Emotional Disparity: Bridging Facial and Speech Emotion Analysis with AI for Enhanced Psychiatric Care

Yedhu Krishnan[1†], Roanek Jena[2†], Manju G.[3*]

[1*]Student, School of Computer Science and Engineering, Vellore Institute of Technology, Kelambakkam, Chennai, 600127, Tamil Nadu, India.
[2*]Student, School of Computer Science and Engineering, Vellore Institute of Technology, Kelambakkam, Chennai, 600127, Tamil Nadu, India.
[3]Faculty, School of Computer Science and Engineering, Vellore Institute of Technology, Kelambakkam, Chennai, 600127, Tamil Nadu, India.

*Corresponding author(s). E-mail(s): manju.g@vit.ac.in;
Contributing authors: edu.yy226@gmail.com; roanek123@gmail.com;
†These authors have equally contributed to this work.

Abstract

The COVID-19 epidemic has increased the use of online treatment, but it has also brought attention to its shortcomings in comparison to in-person therapy sessions. Online therapy is a practical and easily available alternative, but it does not have the same depth of nonverbal clues and subtle emotional expressions as in-person sessions. This study attempts to close this gap by examining the emotional differences that occur between a user's speech and facial expressions during therapy sessions. It is based in the disciplines of Speech Processing, Natural Language Processing (NLP), and Psychology.

The system will independently analyze the other's covert nature-their speech pieces and emotional expressions-teeth to teeth, using a specifically built dataset constructed by machine learning models. These analyses shall then be compared to reveal and characterize the deviations between them, which could signal underlying emotional conflicts or mismatches. Thus, these findings will be summarized further by an assessment model, allowing easier idioms for the therapist reflecting more on clients' emotional states. The project seeks to provide better emotional awareness for online therapy sessions, thus making virtual counseling an even more appropriate option as an alternative to face-to-face therapy.

Keywords: Deep Neural Networks, Large Language Models, Speech Processing, Emotion Analysis

1 Introduction

The rise of online therapy during the COVID-19 pandemic has transformed the landscape of mental health care, offering unprecedented accessibility to individuals seeking psychological support. However, despite its rapid adoption, virtual therapy is often found to be less effective than traditional face-to-face sessions. One of the key limitations lies in the inability of online platforms to capture the full spectrum of emotional cues that are critical for effective therapy—such as facial expressions, vocal tone, and body language. These non-verbal signals are essential for therapists to accurately assess their clients' emotional states and provide appropriate guidance.

This project, operating at the intersection of Speech Processing, Natural Language Processing (NLP), and Psychology, seeks to address this gap by analyzing the emotional disparity between a user's speech and facial expressions. The goal is to independently evaluate both modalities—speech and facial expressions—and identify discrepancies that may reveal hidden or conflicting emotions. A custom dataset and machine learning models will be used to conduct these analyses, with the results synthesized into a summarized output by a text generator model, making it easily interpretable for therapists.

Relevance and Viability: As the demand for online therapy continues to grow, the relevance of improving emotional understanding in virtual sessions is more pressing than ever. By enabling therapists to detect emotional mismatches between what a client says and what their facial expressions convey, this project has the potential to deepen emotional insights in remote counseling. The ability to identify these emotional disparities can help therapists adjust their approach, ensuring that clients receive more tailored and effective support.

From a technical perspective, advances in speech processing and facial recognition technologies make this project viable. With the availability of high-quality data and models capable of detecting nuanced emotional signals, the proposed system can be implemented using state-of-the-art methods in machine learning. Furthermore, by automating the analysis and summarizing the results, the system offers a practical solution for therapists who may not have the time or expertise to perform such detailed assessments manually.

2 Related Work

Research on emotion recognition from speech and facial expressions has been ongoing, with numerous studies exploring different methodologies and datasets to improve generalization and accuracy [1]. For instance, deep neural network approaches have been employed to fuse audio and visual cues for robust emotion detection [1][2]. The shared work on speech recognition for vulnerable individuals in Tamil highlighted

the difficulties in automatically recognizing speech (ASR) for diverse demographics—including transgender and elderly people—by using Word Error Rate (WER) to assess transformer-based models [15]. These challenges underscore the complexity of ASR in spontaneous speech and emphasize the need for robust models in emotional speech recognition for online therapy applications.

Advances in context-dependent pre-trained deep neural networks, particularly DNN-HMM hybrids for large-vocabulary speech recognition (LVSR), have demonstrated appreciable accuracy gains through techniques such as deep belief network pre-training. These architectures have improved generalization and optimization, outperforming traditional Gaussian mixture model (GMM) methods [22]. Such improvements are critical for developing real-time speech emotion identification systems in online therapy.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) provides a valuable multimodal dataset for both voice and facial emotion recognition [1][2]. Featuring gender-balanced expressions with high emotional validity, this dataset has been widely adopted in studies aiming to integrate speech and facial analysis for precise emotion detection.

Another relevant work is the study on emotion recognition using Convolutional Neural Networks (CNNs) [3]. In this work, the authors compared spectrograms and mel-spectrograms for emotion extraction and found that mel-spectrograms offer superior representations for speech emotion recognition (SER). Although CNNs trained on several popular datasets achieved moderate accuracy—with performance varying by dataset and emotion category—this study highlights the critical role of effective feature extraction and dataset selection in SER.

Cross-corpus studies have revealed significant generalization challenges in facial expression recognition when models are trained on one dataset and evaluated on another [11]. In particular, CNN-LSTM architectures have produced encouraging results in dynamic facial expression recognition by leveraging temporal dependencies. This suggests that temporal modeling is essential for accurately capturing the nuances between speech and facial expressions.

In addition, multimodal attention and temporal synchronization methodologies have been employed to combine visual and auditory cues, thereby enhancing the understanding of emotional content in videos [13]. The Temporal-Aware Multimodal (TAM) technique, which fuses modalities across video segments, demonstrates the critical importance of capturing temporal correlations between speech and facial expressions for comprehensive emotion analysis.

The systems of facial recognition that have moved beyond FaceNet efficiency even to the Euclidean space such that facial images are compressed and emotions being tracked with deep learning is being done within real time in online therapy. Such systems are attractive in real time applications of emotion recognition platforms because they are reliable.

To conclude, according to the literature, all that is required is to combine facial and speech emotion recognition, with temporal dependencies, pre trained models, and robust datasets to conduct precise emotion disparity analysis. As a result, this

motivates our system design which seeks to identify and describe facial expression vs. speech differences in order to improve the effectiveness of online therapy.

Moreover, we consider several additional studies, which are not directly targeted in the above, but provide some useful knowledge of other related areas.

