BiLSTM networks to learn both frame-level Low-level Descriptors (LLDs) and their temporal variations.

The feature extraction process was divided into two stages: first, frame-level LLDs such as pitch, voicing probability, MFCCs, and energy-based features were computed. Next, temporal aggregation was applied using various pooling strategies, including mean-pooling, final-frame selection, and an attention-based weighted pooling approach. The proposed local attention mechanism assigned dynamic weights to different time frames, enabling the model to focus on emotionally salient regions of speech while ignoring neutral or silent frames.

The speech emotion recognition (SER) system was rigorously evaluated using the IEMOCAP dataset, a widely accepted benchmark for emotion detection in speech. Results revealed that the proposed system significantly outperformed traditional SVM-based approaches, which heavily depend on handcrafted acoustic features and exhibit limitations in capturing the nuanced variations in emotional speech. By employing an attention-based weighted pooling strategy, the model achieved a weighted accuracy of 63.5% and an unweighted accuracy of 58.8%, exceeding the performance of conventional pooling methods such as mean-pooling and final-frame selection. The attention mechanism enabled the model to dynamically focus on emotionally salient segments within the utterance, rather than treating all frames with equal importance.

The integration of deep RNNs further enhanced the model's ability to generalize across varying emotional expressions by learning temporal dependencies and subtle modulations in pitch, tone, and rhythm attributes that are often lost in traditional feature engineering. This allowed the model to capture dynamic emotional shifts over time, leading to more accurate and context-aware predictions. The findings underscore the value of incorporating local attention mechanisms, which allow the system to filter out irrelevant or neutral speech regions and amplify emotionally rich cues, thereby improving classification accuracy and interpretability. The selective focus on critical emotional features also helps in building more robust models that perform well across diverse speech samples and emotional contexts.

Sun et al. (2021) [16] developed a Multimodal SER model using a Cross- and Self-attention Network (MCSAN) to integrate acoustic and textual information. The proposed approach explicitly modeled both inter-modal interactions (between audio and text) and intra-modal interactions (within each modality) to enhance emotion classification performance. The core architecture consisted of parallel cross-attention and self-attention modules to improve feature fusion and contextual learning.

The cross-attention module facilitated information exchange between audio and text by aligning and propagating relevant emotional cues across modalities. The mechanism enabled the model to capture correlations between speech prosody and linguistic content, addressing challenges in multimodal fusion. Meanwhile, the self-attention module strengthened intra-modal dependencies, refining acoustic and textual representations for more accurate emotion classification. Convolutional and BiLSTM layers were incorporated in the audio encoder to extract temporal and spectral features from MFCCs, while textual embeddings were processed through BiLSTM layers to preserve word-level dependencies.

The model was evaluated on the IEMOCAP and MELD datasets, where it outperformed existing multimodal SER approaches. On IEMOCAP, MCSAN achieved a weighted accuracy of 61.2% and an unweighted accuracy of 56.0%, demonstrating improvements over prior cross-modal attention-based models. On the MELD dataset, the model achieved an F1 score of 59.2%, surpassing semi-supervised approaches that leveraged unlabeled data. The study confirmed that the removal of either the cross-attention or self-attention module reduced classification performance, highlighting the importance of explicitly modeling both intra- and inter-modal dependencies.

The findings demonstrated that the integration of cross- and self-attention mechanisms significantly improved SER performance by effectively capturing complementary information from speech and text. Unlike previous multimodal approaches that either relied on pre-aligned audio-text pairs or independently trained models for each modality, the proposed framework dynamically learned cross-modal relationships without requiring strict alignment constraints.

III. METHODOLOGY

Users interact with the mood recognition system by providing voice input through the web interface, where their recorded speech is processed in real-time for emotion analysis. The system architecture comprises multiple interconnected components, each playing a crucial role in transforming raw audio into meaningful mood insights.

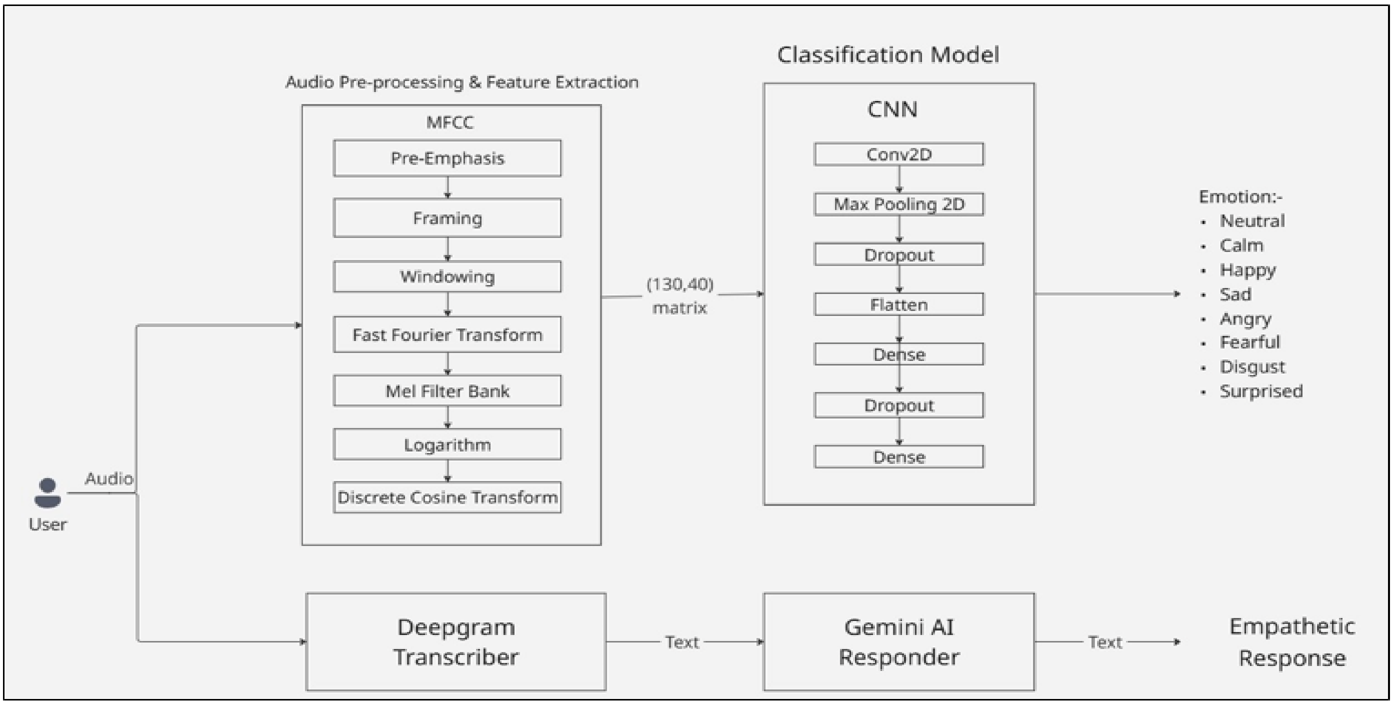


Fig 1 The System Architecture of the Voice Based Mood Recognition System

Figure 1 represents the architecture of the proposed system. The front-end interface, developed as a web-based application, allows users to record their voice, which is then transmitted to the backend for processing. The backend, built using Flask, orchestrates the data flow by handling audio preprocessing, feature extraction, and model inference.

At the core of the system, the AI model extracts key features from speech signals using MFCCs and Mel-Spectrograms, which are subsequently processed through a CNN model. The CNN layers recognize intricate speech patterns, ensuring robust mood classification. The final prediction is generated by fully connected layers and categorized into one of the eight predefined emotional states.

Once the mood classification is complete, the system generates an appropriate response tailored to the user's emotional state. This response is presented through textual feedback, and personalized wellness recommendations. The integration of real-time voice processing with deep learning-based emotion recognition allows the system to offer seamless, intuitive, and personalized mental wellness support, empowering users with actionable insights into their emotional well-being.

A. Module Description

➢ *The Proposed System has been Divided into the following set of modules:*

- *Pre-Processing and Feature Extraction:*

The module handles the initial transformation of raw voice input into a format suitable for emotion recognition. It performs key audio processing tasks such as noise reduction and normalization, frame segmentation of the audio signal, extraction of MFCCs, which effectively capture the tonal and frequency-based characteristics of speech relevant to emotional states. The resulting MFCC feature matrix is passed to the classification model for further analysis.

- *Classification Model:*

The module is responsible for detecting the emotional state of the user based on the extracted features. It uses a CNN architecture trained on a labeled RAVDESS dataset. CNN layers extract hierarchical features and classify emotions. The module acts as the core of the emotion recognition system.

- *Transcription and Empathetic Response:*

The module enhances user experience by understanding and responding to their voice inputs meaningfully. Utilizes the Deepgram API to convert speech to text with high accuracy. The transcribed text is passed to Google Gemini, a large language model, which generates an empathetic and context-aware response tailored to the user's emotional state. The output is presented back to the user in text form via the interface

➢ *User Interface*

The interface is designed to allow users to easily record their voice and receive real-time feedback on their emotional state. The layout is kept minimal and distraction-free to maintain focus on the core functionality, enhancing clarity, usability, and overall user engagement across sessions. The web application serves as the primary interface for users to interact with the mood recognition system. Voice Recording & Submission feature allows users to record their speech, which is then sent to the backend for processing.

The system generates responses based on the predicted mood, including text-based feedback, and personalized wellness suggestions.

The UI Module interacts with the Backend Processing Module to transmit recorded audio and receive analysis results.

- *Backend Processing Module*

The backend processing module is responsible for handling audio data, managing communication between the UI and AI model, and returning mood predictions to the user interface. It ensures seamless integration of various components by preprocessing the audio input, extracting relevant features, and feeding them into the AI model for analysis. The module also handles the execution of machine learning algorithms, managing real-time data flow, and providing timely, accurate mood predictions that are then transmitted back to the UI for user display.

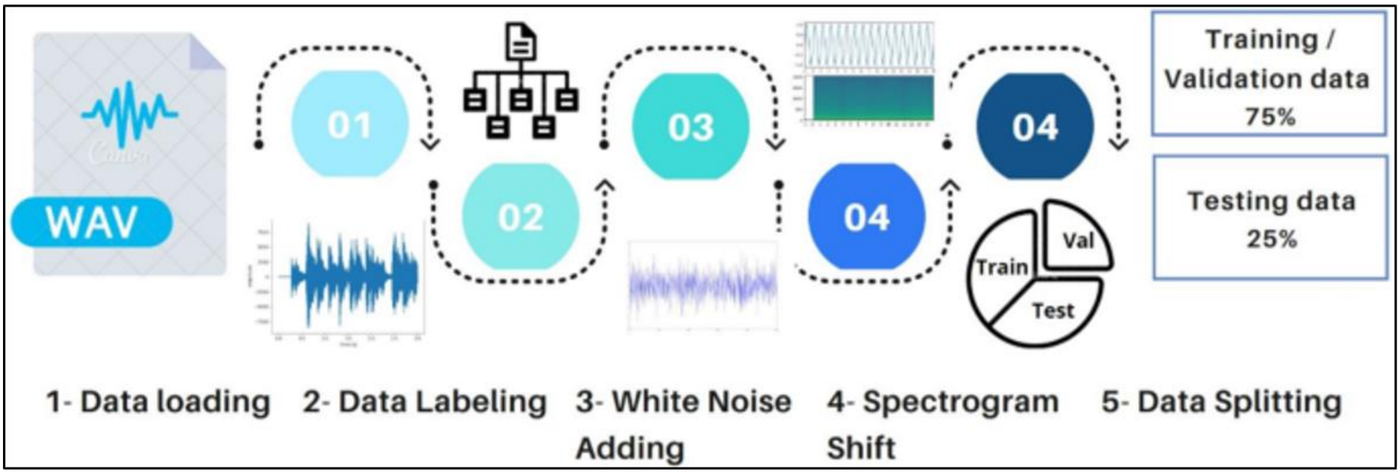


Fig 2 Data Preparation Phase

Extracting meaningful features from raw audio signals is crucial for accurate emotion classification as shown in Figure 2. The first step involves waveform conversion, which transforms raw audio into a structured data format. This is followed by the extraction of MFCCs, which capture essential frequency-based features that help differentiate emotions. Additionally, Mel-Spectrogram features are utilized to analyze variations in speech tone and pitch, providing a comprehensive spectral representation of the audio signal. To ensure consistency and improve model performance, standardization techniques are applied, which help maintain uniformity in feature representation across different samples. To enhance the diversity of the dataset and improve model generalization, various data augmentation techniques are applied.

- *White Noise Addition:*

One of the key techniques used is white noise addition, where controlled levels of noise are introduced to simulate real-world environments with background disturbances. This ensures that the model is trained to recognize emotions even in noisy conditions. The Signal-to-Noise Ratio (SNR) is carefully controlled at 20dB to maintain clarity while enhancing robustness. By adding white noise, the model becomes better equipped to handle real-world scenarios where background noise is present, improving its accuracy in emotion classification. This enhancement allows the model to generalize better across diverse environments, making it more robust to variations in input quality.

- *Spectrogram Shifting:*

Another augmentation technique applied is spectrogram shifting, which involves adjusting the spectrogram along the time axis to simulate variations in speech timing due to different emotional states. Emotions such as anger or sadness can influence the pace of speech, and shifting the spectrogram helps the model learn more invariant features.

- *Data Splitting and Model Training:*

Once the data has been preprocessed and augmented, it is split into training, validation, and testing sets. 80% of the dataset is allocated for training and validation, while the remaining 20% is used for testing. This division ensures that the model is trained on a substantial portion of the data while also being evaluated on unseen samples to assess its performance accurately. The preprocessed and structured data is then fed into the SER model, which learns to classify emotions based on extracted features.

- *ML Model Module*

- *Feature Extraction:*

Audio recordings are transformed into structured features that the model can process.

- *Responsibilities:*

- Extracting MFCC and Mel-Spectrogram features from voice inputs.
- Capturing pitch, tone, and speech variations that indicate emotional states.
- Preparing structured data for CNN model processing.

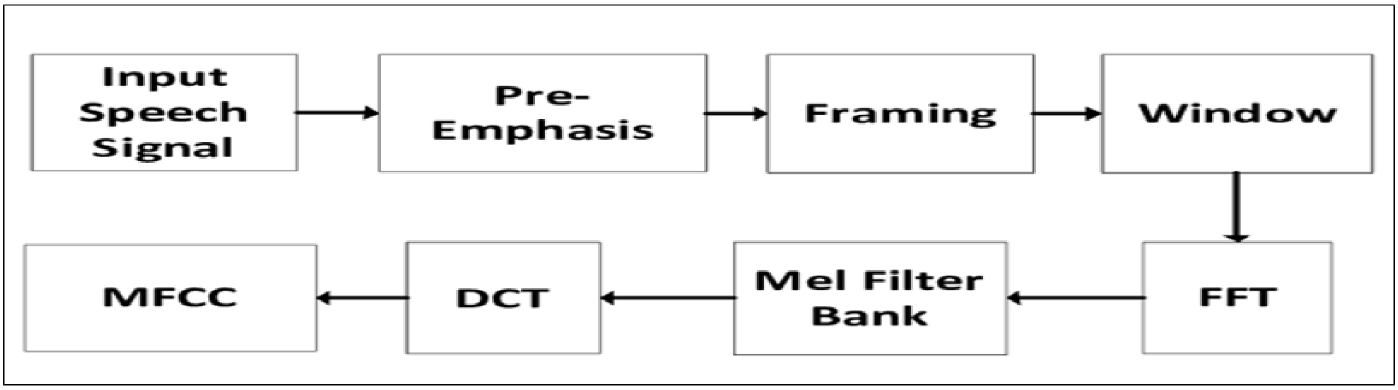


Fig 3 Mel Frequency Cepstral Coefficients

MFCCs are widely used in speech processing because they provide an accurate estimation of speech while maintaining computational efficiency. This technique transforms short-term power spectra of sound into a linear cosine transform of the log power spectrum on a nonlinear Mel scale. The MFCC technique represents speech signals by converting their short-term power spectrum into a linear cosine transform of the logarithmic power spectrum on a nonlinear Mel scale of frequency, as illustrated in Figure 3. The Mel scale reflects human auditory perception, particularly its sensitivity to lower frequencies. The steps involved in computing MFCCs include pre-emphasis, framing, windowing, Fast Fourier Transform (FFT), Mel filter bank application, logarithm scaling, and Discrete Cosine Transform (DCT).