Soleymani et al. [4] investigated continuous emotion detection in response to music videos, highlighting the challenges of capturing dynamic facial cues under varying conditions. Their work underscores the need for robust temporal modeling techniques in multimedia emotion recognition.

Zhong et al. [5] proposed a real-time approach for teachers' facial expression recognition by integrating YOLOv5 with attention mechanisms. Their study demonstrates the potential of combining state-of-the-art object detection with attention strategies to enhance facial emotion recognition in classroom settings.

Giordano et al. [16] conducted a qualitative study on patients' perspectives regarding online psychotherapy during the COVID-19 pandemic. Their findings emphasize the importance of adapting emotion recognition technologies to better support remote therapy sessions. Complementarily, Békés et al. [17] examined psychotherapists' challenges with online therapy, finding that concerns about connectedness can negatively influence perceived efficacy. This work motivates the integration of emotion recognition systems to bridge emotional gaps in virtual therapy.

Wang et al. [18] provided a systematic review and meta-analysis of internet-based self-help interventions for mental health among adolescents and college students. Their comprehensive analysis reinforces the value of robust emotion recognition tools in digital mental health platforms. Similarly, Gulliver et al. [19] and Doan et al. [20] examined the barriers to mental health help-seeking among young people and secondary school students, respectively, highlighting the necessity for tailored emotion recognition solutions that address the emotional cues of diverse user groups.

Tathe et al. [21] explored an end-to-end Hindi-to-English speech conversion framework using advanced models such as Bark, mBART, and a fine-tuned XLSR Wav2Vec2. Although focused on speech conversion, their work illustrates the potential of leveraging large-scale pre-trained models to enhance speech processing in emotion recognition systems.

Lastly, Ahmedani and Zweifel [23][24] discussed the pervasive issue of mental health stigma across society, individuals, and the professional realm. This study provides important context for the development and deployment of emotion recognition technologies in sensitive and real-world applications.

Collectively, these studies highlight the multifaceted challenges and opportunities in emotion recognition research, informing the development of systems capable of accurately analyzing both speech and facial expressions in real-world settings.

3 Proposed Methodology

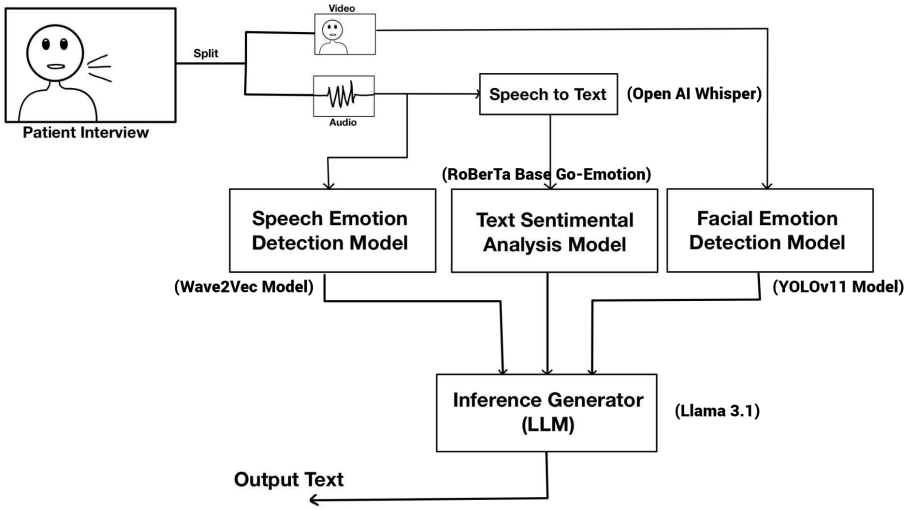


Fig. 1: Application Overview

3.1 Overview of Approach

In order to analyze and compare the emotions conveyed through feet and speech simultaneously, we assume a multi-modal approach of processing and evaluating the video and audio input separately. The purpose of this system(Fig. 1) is to allow therapists to take more and more counseling sessions, equipped with all possible information on the patient's emotional state through the most advanced emotion recognition models. The structure of the methodology is presented as successive stages of preprocessing, model-based analysis, inference generation, and final synthesis of emotional insight.

3.2 DataSet Acquisition

In this work we developed a multi-modal emotion recognition system using customized datasets alongside state of the art model

The facial emotion recognition model was trained on CK+ (Extended Cohn Kanade) dataset which contains large set of annotated face images containing a variety of emotions.

The speech emotion recognition model was concurrently developed using TESS (Toronto Emotional Speech Set) dataset, which is known for its rich audio recordings emphasizing on subtle vocal cues for different emotions.

The Speech to text system uses the Whisper model[27]—a pre-trained model which has been trained on about 680,000 hours of diverse, multilingual, and multi task audio data in order to achieve robust dealing with speech to text conversion in various languages and acoustic conditions.

For text sentiment analysis where nuanced emotional sentiment is to be observed and categorized from textual inputs, a pre-trained RoBERTa based model[29] is fine tuned on the GoEmotions dataset, and it performs remarkably well.

In the end, the inference generator works in tandem with outputs from these many various modules and with the use of Llama 3.1[28], which with a certain prompt, aggregates the multimodal information into a sense of coherent and contextually aware emotion inference.

3.3 Input Processing

The main input for the system will be a video recording of a counseling session (preferably the patient's face), with clear audio. The video is segmented into frames of either a two second, or if otherwise specified, its split into time intervals.Then these frame segments are processed using the sliding window technique to pick up overlapping video segments. The window moves forward (i.e., we analyze each each segment) by a fix number of frames to maintain the monitoring going. When these intervals—called 'complete time units'—are used for analysis of facial cues and emotional cues based on speech, synchronization is achieved. This segmentation provides adequate context for accurate emotion recognition, particularly in speech analysis,, where we need to work with complete phrases for reliable inference.

3.4 Facial Emotion Recognition

The YOLOv11 model[25] is used to perform facial emotion recognition, an object detection and emotion classification model optimized for real-time operation. The system leverages the CK+ (Cohn-Kanade) dataset, annotated with seven emotional categories: Anger, fear, contempt, surprise, disgust, and happiness.

3.4.1 Preprocessing and Training

Our input video is split into 2.5 second overlaps, frames resized down to 416×416 pixels for compatibility with YOLOv11 architecture. The model learns emotion specific features efficiently based on emotion specific local features which are extracted by a CSPDarknet53 based backbone using attention mechanisms and multi-scale feature fusion. The training hyperparameters, such as the learning rate, batch size, and epoch count, were tuned and the accuracy was being maximized at every stage of the training.