The proposed model is composed of three primary components: the convolutional feature extraction block, the fully connected classification block, and regularization through dropout layers. The feature extraction block includes two convolutional layers with 32 and 64 filters, respectively. Each convolutional layer is followed by a ReLU activation function and a max pooling layer, which reduces the spatial dimensions of the feature maps while preserving critical information. Dropout layers are introduced after the pooling and dense layers to minimize overfitting during training. The flattened output is passed to a dense layer with 128 units, followed by another dropout layer, and finally to an output layer with 8 units corresponding to the emotion classes.

Finally, the classification component is responsible for predicting the emotion category of the input speech. It consists of a dense output layer, where the number of neurons corresponds to the total number of emotion classes in the dataset. A Softmax activation function is applied to normalize the output probabilities, ensuring that the model assigns an appropriate likelihood to each emotion category. By focusing on spatial features extracted through the CNN layers, the architecture supports accurate emotion classification based on relevant patterns in the input.

➢ *Response Generation Module*

Once the model predicts the user’s mood, the backend formulates appropriate responses and sends them to the UI module.

➢ *Responsibilities:*

- Structuring mood prediction results for display.
- Sending relevant recommendations based on the classified emotion.
- Ensuring secure data transmission between the UI and AI model.

➢ *Real-time SER System*

Developing a deep learning model for recognition tasks, including Speech Emotion Recognition (SER), is a complex task that extends beyond achieving high recognition rates and strong evaluation metrics. To fully validate an SER system, it must be tested in real-time environments to ensure its practical applicability. While achieving high learning accuracy (even up to 100%) may seem impressive, it is crucial that the system maintains its performance when deployed in real-time applications, where factors like noise and variability can affect its reliability.

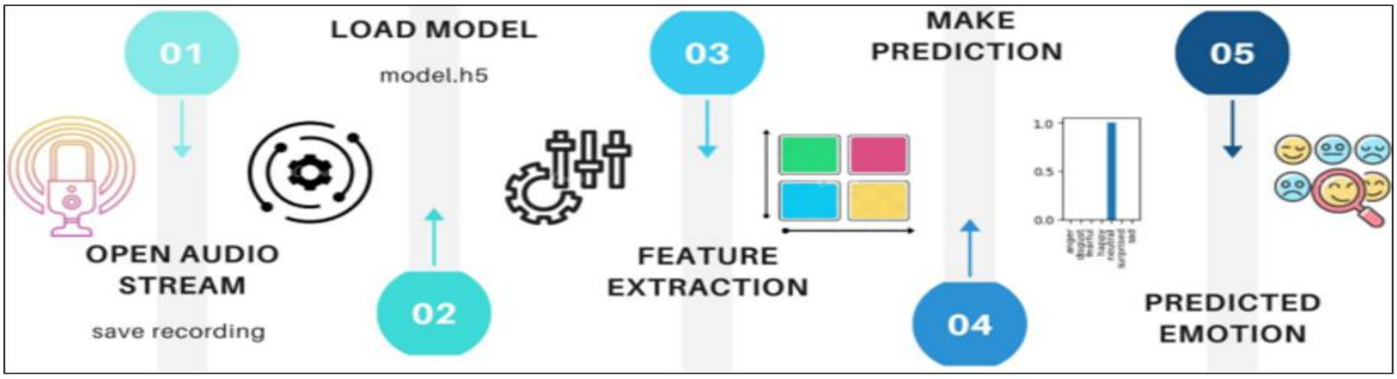
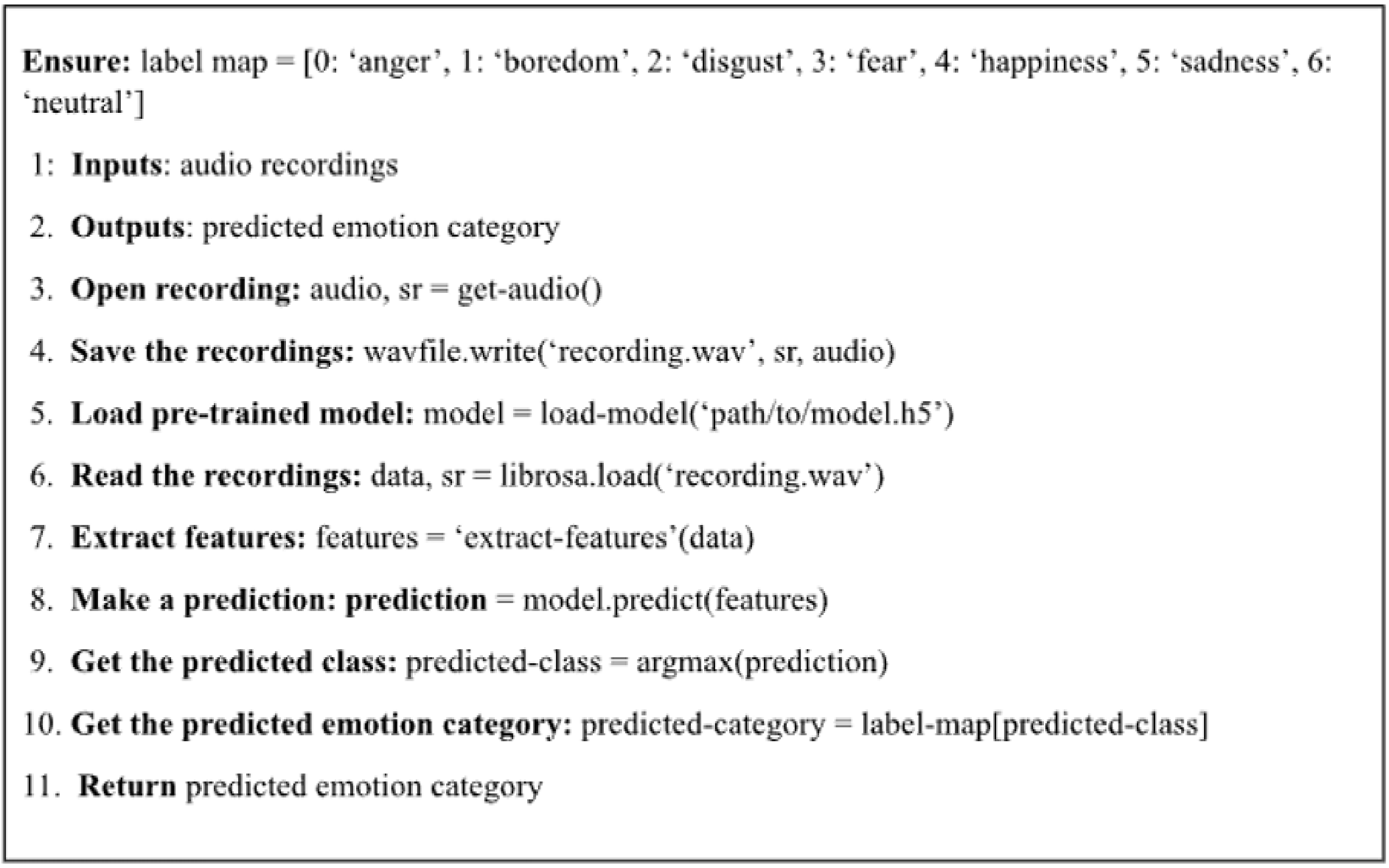


Fig 4 Speech Emotion Recognition Process

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may894

As illustrated in Figure 4, the real-time SER process begins with inputting a live audio stream. The pre-trained deep learning model, built on a CNN architecture, is then loaded, as it has demonstrated superior performance in terms of accuracy, precision, recall, and other evaluation metrics.

Once the model is loaded, the same feature extraction techniques applied to stored recordings are utilized for the live audio input. Finally, the extracted features are processed through the model to generate emotion predictions. Real-time testing involves evaluating the system using live audio streams, providing a more accurate representation of how it would perform in real-world scenarios. The outcome of these tests determines whether the system functions correctly and efficiently under real-time conditions. Thus, the ultimate success of an SER system relies on its ability to operate accurately, reliably, and efficiently in real-time settings.



```
Ensure: label map = [0: 'anger', 1: 'boredom', 2: 'disgust', 3: 'fear', 4: 'happiness', 5: 'sadness', 6: 'neutral']
1:  Inputs: audio recordings
2.  Outputs: predicted emotion category
3.  Open recording: audio, sr = get-audio()
4.  Save the recordings: wavfile.write('recording.wav', sr, audio)
5.  Load pre-trained model: model = load-model('path/to/model.h5')
6.  Read the recordings: data, sr = librosa.load('recording.wav')
7.  Extract features: features = 'extract-features'(data)
8.  Make a prediction: prediction = model.predict(features)
9.  Get the predicted class: predicted-class = argmax(prediction)
10. Get the predicted emotion category: predicted-category = label-map[predicted-class]
11. Return predicted emotion category
```

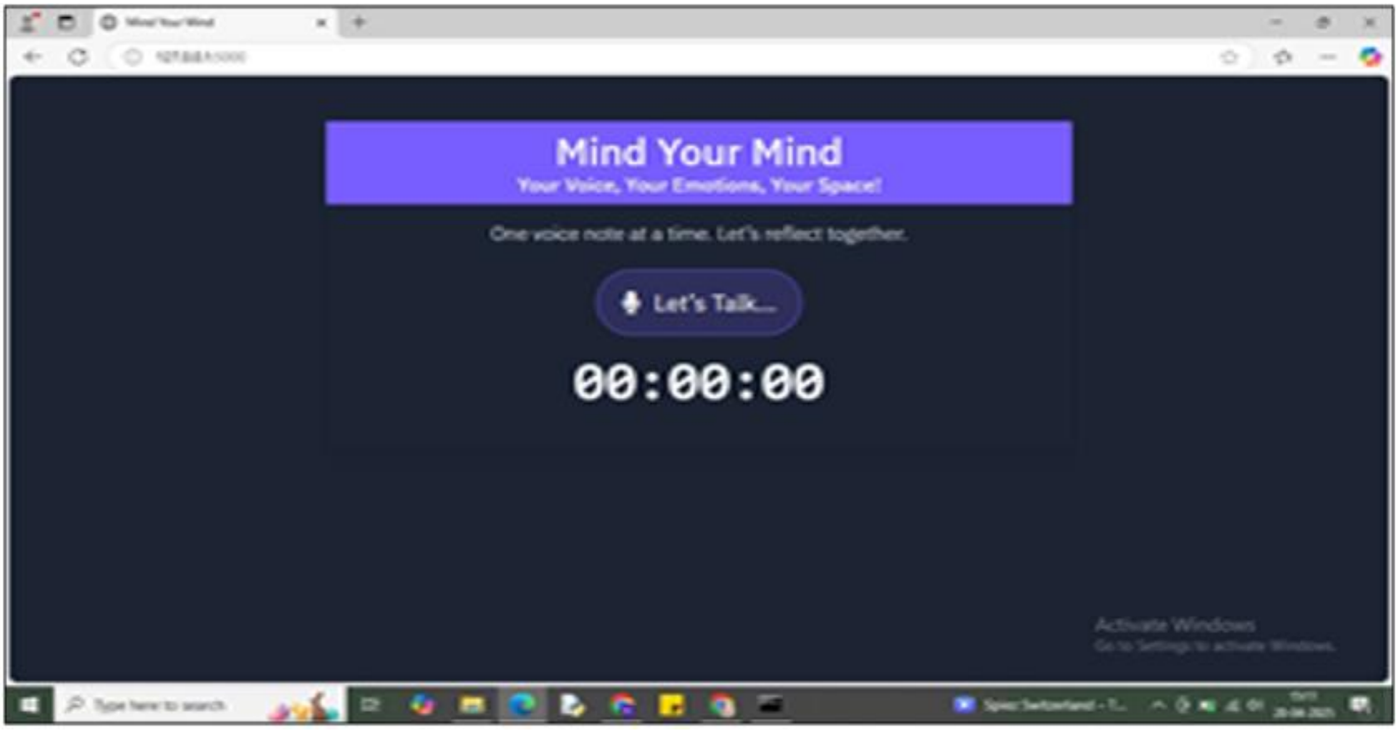
Fig 5 Ensure: label map

Algorithm 1 provides a detailed breakdown of the steps involved in this real-time SER process, where an audio recording is taken as input and classified into its corresponding emotional category.

The AI Model Module ensures high-accuracy mood classification by utilizing advanced deep learning techniques. It processes the extracted audio features using a CNN architecture, which excels in spatial feature extraction. This approach enables the model to analyze the acoustic properties of speech, leading to more precise mood classification. By focusing on both low-level and high-level speech features, the CNN efficiently captures nuanced emotional cues that are critical for mood recognition. Additionally, the model's ability to process features in real-time allows for quick, adaptive responses to emotional shifts. By integrating deep learning with real-time processing capabilities, the AI Model Module delivers meaningful and context-aware mood insights, empowering users with a more personalized and interactive emotional well-being experience.

IV. RESULT

This chapter explores the actual implementation of all the modules in the proposed system and presents the results obtained from the implementation process.

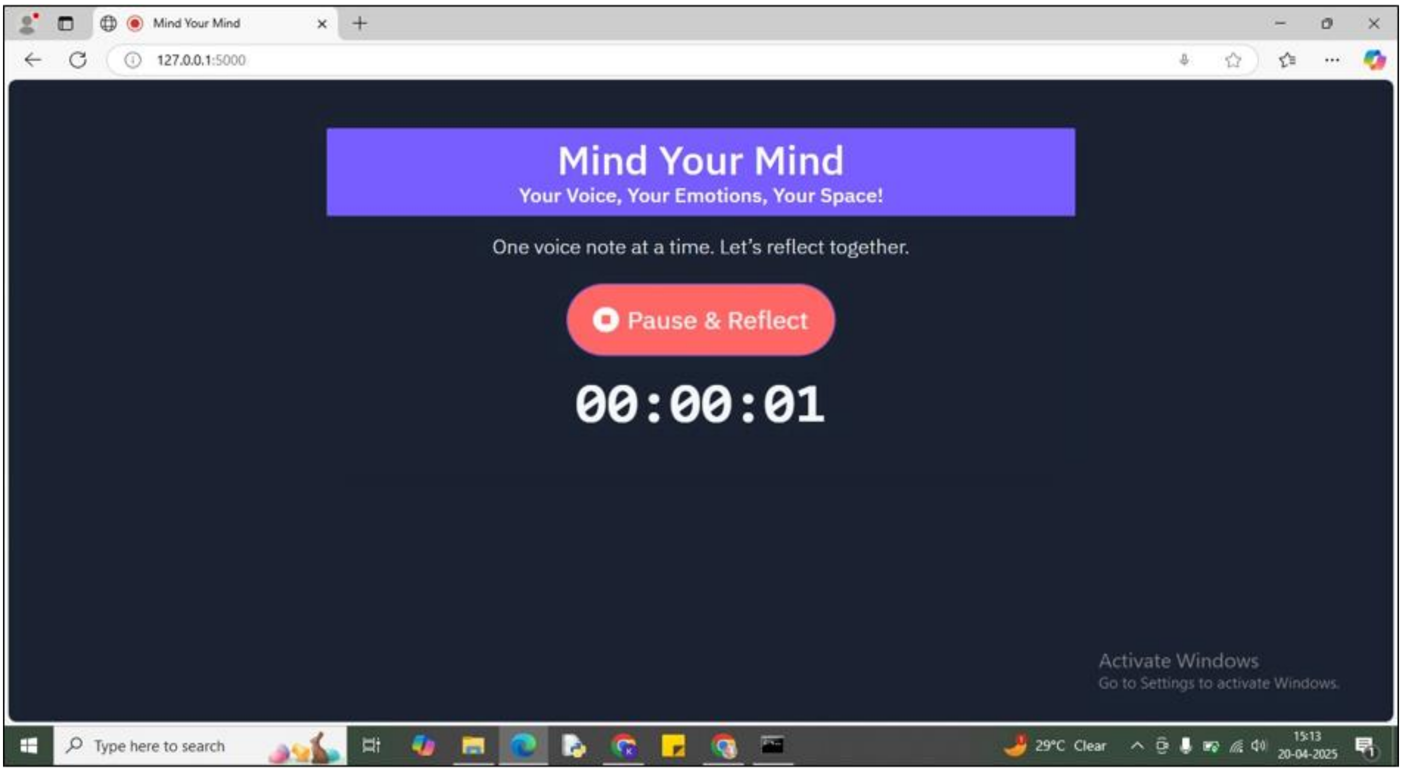


Mind Your Mind
Your Voice, Your Emotions, Your Space!

One voice note at a time. Let's reflect together.