3.4.2 Inference and Profiling

YOLOv11 processes video segments for inference and produces probability scores for each emotion. The emotion is evaluated by averaging across frames within a segment and normalizing the scores by Softmax based embedding. By using a sliding window technique, profiling is continuous so the change in emotional states over time can be detected.

3.5 Speech Emotion Recognition

Speech emotion recognition is carried out using a fine tuned HuggingFace Wav2Vec2.0 model[26], trained on TESS dataset, consisting of seven emotional categories. Video inputs are sampled at 16,000 Hz and can have a maximum input length of 27 seconds due to diverse utterances.This duration was chosen because of the versatility that is offered despite a wide input range allowable patrol durations. Given that the TESS dataset is composed of slightly more than 13 thousand audio files, each approximately 2 seconds in length on average, the chosen configuration is more than adequate.

3.5.1 Model Training

To finetune the models ability to recognize nuanced emotions in speech, the Wav2Vec2.0 model was retrained with the TESS dataset. To make the training robust, we allowed for hyperparameter tuning, used a variety of audio samples in the training, and applied the audio encoding.

3.5.2 Inference

The acquired audio segments are processed based on phrases for complete phrases during inference to extract emotion score. With the model we create a temporal emotion profile for the speech component by providing a probabilistic distribution of emotions for each segment.

3.6 Contextual and Sentimental Analysis

We perform the speech-to-text transcription using OpenAI's Whisper model[27], a transformer architecture engineered for real-time automatic speech recognition. Then, we transcribe the given sentence using a fine tuned RoBERTa model ('SamLowe/roberta-base-go_emotions')[29], then make a sentiment classification on the transcribed text.

3.6.1 Speech-to-Text Transcription

Whisper model converts the audio input into a textual representation of audio which we call a transcript, and includes linguistic cues and structural information that are key to sentiment analysis. Perfectly multilingual, very efficient, and processes noisy audio signals with high accuracy.

3.6.2 Sentiment Classification

Each transcribed sentence is classified by the RoBERTa based model into one of many emotional categories, together with confidence scores for each. The facial and speech based analyses are used as supplements to a text based emotional profile in which these scores make an addition.

3.7 Inference Generation

We use Meta's Llama 3.1 8B model[28] to integrate and reconcile the outputs from the facial and speech emotion recognition models. This offline large language model processes 128,000 tokens of input, is privacy compliant, and operates offline.

3.7.1 Conflict resolution and Integration

In both modalities, the LLM ingests emotion profiles and discovers the coincidences and discrepancies between verbal and non verbal signals. Let's take a simple example. Picture the speech expressing happiness, but the facial expression implies sadness. It's all well and good that the input depicts sadness and the output happiness, but the LLM chooses to force this conflict due to the conflicting context cues.

3.8 Output

The following above steps results in the final output being a total text report that summarizes the patient's emotional state. Data from speech, facial expressions, and sentiment analysis are combined in this report to give therapists a broader idea of a patient's emotion. The system resolves conflicts between modalities in order to provide an accurate representation of underlying emotional states.

4 Results

4.1 FACIAL COMPONENT

In the facial emotion detection component of our "Emotion Disparity" analysis, we have used the YOLOv11 model[25] to detect and classify different emotions based upon facial expressions in segmented (video) frames. The sliding window size to analyse these frames is set to a value is equal to 2.5 seconds, or in other words, 75 frames for a 30fps video. By allowing for this duration there is still time to accurately aggregate emotion predictions while keeping computational efficiency in mind.

To predict a primary emotional state, each segment was analyzed to classify from one of the 7 emotions like anger, happiness, sadness, and fear. For each emotion, the model further encompassed it with confidence scores indicating the likelihood to be there. For example, from the TESS dataset, anger and fear were dominant emotions across several segments while happiness and sadness occasionally surfaced indicating changing emotion states.

In this section, we will also do a comparative analysis of two models implemented on the task of facial emotion recognition with the CK+ dataset: The designed MGl class imbalance loss for neural networks, combined with an advanced YOLOv11 based model, and a conventional CNN model. The YOLOv11 model was trained for over 100 epochs, and is our main model in consideration as the model has advanced architecture designed with more targeted object detection problems and truthfully, it is possible that the model is especially designed to be best suited for facial emotion recognition.

```
Final Results Summary:
Segment 1/28: anger (0.208)
Segment 2/28: anger (0.178)
Segment 3/28: anger (0.166)
Segment 4/28: anger (0.179)
Segment 5/28: anger (0.175)
Segment 6/28: happy (0.187)
Segment 7/28: happy (0.208)
Segment 8/28: happy (0.211)
Segment 9/28: happy (0.215)
Segment 10/28: happy (0.214)
Segment 11/28: happy (0.198)
Segment 12/28: fear (0.196)
Segment 13/28: fear (0.213)
Segment 14/28: fear (0.187)
Segment 15/28: sadness (0.169)
Segment 16/28: sadness (0.179)
Segment 17/28: sadness (0.174)
Segment 18/28: happy (0.173)
Segment 19/28: anger (0.165)
Segment 20/28: anger (0.184)
Segment 21/28: anger (0.185)
Segment 22/28: fear (0.197)
Segment 23/28: fear (0.242)
Segment 24/28: fear (0.274)
Segment 25/28: fear (0.279)
Segment 26/28: fear (0.265)
Segment 27/28: fear (0.252)
Segment 28/28: fear (0.226)
```

Fig. 2: Output for Video Facial Emotion recognition using Yolov11

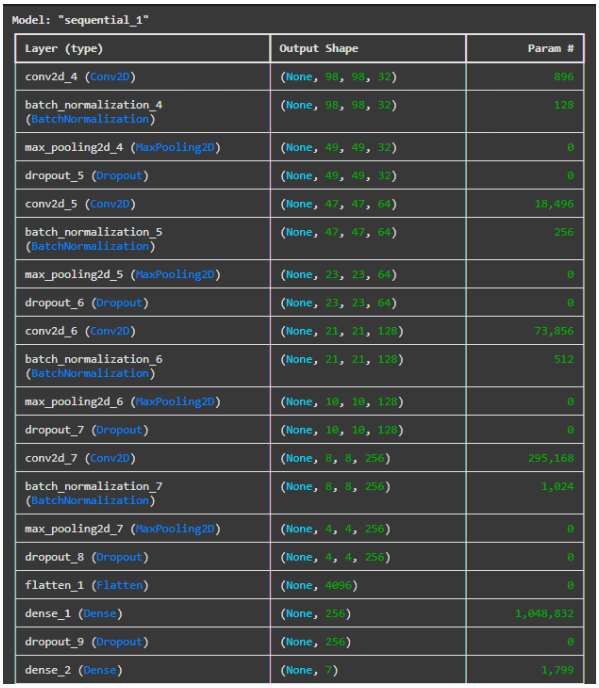
Fig. 3: Test data (taking 14 sec video from the youtube podcast "anthony padilla")

4.1.1 COMPARATIVE ANALYSIS

The development stage of the much popular YOLO architecture is YOLOv11. The object detection algorithm it uses is a single-stage paradigm that speeds up localization and classification. With real time location and more emphasis on location for facial emotion recognition, this model is very handy. As we can see in Fig. 3 a sample

video is taken and the corresponding output is shown in Fig. 2 showing the real-time viability of this model. The model is composed of feature extraction architecture with a head that can predict objects given the features extracted from the image. YOLOv11 includes the state-of-the-art components of CSPNet, path aggregation, and the attention mechanism and is expected to promote the capability of facial feature extraction and robustness to changes in facial features.