Fig 6 MYM Home Page

Figure 6 illustrates the voice recording interface of the "Mind Your Mind" application. This page acts as the primary interaction point where users begin their emotional journaling journey. The interface is designed with a soothing visual theme, featuring a prominently displayed banner with the application's tagline "Your Voice, Your Emotions, Your Space!" which reinforces the platform's goal of providing a safe, personal space for emotional expression. At the center of the interface is a microphone-enabled button labeled "Let's Talk...", inviting users to start recording their voice inputs. Upon activation, a timer is initiated to track the duration of the recording, helping users stay aware of the time they spend reflecting. This setup encourages mindful interaction and ensures the experience is intuitive, accessible, and focused on self-expression. The simplicity and clarity of the layout support a user-friendly experience, making it easy even for first-time users to navigate and engage with the emotional logging feature.

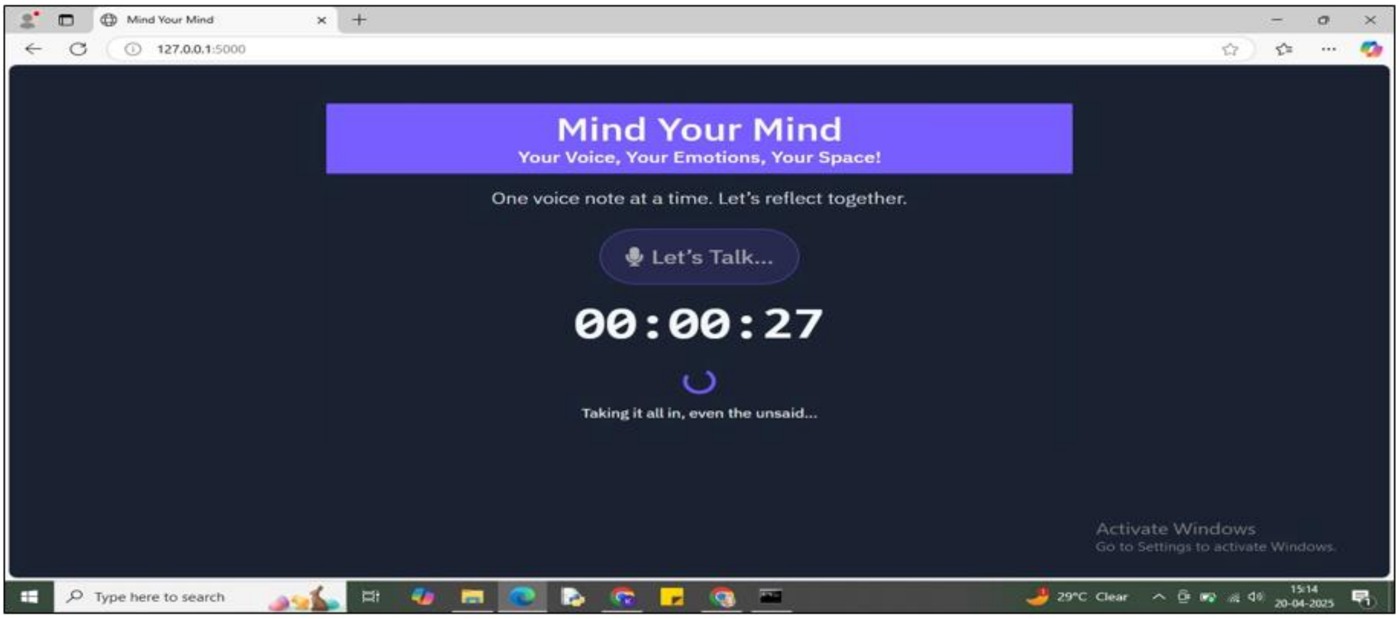


Mind Your Mind
Your Voice, Your Emotions, Your Space!

One voice note at a time. Let's reflect together.

Fig 7 Active Voice Recording Interface

Figure 7 displays the "Pause & Reflect" interface of the Mind Your Mind application during an active voice journaling session. The screen features a prominent header with the application's name and tagline: "Your Voice, Your Emotions, Your Space!" emphasizing a safe and personal environment. Beneath the header, a supportive prompt encourages mindful reflection: "One voice note at a time. Let's reflect together." At the centre, a large red button labelled "Pause & Reflect" allows users to halt their recording at any moment, reinforcing the app's focus on intentional and paced self-expression. Just below the button, a real-time digital timer shows the elapsed time of the current recording session, enhancing the user's awareness of their journaling duration.



Mind Your Mind

Your Voice, Your Emotions, Your Space!

One voice note at a time. Let's reflect together.

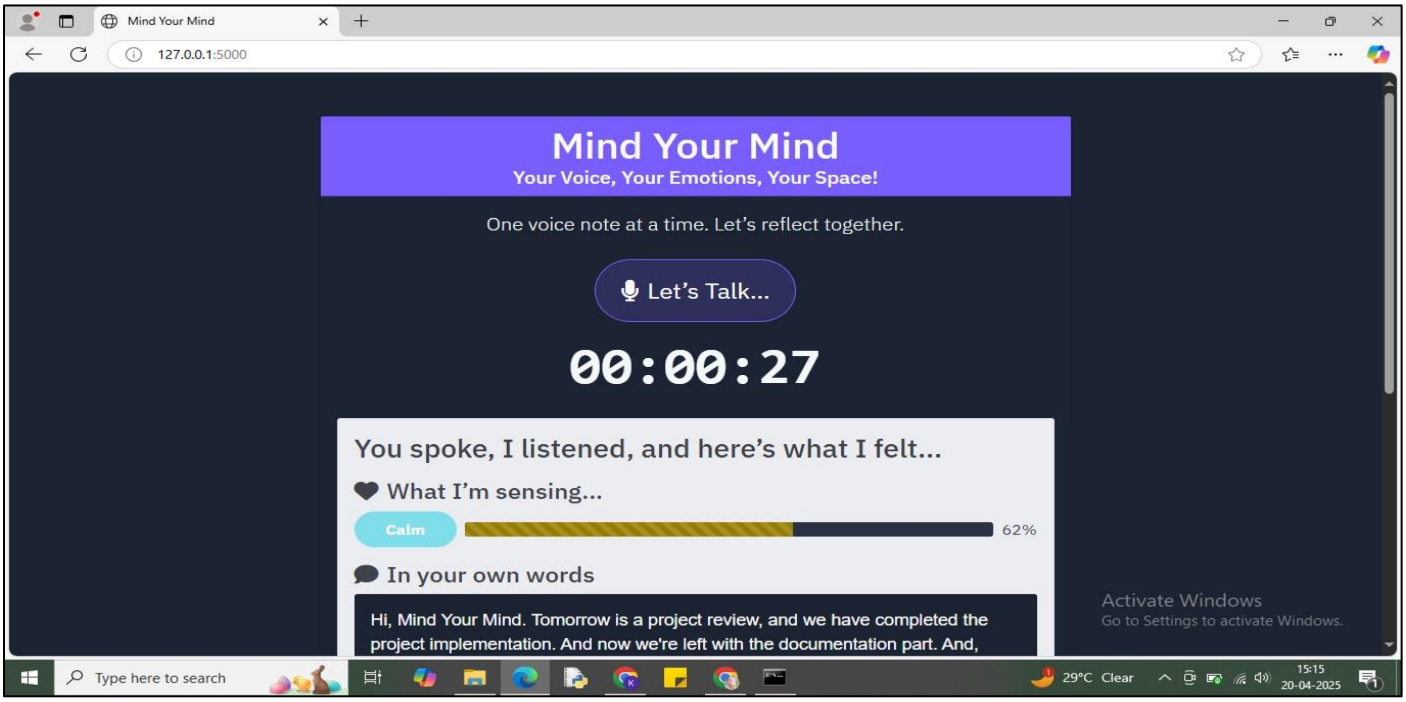
00:00:27

Taking it all in, even the unsaid...

Activate Windows
Go to Settings to activate Windows.

Fig 8 Audio Processing Interface

Figure 8 showcases the voice recording processing interface within the "Mind Your Mind" application. Once the user completes the recording process by clicking the "Pause & Reflect" button, the application transitions into a processing state. The phrase "Taking it all in, even the unsaid…" appears below the timer, gently reinforcing the app's empathetic purpose to provide a non-judgmental space that values every thought and emotion, spoken or unspoken. The progress indicator confirms the user that the system is actively capturing their voice input.



Mind Your Mind

Your Voice, Your Emotions, Your Space!

One voice note at a time. Let's reflect together.

00:00:27

You spoke, I listened, and here's what I felt...

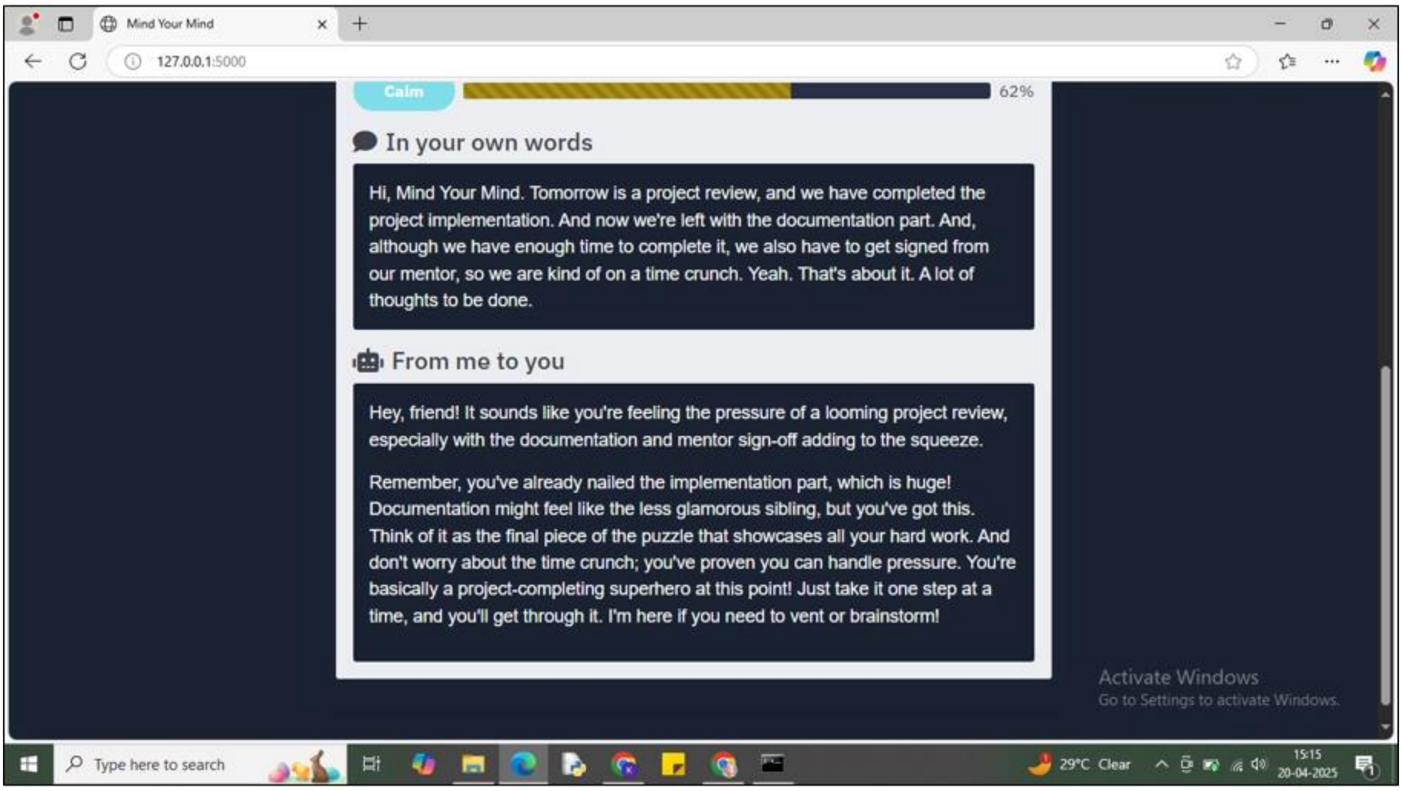
Hi, Mind Your Mind. Tomorrow is a project review, and we have completed the project implementation. And now we're left with the documentation part. And,

Activate Windows
Go to Settings to activate Windows.

Fig 9 Emotion Recognition and Feedback Interface

Figure 9 presents the reflective feedback interface within the Mind Your Mind application following a completed voice note. After the user finishes speaking by interacting with the “Let’s Talk…” button, the application transitions into an analysis phase, offering both emotional interpretation and transcription.

This interface acts as the starting point of the user’s emotional journey in the app, leading into the subsequent phases of analysis and reflection. It sets the tone for a mindful experience, emphasizing the importance of uninterrupted self-expression before any interpretation or feedback is introduced.



In your own words

Hi, Mind Your Mind. Tomorrow is a project review, and we have completed the project implementation. And now we're left with the documentation part. And, although we have enough time to complete it, we also have to get signed from our mentor, so we are kind of on a time crunch. Yeah. That's about it. A lot of thoughts to be done.

From me to you

Hey, friend! It sounds like you're feeling the pressure of a looming project review, especially with the documentation and mentor sign-off adding to the squeeze.

Remember, you've already nailed the implementation part, which is huge! Documentation might feel like the less glamorous sibling, but you've got this. Think of it as the final piece of the puzzle that showcases all your hard work. And don't worry about the time crunch; you've proven you can handle pressure. You're basically a project-completing superhero at this point! Just take it one step at a time, and you'll get through it. I'm here if you need to vent or brainstorm!

Fig 10 Empathetic Feedback Interface

Figure 10 displays the final stage of the reflective journaling process in the Mind Your Mind application. After recording a voice note and receiving emotional analysis, the user is presented with both a transcription of their message and a personalized AI-generated empathetic response. The section titled “In your own words” provides the user with a clear, accurate text representation of their spoken reflections, ensuring they can review and reconnect with their original thoughts.

Following this, the “From me to you” segment acts as a supportive and empathetic voice from the system. It delivers an encouraging message based on the user's emotional tone and content, offering reassurance, validation, and motivational suggestions. This approach nurtures emotional well-being by making users feel heard and supported, effectively mimicking a compassionate human response.

V. PERFORMANCE EVALUATION

The performance evaluation of a machine learning model is essential to assess its generalization capability and robustness, especially for real-world applications. In the context of a speech sentiment analysis model, performance evaluation provides insight into how well the model can classify emotions from speech signals. For this project, we used the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset to train and test our model. The model was evaluated across multiple performance metrics: precision, recall, F1-score, and overall accuracy.

A. *Metrics Used for Evaluation*

The model was evaluated using the following standard classification metrics:

➢ *Precision:*

The ratio of correctly predicted positive observations to the total predicted positives. It indicates the accuracy of positive predictions made by the model.

$$\text{Precision} = TP \div (TP + FP) \qquad 6.1$$

TP = True Positives

FP = False Positives

➢ *Recall (Sensitivity):*

The ratio of correctly predicted positive observations to all observations in actual class. It shows how well the model can detect positive classes.

$$\text{Recall} = TP \div (TP + FN) \qquad 6.2$$

TP = True Positives

FN = False Negatives

- *F1-Score:*

The harmonic mean of precision and recall, providing a balance between them. It is particularly useful when the class distribution is imbalanced.

$$\text{F1-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \qquad 6.3$$

- *Accuracy:*

The overall correctness of the model, measured as the ratio of correct predictions to the total number of predictions made, reflecting the model's general performance across all emotion categories.

$$\text{Accuracy} = (TP + TN) \div (TP + TN + FP + FN) \qquad 6.4$$

TN = True Negatives

B. Classification Report

The following table presents a comprehensive classification report for the emotion recognition model. It outlines key performance metrics like precision, recall, F1-score, and support for each emotion category. These metrics provide valuable insight into the model’s ability to accurately identify and differentiate between various emotional states, highlighting both its strengths and areas for potential improvement.

Table 1 Classification Report of Emotions

	Precision	**Recall**	**F1-score**	**Support**
Angry	0.90	0.97	0.94	76
Calm	0.92	0.95	0.94	77
Disgust	1.00	0.87	0.93	77
Fearful	0.90	0.95	0.92	77
Happy	0.94	0.88	0.91	77
Neutral	0.86	1.00	0.93	38
Sad	0.92	0.90	0.91	77
Surprised	0.96	0.95	0.95	77
Accuracy			0.93	576
Macro avg	0.93	0.93	0.93	576
Weighted avg	0.93	0.93	0.93	576

Table 1 presents a detailed classification report evaluating the performance of the emotion recognition model used in the application. The report includes standard evaluation metrics such as precision, recall, F1-score, and support for each of the eight emotion categories: angry, calm, disgust, fearful, happy, neutral, sad, and surprised. These metrics provide a comprehensive assessment of the model's ability to accurately detect and classify emotions, highlighting its strengths and potential areas for improvement in various emotional contexts.