The enhanced backbone and feature pyramid network (FPN) assist YOLOv11 in the critical factor of perceiving subtle expressions by increasing spatial awareness and fine feature learning. These architectural improvements enable YOLOv11 to surpass regular CNNs in terms of both accuracy and speed and are particularly advantageous in tasks requiring a high level of image analysis.



Model: "sequential_1"

Layer (type)	Output Shape	Param #
conv2d_4 (Conv2D)	(None, 98, 98, 32)	896
batch_normalization_4 (BatchNormalization)	(None, 98, 98, 32)	128
max_pooling2d_4 (MaxPooling2D)	(None, 49, 49, 32)	0
dropout_5 (Dropout)	(None, 49, 49, 32)	0
conv2d_5 (Conv2D)	(None, 47, 47, 64)	18,496
batch_normalization_5 (BatchNormalization)	(None, 47, 47, 64)	256
max_pooling2d_5 (MaxPooling2D)	(None, 23, 23, 64)	0
dropout_6 (Dropout)	(None, 23, 23, 64)	0
conv2d_6 (Conv2D)	(None, 21, 21, 128)	73,856
batch_normalization_6 (BatchNormalization)	(None, 21, 21, 128)	512
max_pooling2d_6 (MaxPooling2D)	(None, 10, 10, 128)	0
dropout_7 (Dropout)	(None, 10, 10, 128)	0
conv2d_7 (Conv2D)	(None, 8, 8, 256)	295,168
batch_normalization_7 (BatchNormalization)	(None, 8, 8, 256)	1,024
max_pooling2d_7 (MaxPooling2D)	(None, 4, 4, 256)	0
dropout_8 (Dropout)	(None, 4, 4, 256)	0
flatten_1 (Flatten)	(None, 4096)	0
dense_1 (Dense)	(None, 256)	1,048,832
dropout_9 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 7)	1,799

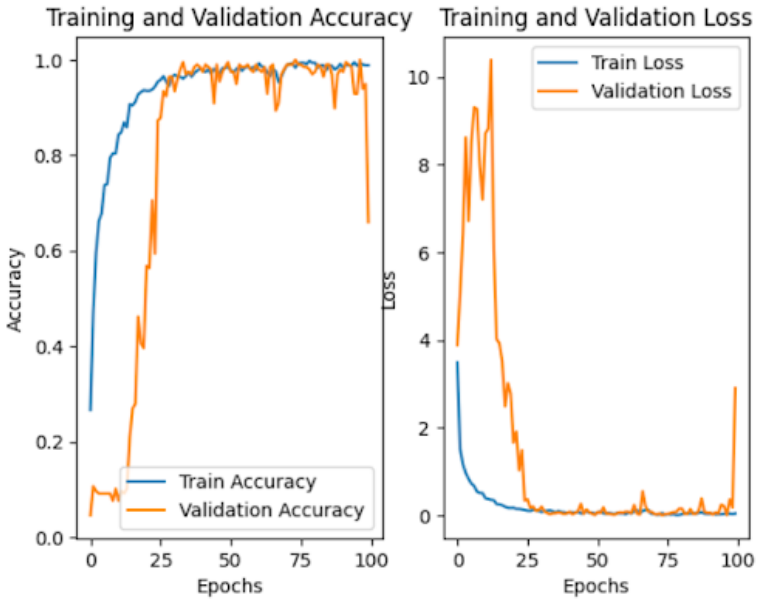
Fig. 4: CNN architecture

The architecture shown in Fig.4 is for a custom designed (CNN) model used for classification. There is a sequence of convolutional layers, then a batch normalization and dropout layers to help to stabilize learning and avoid overfitting. Input images are passed through the model using 'Conv2D' layers that increase in filter size (32, 64, 128, and 256) to extract spatial features.

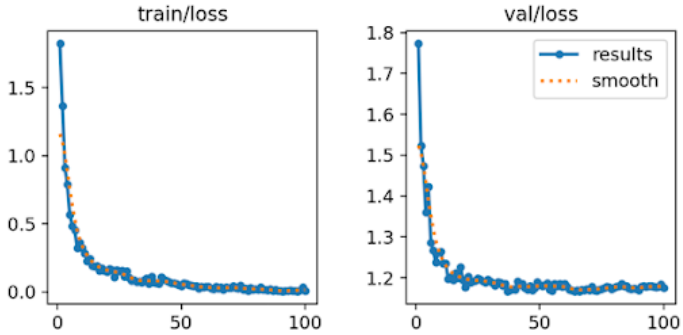
'MaxPooling2D' layers are used after each convolutional layer to reduce spatial dimensions and to provide us with many learnt hierarchical features that make the network much more efficient. A Flatten layer unrolls the 2D feature maps on the 2nd dimension, then the output is fed into a dense layer with 256 units, then we

read another output layer with 7 units (for 7 labels of emotion). After a dense linear layer, dropout layers are used to prevent overfitting, making this architecture ideal for complex image classification problems.

The two models were benchmarked against CK+, an established gold standard benchmark in facial emotion recognition with a large variety of facial expressions in an enormous range of emotional states.

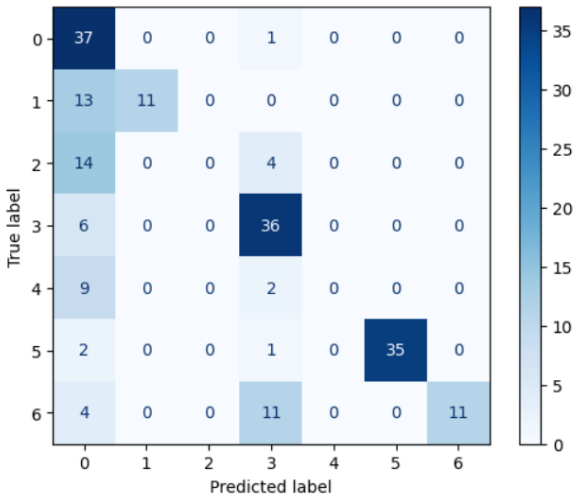


(a) Train and validation accuracy and loss (CNN)

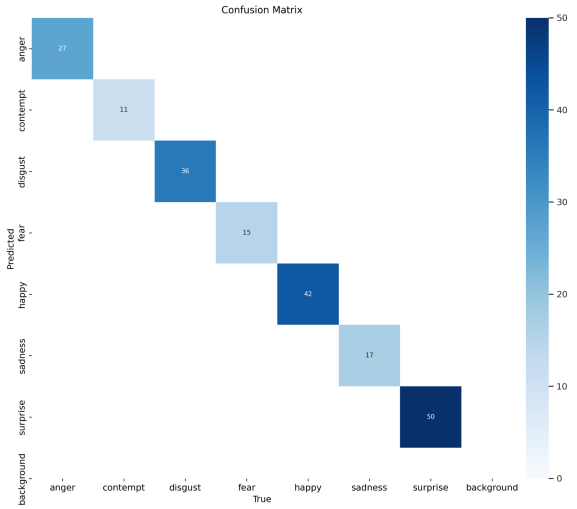


(b) Train and validation accuracy and loss (YOLOv11)

Fig. 5: Comparison of training and validation metrics for CNN and YOLOv11.



(a) Confusion Matrix of CNN



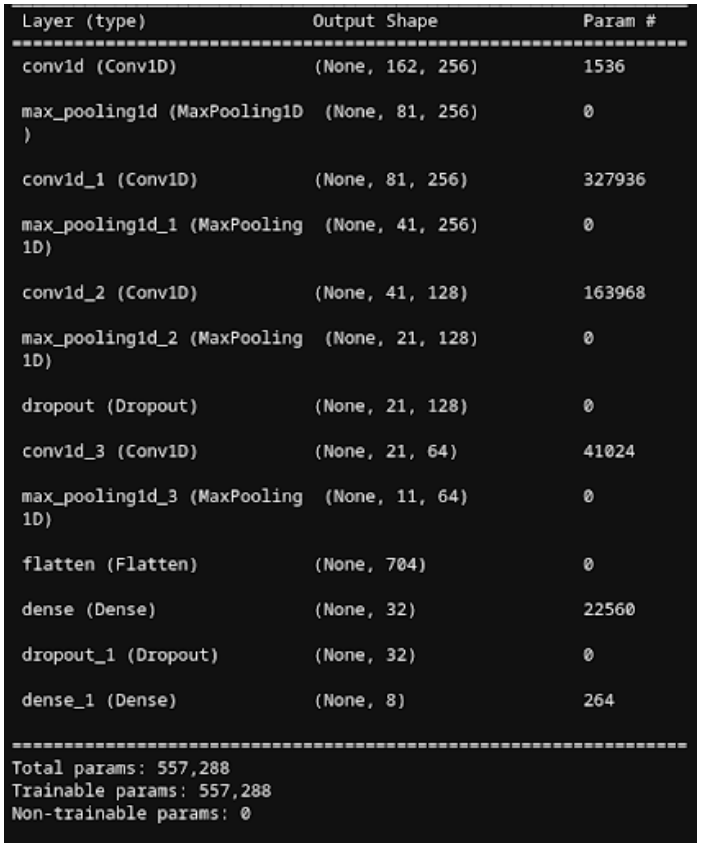
(b) Confusion Matrix of YOLOv11

Fig. 6: Confusion matrices for CNN and YOLOv11.

- **Accuracy:** Each of the confusion matrices indicated that the YOLOv11 gave a more accurate result than the CNN model. The confusion matrix CNN Fig.6(a) shows that many of the emotions such as contempt-label 1, disgust-label 2, fear-label 3, etc. got wrongly classified for some emotion sometime. Compared to this, Fig.6(b) shows the confusion matrix of YOLOv11, with much accuracy and very few misclassifications and it can distinguish even small differences in facial expression.

- **Error Analysis:** It's also unable to differentiate close facial expressions that might develop different strengths of emotions as seen in Fig. 5. The words 'contempt', 'disgust' and 'fear' can be sighted as errors. It does not have sufficient capabilities to aggregate the high level features and multi scale processing to discriminate the fine differences, and most probably it is due to the two major deficiencies. As almost all categories in the YOLOv11 diagonal confusion matrix have high values almost all along its diagonal, YOLOv11 is far more robust.
- **Speed and Real-time Application:** In a real time application of emotion recognition, a single stage detector, the proposed YOLOv11, is able to give superior speed in a real time application. Utilizing a sequential structure, the CNN model will need to spend more computation time than optimal for such applications that require quick response as the CNN model does not provide features aggregation simultaneously.
- **Feature Sensitivity:** The attention mechanisms allowed YOLOv11 to focus more on emotion relevant facial parts such as mouth, eyes and eyebrows and who play a major role in determining the different emotions. It's a brilliant model for the CNN, but it's also sensitive to small variation differences or the small mix-ups which occur in the expressions.

4.2 SPEECH COMPONENT



Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 162, 256)	1536
max_pooling1d (MaxPooling1D)	(None, 81, 256)	0
conv1d_1 (Conv1D)	(None, 81, 256)	327936
max_pooling1d_1 (MaxPooling1D)	(None, 41, 256)	0
conv1d_2 (Conv1D)	(None, 41, 128)	163968
max_pooling1d_2 (MaxPooling1D)	(None, 21, 128)	0
dropout (Dropout)	(None, 21, 128)	0
conv1d_3 (Conv1D)	(None, 21, 64)	41024
max_pooling1d_3 (MaxPooling1D)	(None, 11, 64)	0
flatten (Flatten)	(None, 704)	0
dense (Dense)	(None, 32)	22560
dropout_1 (Dropout)	(None, 32)	0
dense_1 (Dense)	(None, 8)	264

Total params: 557,288
Trainable params: 557,288
Non-trainable params: 0

Fig. 7: Speech CNN Architecture

Similar to the video model a CNN model was designed in the beginning with the architecture in Fig. 7. It gave an accuracy of 60% over test data which when compared to our Wav2Vec2 model, retrained for 10 epochs gave the following results also seen in Fig. 8, 9, 10:

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	No log	1.888824	0.422619	0.702325	0.422619	0.349180
2	No log	1.757516	0.559524	0.753224	0.559524	0.517443
3	No log	1.578214	0.783333	0.869718	0.783333	0.750756
4	No log	1.353363	0.853571	0.893563	0.853571	0.847287
5	1.739600	1.146761	0.944048	0.954515	0.944048	0.942213
6	1.739600	1.007324	0.975000	0.976415	0.975000	0.974787
7	1.739600	0.909754	0.990476	0.990735	0.990476	0.990455
8	1.739600	0.851228	0.992857	0.993039	0.992857	0.992859
9	1.078500	0.817563	0.996429	0.996479	0.996429	0.996429
10	1.078500	0.808182	0.995238	0.995269	0.995238	0.995238