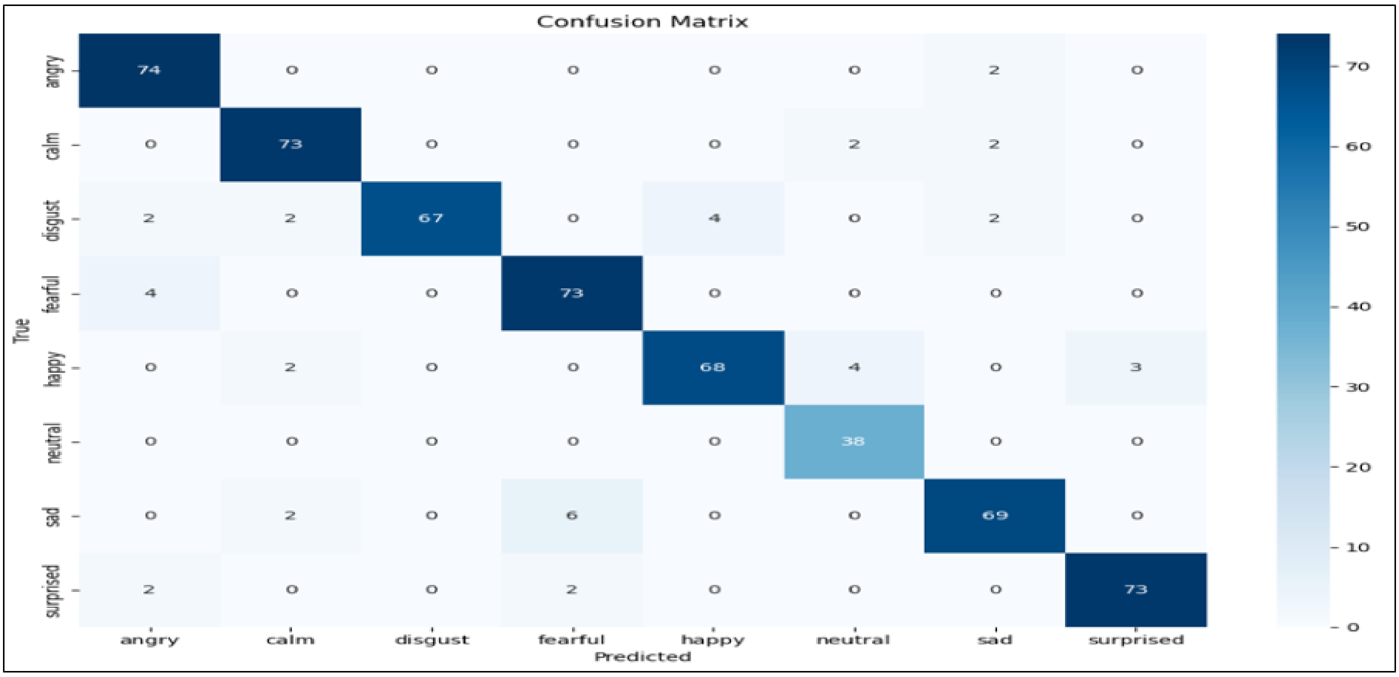


Fig. 11 Emotion Classification - Confusion Matrix

Figure 11 illustrates the confusion matrix for the emotion classification model, offering insight into how well the system differentiates between various emotional states. The matrix presents true labels on the vertical axis and predicted labels on the horizontal axis, with each cell indicating the number of instances classified into each category.

C. Model Performance Analysis

The model achieved an overall accuracy of 93%, indicating a strong ability to correctly classify the speech sentiment for most inputs. This is a high level of accuracy, suggesting that the model is well-calibrated and performs well across all emotion classes.

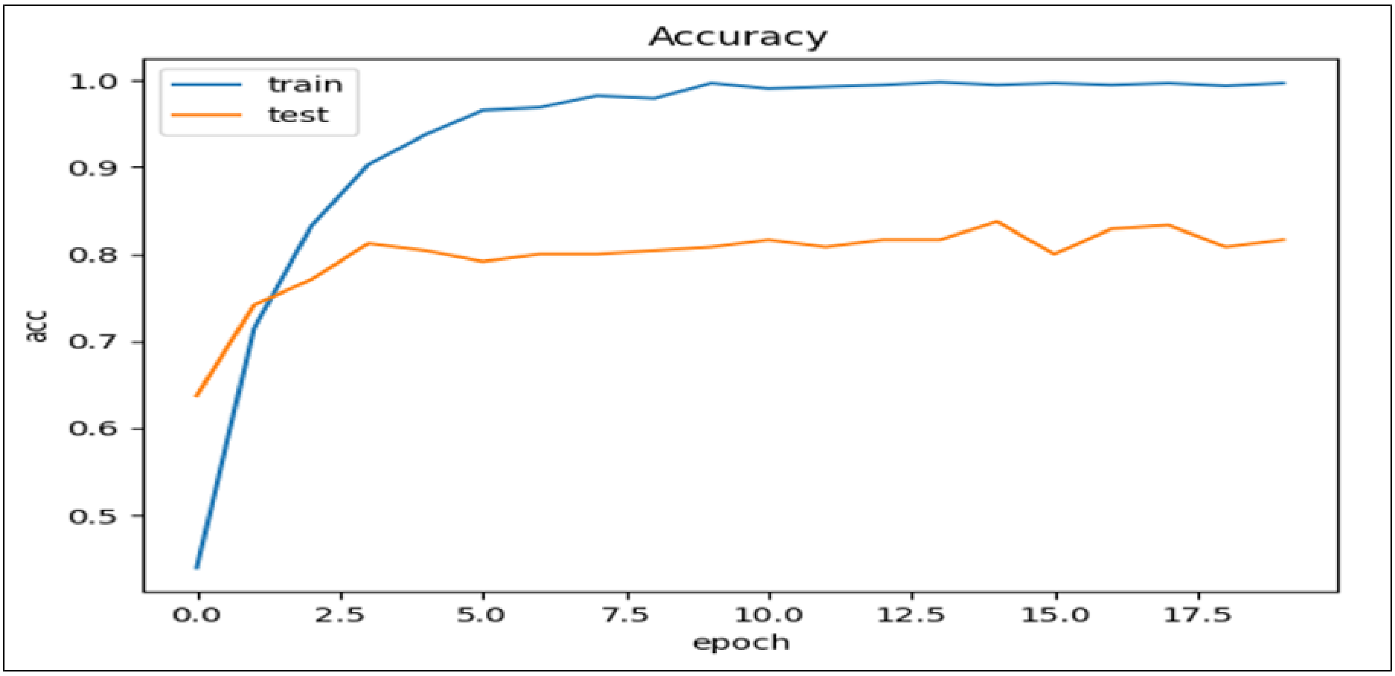


Fig 11 Training and Testing Accuracy over Epochs

Figure 11 illustrates the accuracy progression of both the training and testing datasets across 20 training epochs. The blue line represents the training accuracy, while the orange line depicts the testing accuracy. This visual comparison highlights how the model's performance evolves over time and helps identify any signs of overfitting or underfitting.

➢ *Emotion-Specific Performance:*

- *Angry:*

The model shows a high precision of 0.90 and recall of 0.97, resulting in a solid F1-score of 0.94. This suggests that the model correctly identifies angry emotions with high accuracy, though there is a slight trade-off with false positives.

- *Calm:*

Precision is 0.92 and recall is 0.95, with an F1-score of 0.94, showing that the model has a balanced performance in recognizing calm emotions.

- *Disgust:*

Although the model has perfect precision (1.00), its recall is 0.87, resulting in a lower F1-score of 0.93. This indicates that while the model is highly confident in predicting disgust, it misses a few instances, leading to a slightly lower recall.

- *Fearful:*

Precision and recall for fearful emotions are both relatively high (0.90 and 0.95 respectively), yielding an F1-score of 0.92, showing reliable detection with minimal errors.

- *Happy:*

A precision of 0.94 and recall of 0.88 results in an F1-score of 0.91, showing that the model is good at predicting happiness but may miss a few cases (lower recall).

- *Neutral:*

Despite a high recall (1.00), the precision of 0.86 for neutral emotions results in a slightly lower F1-score (0.93). The perfect recall suggests that the model is sensitive to neutral emotions, though it may generate more false positives.

- *Sad***:**

Precision of 0.92 and recall of 0.90 yield an F1-score of 0.91, indicating good detection of sadness with a balanced trade-off between false positives and false negatives.