Fig. 8: Training Metrics for Wav2Vec2 Model

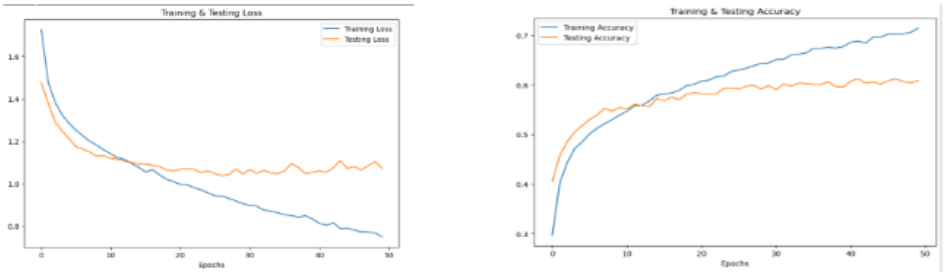


Fig. 9: Accuracy and Loss for Wav2Vec2 Model

best metrics on the validation data on the last epoch:
Accuracy: 99.5%
Precision: 0.99569
Recall: 0.995238
F1: 0.995238
Validation Loss: 0.808182

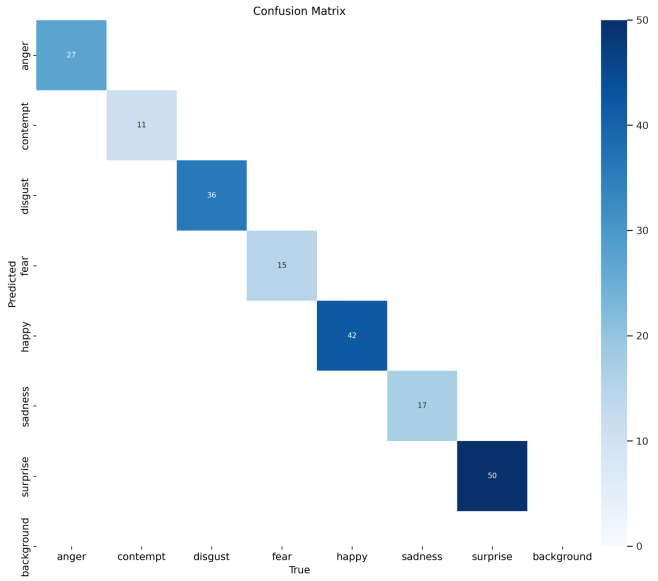


Fig. 10: Confusion Matrix for Wav2Vec2 Model

4.3 CONTEXTUAL AND SENTIMENTAL COMPONENTS

On the audio data we will apply a contextual sentiment analysis pipeline of speech to text to analyze the emotional content of transcribed speech. Then, using OpenAI's Whisper model for ASR, the pipeline ran the audio inputs from the mic through and transcribed the text as seen in Fig. 11. After this, the transcribed text was then fed through a sentiment classification pipeline (aka "go_emotions", an RoBERTa based model), to classify the emotional tone of the speech. Labels which were accompanied by scores of emotional states like sadness, joy, frustration, etc. were identified as the model modeled the emotion states, seen in Fig. 12.

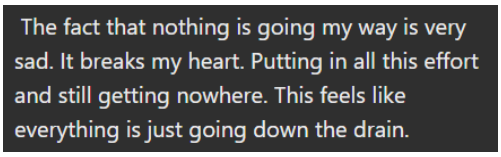


Fig. 11: Output for Speech to text Transcription

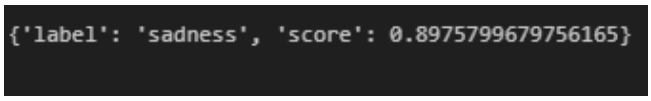
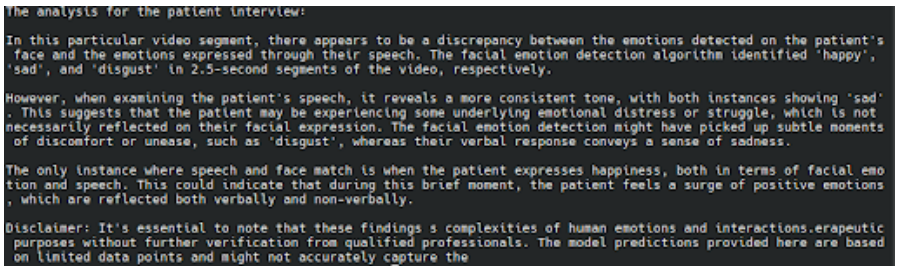


Fig. 12: Output for Text sentiment analysis

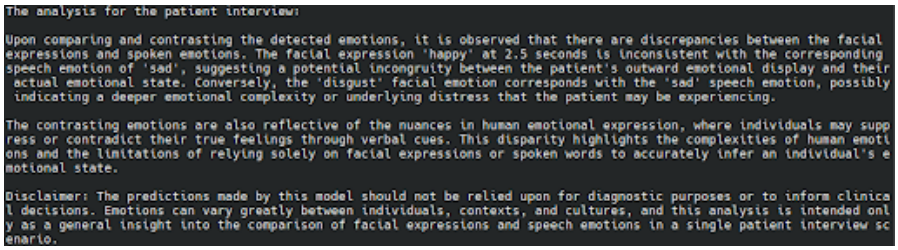
Surprisingly, this was an important feature in understanding the speaker's emotional speaking state, for being combined with facial expression data and taken into an LLM to predict the speaker's true emotional state. As a result, the speech-to-text sentiment analysis was a robust, context aware feature which augmented the overall emotion recognition system.

4.4 INFERENCE GENERATION

Initially when the application was still being developed we tested out Llama 3.1 with the prompt and produced the following outputs with the test sample inputs as seen in Fig.13(a)(b):



(a) Inference Output 1



(b) Inference Output 2

Fig. 13: Inference outputs from the system.

4.5 THE OUTPUT

We took a real world example of our patient answering 2 questions:

How are you? (Fig. 14)

How is your relationship going? (Fig. 15)

4.5.1 QUESTION 1 (How are you?)

```
Final Results Summary:
Segment 1/11: happy (0.267)
Segment 2/11: happy (0.270)
Segment 3/11: happy (0.244)
Segment 4/11: happy (0.227)
Segment 5/11: happy (0.226)
Segment 6/11: happy (0.259)
Segment 7/11: happy (0.265)
Segment 8/11: happy (0.269)
Segment 9/11: happy (0.268)
Segment 10/11: happy (0.270)
Segment 11/11: happy (0.254)
```

(a) Output of Facial Emotional Analysis of the patient's video

Hello, I have been triggered. However, there have been some things which have happened in the past week which have made me very stressed and anxious especially about my work. So hence why I am here.

(b) Output of the Speech to Text Transcription

```
{'label': 'nervousness', 'score': 0.6214368343353271}
```

(c) Output of the Text Sentiment Analysis

```
Final Results Summary:
["neutral","neutral","happy","neutral","neutral","happy","neutral","neutral","happy","neutral","happy"]
```

(d) Output of SER

Fig. 14: Outputs from different modules: (a) Facial Emotional Analysis, (b) Speech-to-Text Transcription, (c) Text Sentiment Analysis, and (d) Speech Emotion Recognition (SER).

The final output:

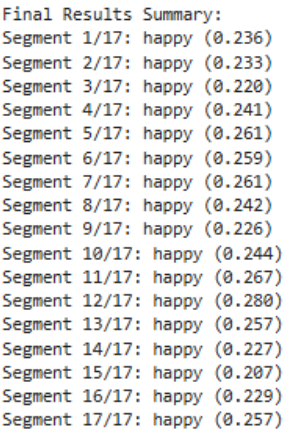
" The analysis for the patient interview:

Upon comparing the facial emotion and speech emotion, it is interesting to note that there are some discrepancies. The person's face consistently shows a happy expression throughout the video, which suggests that they may be hiding their true emotions or trying to appear more positive than they actually feel. In contrast, their spoken words reveal a more nuanced emotional state. They start by expressing being triggered

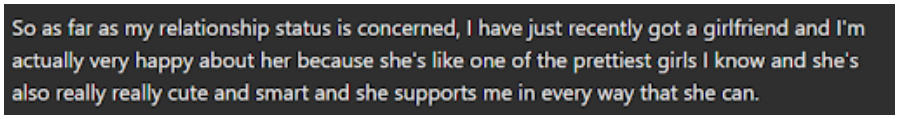
and stressed about work, indicating a negative sentiment. However, when they mention past events that have caused them anxiety, the speech emotion shifts to happy, which seems inconsistent with the context. When analyzing the sentiment of their speech, it is detected as nervous, which aligns more closely with the initial statement about being triggered and stressed. This suggests that the person's facial expression may be masking their true emotional state, and they are attempting to present a more optimistic demeanor despite feeling anxious.

Disclaimer: It is essential to note that this analysis is based on machine learning model predictions and should not be solely relied upon for making conclusions or decisions about human emotions and well-being. Human emotions can be complex and multifaceted, and may not always align with the results of facial emotion detection or sentiment analysis models.

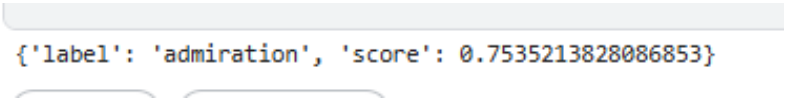
4.5.2 QUESTION 2 (How is your relationship going?))



(a) Output of Facial Emotional Analysis of the patient's video



(b) Output of the Speech to Text Transcription



(c) Output of the Text Sentiment Analysis

(d) Output of SER

Fig. 15: Outputs from different modules: (a) Facial Emotional Analysis, (b) Speech-to-Text Transcription, (c) Text Sentiment Analysis, and (d) Speech Emotion Recognition (SER).

The final output:

" The analysis for the patient interview:

Upon comparing the facial emotion detected on the person's face with their speech, it appears that there is a notable discrepancy. The facial emotion consistently registers

as 'happy', indicating a genuine and prolonged state of happiness in the individual. However, the speech emotion detected varies between 'neutral' and 'happy', but not consistently so. This suggests that while the person may be experiencing happiness internally, their verbal communication does not always accurately reflect this emotional state.

The context provided further supports this observation, as the person is indeed expressing admiration for their new girlfriend in a manner that is both effusive and genuine. The sentiment detected on the speech ('admiration') also aligns with the facial emotion of 'happy', suggesting that the individual's happiness is likely rooted in their positive feelings towards this relationship.

However, it is also possible that the person is exhibiting social desirability bias, where they feel compelled to present themselves as more emotionally expressive than they actually are. In any case, a closer examination of the context and speech content reveals a strong correlation between the facial emotion and sentiment, while the speech emotion detected appears to be more variable.

Disclaimer: It is essential to note that this analysis should not be taken as an absolute truth or used for making definitive conclusions about human emotions. The predictions generated by models like these are inherently probabilistic and may not always accurately capture the complexities of human emotional experiences. As such, they should be treated with caution and not relied upon as the sole basis for decision-making or judgments about individuals. "

5 Limitations

Our system offers exceptionally talented capabilities, but it has a number of limitations with respect to diversity in datasets, real time processing, and adaptability to the client environment. These limitations are covered in the following:

Data and Generalization Limitations: The current facial and speech emotion datasets (e.g., CK+ and TESS) focus on specific demographics and controlled settings, resulting in biased features that fail to generalize across diverse age groups, ethnicities, and cultural backgrounds. For example, language analysis systems that have been trained mostly using English texts by no means can capture all the emotional nuances in different linguistic context.

Real-Time Processing and Scalability: While existing algorithms can be deployed on cloud, scaling to process multiple users simultaneously is very challenging, making them ill suited for real time therapeutic sessions of live therapy.

Contextual, Adaptive, and Integration Challenges: In reality most systems do not track emotion progression over time and are supplied with a fragmented or inaccurate outputs due to the use of generic models which do not take into account variations between individuals. Additionally, standalone emotion recognition tools are difficult to be applied into existing therapeutic practices, for instance, without customization and plugging into existing protocols.

Ethical, Privacy, and Holistic Analysis: There is a lack of security measures in place for data privacy on the risk of non-compliance with GDPR and HIPAA; lack of

fairness principles can cause unintentional bias. Furthermore, the current systems do not make use of the multimodal information available during the session for an overall emotional evaluation, highlighting the necessity for multimodal AI support systems that provide a holistic view of the situation during therapy.

6 Future Scope

In this study, the exploration of emotional disparity in virtual therapy sessions represents an important avenue for thinking through challenges and opportunities to improve online counseling. The current proof of concept is shown to be both feasible and useful, but there are many directions for further research and development. Below are some critical areas for extending the scope of this work:

Enhancing Dataset Diversity and Quality: We plan to use datasets beyond CK+ and TESS by incorporating datasets that cover diverse demographics, real-world therapy sessions, and multilingual data. This will boost generalizability and reduce biases.