- *Surprised:*

The highest-performing emotion, with a precision of 0.96, recall of 0.95, and an F1-score of 0.95. This indicates the model performs very well in detecting surprise with minimal errors.

D. Macro and Weighted Averages

- *Macro Average:*

The model achieves a macro average precision, recall, and F1-score of 0.93, indicating that it performs equally well across all classes, suggesting a well-balanced performance, where the model doesn't favor any particular class over others. A macro average considers the performance of each class individually before averaging the scores, which makes it particularly useful in scenarios where class imbalance might otherwise skew the results. By achieving such a high score, the model demonstrates both effectiveness and fairness in its predictions across all categories.

- *Weighted Average:*

The weighted average precision, recall, and F1-score are also 0.93, which takes into account the class distribution. The result indicates that the model performs consistently well even in the presence of class imbalances (e.g., fewer samples of neutral emotions), suggesting that the model is effectively handling underrepresented classes while maintaining high overall performance.

E. Model Strengths and Weaknesses

- *Strengths:*

The model exhibits high accuracy and reliable classification across various emotional categories. Notably, the high recall for neutral emotions demonstrates the model's ability to effectively detect emotions with low representation in the dataset. Additionally, the model performs exceptionally well with both high precision and recall for surprised and disgusted emotions, indicating strong performance in accurately identifying these emotional states.

- *Weaknesses:*

The model shows a slight imbalance between precision and recall in some categories like neutral and disgust. While precision is relatively high, the recall in these categories could be improved. Specifically, there is room for improvement in the recall for the disgust category to ensure more complete detection of emotions, thereby reducing false negatives and improving the model's ability to accurately identify and classify these emotions.

F. CNN vs LSTM Classification Report

The CNN vs LSTM Classification Report provides a direct comparison of the performance of CNN and LSTM networks in the task of speech emotion recognition across various emotional categories.

Table 2 CNN vs LSTM Classification Report of Emotions

Emotion	Precision (CNN)	Recall (CNN)	F1-Score (CNN)	Precision (LSTM)	Recall (LSTM)	F1-Score (LSTM)
Neutral	0.84	0.81	0.82	0.85	0.84	0.84
Calm	0.79	0.76	0.77	0.82	0.80	0.81
Happy	0.88	0.86	0.87	0.89	0.88	0.88
Sad	0.82	0.79	0.80	0.83	0.82	0.82
Angry	0.85	0.88	0.86	0.87	0.89	0.88
Fearful	0.80	0.78	0.79	0.82	0.80	0.81
Disgust	0.77	0.74	0.75	0.80	0.78	0.79
Surprised	0.86	0.89	0.87	0.88	0.90	0.89
Accuracy			0.82			0.85
Macro Avg	0.83	0.82	0.82	0.85	0.84	0.84
Weighted Avg	0.83	0.82	0.82	0.85	0.84	0.84

Table 2 compares the performance of CNN and LSTM models on a speech emotion recognition task. It provides precision, recall, F1-score, and support for each of the eight emotion categories (angry, calm, disgust, fearful, happy, neutral, sad, surprised), offering valuable insights.

VI. CONCLUSION

MYM has successfully developed an audio-based mood recognition system using deep learning techniques. The system employs advanced spectral feature extraction methods, such as MFCC and Mel Spectrogram, to capture key acoustic properties from user-recorded speech. A CNN-based model was implemented to analyze these features, effectively identifying mood-related patterns in the input data.

Unlike traditional real-time SER systems, this approach focuses on recorded audio rather than live input, ensuring reliable processing and analysis. The model was trained and tested on benchmark datasets such as RAVDESS, where CNN demonstrated strong performance in emotion classification. The integration with transcription via Deepgram and response generation via Gemini further enhances the user experience, providing meaningful emotional insights in a non-real-time but interactive setting.

FUTURE WORKS

The future trajectory of the project focuses on enhancing emotion recognition and building a more intuitive user experience. Planned updates include an interactive dashboard for emotional trends, voice journaling history, mood reflections, and achievement badges. The system will also integrate adaptive noise cancellation and vocal trait adjustments for improved, personalized emotion detection.

REFERENCES

[1]. Barhoumi, Chawki, and Yassine BenAyed. "Real-time speech emotion recognition using deep learning and data augmentation." Artificial Intelligence Review 58.2 (2024): 49.

[2]. Sayis, Batuhan, and Hatice Gunes. "Technology-assisted journal writing for improving student mental wellbeing: Humanoid robot vs. voice assistant." Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction.

[3]. Chawla, Shreya, and Sneha Saha. "Exploring perceptions of psychology students in Delhi-NCR Region towards using mental health apps to promote resilience: a qualitative study." BMC Public Health 24.1 (2024): 2000.

[4]. Olawade, David B., et al. "Enhancing mental health with Artificial Intelligence: Current trends and future prospects." Journal of Medicine, Surgery, and Public Health (2024): 100099.

[5]. Liu, Z. "Online hate speech on Twitter from the perspective of pragmatics." International Journal of Social Sciences and Public Administration 4.1 (2024): 322-326.

[6]. Simmons, Natalie, Lewis Goodings, and Ian Tucker. "Experiences of using mental health Apps to support psychological health and wellbeing." Journal of Applied Social Science 18.1 (2024): 32–44.

[7]. Aipenova, Aziza, and Seitmukhanova Almira. "Mental Health in the Digital Age: Balancing Connectivity and Well-Being." Journal of Spirituality in Mental Health. 7 (2024).

[8]. Hamdoun, Salah, et al. "AI-based and digital mental health apps: Balancing need and risk." IEEE Technology and Society Magazine 42.1 (2023): 25–36.

[9]. Tucker, Ian, Katherine Easton, and Rebecca Prestwood. "Digital community assets: Investigating the impact of online engagement with arts and peer support groups on mental health during COVID-19." Sociology of Health & Illness 45.3 (2023): 666–683.

[10]. Pan, Jiahui, et al. "Multimodal emotion recognition based on facial expressions, speech, and EEG." IEEE Open Journal of Engineering in Medicine and Biology (2023).

[11]. Chamishka, Sadil, et al. "A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling." Multimedia Tools and Applications 81.24 (2022): 35173–35194.

[12]. Ravuri, Vinesh, Ricardo Gutierrez-Osuna, and Theodora Chaspari. "Preserving mental health information in speech anonymization." 2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). IEEE, 2022.

[13]. Andayani, Felicia, et al. "Hybrid LSTM-transformer model for emotion recognition from speech audio files." IEEE Access 10 (2022): 36018–36027.

[14]. Mertens, Esther CA, et al. "Parallel changes in positive youth development and self-awareness: The role of emotional self-regulation, self-esteem, and self-reflection." Prevention Science 23.4 (2022): 502–512.

[15]. Kakuba, Samuel, Alwin Poulose, and Dong Seog Han. "Deep learning-based speech emotion recognition using multi-level fusion of concurrent features." IEEE Access 10 (2022): 125538–125551.

[16]. Sun, Licai, et al. "Multimodal cross-and self-attention network for speech emotion recognition." ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.

[17]. Muppidi, Aneesh, and Martin Radfar. "Speech emotion recognition using quaternion convolutional neural networks." ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.

[18]. Sajjad, Muhammad, and Soonil Kwon. "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM." IEEE access 8 (2020): 79861-79875.

[19]. Zhang, Jianhua, et al. "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review." Information fusion 59 (2020): 103-126.

[20]. Yao, Zengwei, et al. "Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN." Speech Communication 120 (2020): 11-19.

[21]. Nasri, M. A., et al. "Face emotion recognition from static image based on convolution neural networks." 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). IEEE, 2020.

[22]. Mellouk, Wafa, and Wahida Handouzi. "Facial emotion recognition using deep learning: review and insights." Procedia Computer Science 175 (2020): 689–694.

[23]. Twenge, Jean M., et al. "Age, period, and cohort trends in mood disorder indicators and suicide-related outcomes in a nationally representative dataset, 2005–2017." Journal of Abnormal Psychology 128.3 (2019): 185.

[24]. Mirsamadi, Seyedmahdad, Emad Barsoum, and Cha Zhang. "Automatic speech emotion recognition using recurrent neural networks with local attention." 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

[25]. Huang, Yongrui, et al. "Fusion of facial expressions and EEG for multimodal emotion recognition." Computational Intelligence and Neuroscience 2017.1 (2017): 2107451.