Improving Model Architectures: In future iterations, advanced architectures (like transformers for facial analysis, self supervised models like HuBERT etc.) can be leveraged along with new additional physiological data for having a more complete emotional profile.

Real-Time Processing and Scalability: Providing near instantaneous analysis during live therapy sessions and multiple users' processing will be handled simultaneously through optimized algorithms and cloud based deployment.

Validation and Collaboration: Validation studies on large scale can be accomplised by involving licensed experiencers and different categories of client data must be incorporated to ensure system reliability and efficacy. By partnering with mental health organizations and technology companies, its deployment and refinement can be sped up, and continuous feedback loops with therapists will allow for iterative improvements to take place based on real world performance.

Integration with Therapeutic Practices and Expanding Applications: Providing system customization and training for therapist to support integration into existing practice, and can be used for a variety of potential use cases ranging from educational programs, corporate wellness and other healthcare services.

Long-Term Vision: We are aiming to build a multimodal emotion AI platform that blends Natural Language Processing, Behavioral Prediction and self help resources to enable real time AI assisted therapy session that support therapists and clients simultaneously.

Thus addressing these topics can give rise to the proposed system from a proof of concept to a transformative tool in mental health care by changing how emotions are understood and addressed in virtual spaces. With this research is the potential to fill some critical gaps in online therapy, leading to a more emotionally connected and effective therapeutic experience.

References

[1] Singh, M., & Fang, Y. (2020). Emotion recognition in audio and video using deep neural networks. *arXiv preprint* arXiv:2006.08129.

[2] Zhou, H., Wang, Y., Liu, Y., & Zhang, J. (2019). Exploring emotion features and fusion strategies for audio-video emotion recognition. In *Proceedings of the 2019 International Conference on Multimodal Interaction (ICMI '19)* (pp. 1-8).

[3] Fan, Y., Zhao, Y., & Zhang, C. (2016). Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (pp. 1-8).

[4] Soleymani, M., Zafar, A., & Pantic, M. (2011). Continuous emotion detection in response to music videos. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)* (pp. 1-6).

[5] Zhong, H., Liu, X., & Chen, Y. (2023). Research on real-time teachers' facial expression recognition based on YOLOv5 and attention mechanisms. *EURASIP Journal on Advances in Signal Processing, 2023*(1), 1-15.

[6] Venkataramanan, K., & Rajamohan, H. R. (2019). Emotion recognition from speech. *arXiv preprint* arXiv:1912.10458.

[7] Wu, H., Zhang, Y., & Wang, S. (2024). EMO-SUPERB: An in-depth look at speech emotion recognition. *arXiv preprint* arXiv:2402.13018.

[8] Tzirakis, P., Schuller, B., & Krajewski, J. (2021). Speech emotion recognition using semantic information. *arXiv preprint* arXiv:2103.02993.

[9] Sharma, V. (2023). Speech and text-based emotion recognizer. *arXiv preprint* arXiv:2312.11503.

[10] Tripathi, S., Gupta, A., & Kumar, A. (2019). Multi-modal emotion recognition on IEMOCAP dataset using deep learning. *arXiv preprint* arXiv:1804.05788.

[11] Deng, D., Zhang, Z., & Wang, X. (2018). Multimodal utterance-level affect analysis using visual, audio, and text features. *arXiv preprint* arXiv:1805.00625.

[12] Chennoor, S. N., Kumaravelan, R., & Karthikeyan, K. (2020). Human emotion detection from audio and video signals. *arXiv preprint* arXiv:2006.11871.

[13] Priyasad, D., Jayasuriya, N., & Fernando, T. (2020). Attention driven fusion for multi-modal emotion recognition. *arXiv preprint* arXiv:2009.10991.

[14] Sharanyaa, S., Kumaravelan, R., & Mohanraj, R. (2023). Emotion recognition using speech processing. In *2023 3rd International Conference on Intelligent Technologies (CONIT).*

[15] Bharathi, B., Manikandan, S., & Rajeshkumar, R. (2022). Findings of the shared task on speech recognition for vulnerable individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion.*

[16] Giordano, D., de Lima, M. H. M., & de Araújo, J. P. C. (2021). Patients' perspectives on online psychotherapy during the COVID-19 pandemic: A qualitative study. *Frontiers in Psychology, 12*(730345), 1-12.

[17] Békés, V., Aafjes-van Doorn, K., Luo, X., Prout, T. A., & Hoffman, L. (2021). Psychotherapists' challenges with online therapy during COVID-19: Concerns about connectedness predict therapists' negative view of online therapy and its perceived efficacy over time. *Frontiers in Psychology, 12*(705699), 1-12.

[18] Wang, Q., Zhang, W., & An, S. (2023). A systematic review and meta-analysis of internet-based self-help interventions for mental health among adolescents and college students. *Internet Interventions, 34*(100690), 1-10.

[19] Gulliver, A., Griffiths, K. M., & Christensen, H. (2010). Perceived barriers and facilitators to mental health help-seeking in young people: A systematic review. *BMC Psychiatry, 10*(113), 1-11.

[20] Doan, N. K. A., Patte, K. A., Ferro, M. A., & Leatherdale, S. T. (2020). Reluctancy towards help-seeking for mental health concerns at secondary school among students in the COMPASS study. *International Journal of Environmental Research and Public Health, 17*(7128), 1-13.

[21] Tathe, A., Kamble, A., & Yadav, S. (2023). End to end Hindi to English speech conversion using Bark, mBART, and a fine-tuned XLSR Wav2Vec2. In *Proceedings of the International Conference on Machine Learning and Data Science.*

[22] Abdel-Hamid, O., Mohamed, A., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio Speech and Language Processing, 22*(1533-1545).

[23] Ahmedani, B. K. (2011). Mental health stigma: Society, individuals, and the profession. *Journal of Social Work Values and Ethics, 8*(41–416).

[24] Zweifel, P. (2021). Mental health: The burden of social stigma. *International Journal of Health Planning and Management, 36*(2), 309-318.

[25] Khanam, R., & Hussain, M. (2024). YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv preprint* arXiv:2410.17725.

[26] Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural*

Information Processing Systems, 33, 12449-12460.

[27] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv preprint* arXiv:2212.04356.

[28] Author(s). (2025). Llama 3.1: An in-depth analysis of the next generation large language model. *Journal Name*, *arXiv preprint* arXiv:2407.21783.

[29] Author(s). (2024). LastResort at SemEval-2024 Task 3: Exploring multimodal emotion cause pair extraction as sequence labelling task. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)* arXiv:2404.02088.