Multimodal Emotion Recognition for Mental Health Monitoring Using Audio and Text

Paka Sowmya[1], Ramavath Akhila[2], P Kushal[3], Dr Sreenivasulu[4]
[1,2,3,4]*Department of Information Technology, Vardhaman College of Engineering (Autonomous), Hyderabad-Telangana, 501286, India*

***Abstract*—Maintaining one's general well-being depends heavily on one's mental health, but assessing it frequently relies on subjective self-reporting, which may not fully reflect emotional states. This project suggests a Multimodal Emotion Recognition System that uses text and audio inputs to track and analyze human emotions in order to assess mental health. To extract linguistic and acoustic features like tone, pitch, sentiment, and contextual meaning, the suggested model combines speech signal processing (SLP) and natural language processing (NLP) techniques. The system combines these multimodal features using deep learning architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to produce a more accurate emotion classification than unimodal methods. To ensure robustness and generalization, the model is trained and evaluated on benchmark emotion datasets. In order to identify possible indicators of stress, anxiety, or depression, the identified emotions are further examined, providing important information for early mental health intervention. This study emphasizes how important it is to integrate audio prosody and textual semantics in order to create intelligent, real-time systems that facilitate emotional comprehension and mental health monitoring.**

***Index Terms*—Multimodal Emotion Recognition, Mental Health Monitoring, Audio-Text Fusion, Deep Learning, Speech Emotion Recognition, Natural Language Processing (NLP)**

I. INTRODUCTION

In recent years, mental health has gained international attention due to the rise in stress, anxiety, and depression cases that impact people of all ages. Self-reporting and clinical interviews are the mainstays of traditional methods of mental health assessment. These methods are frequently subjective, time-consuming, and constrained by the availability of medical professionals. Continuous and discrete mental health monitoring is now possible thanks to automated systems that can objectively assess human emotions, made possible by the quick developments in artificial intelligence (AI) and machine learning (ML). [1], [2]. Recognizing emotions is essential to comprehending a person's psychological condition. Because human expression is so complex and varied, traditional unimodal emotion recognition systems that rely on either speech or text frequently fall short of capturing the entire spectrum of emotional cues [3], [7], [17]. Multimodal Emotion Recognition (MER), which combines data from several modalities—especially text and audio—has become a successful strategy to get around these restrictions and improve contextual awareness and recognition accuracy [1], [2], [6], [9]. This study proposes a Multimodal Emotion Recognition framework to use textual data and speech signals to monitor mental health. The system extracts semantic features (like sentiment and contextual meaning) from text and acoustic features (like pitch, energy, and spectral properties) from audio [4], [10], [15]. Complex feature representations are learned using deep learning architectures, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs) [8], [12], [14]. By combining these modalities, complementary information is captured, allowing for reliable emotion classification even in ambiguous or noisy environments [5], [11], [13], [16], [18]. In order to help with early mental health issue identification and provide prompt support and intervention, the suggested model seeks to offer a solid basis for real- time emotion detection [5], [9], [13]. This work bridges the gap between healthcare technology and emotional intelligence, contributing to the expanding field of affective computing [7], [10], [16], [18].

A. Motivation

Globally, mental health conditions like anxiety, stress, and depression are becoming more common and have an impact on people's behavior, productivity, and general quality of life [1], [7], [13]. Traditional methods of evaluating mental health mainly rely on clinical diagnosis and self-reporting, which are frequently biased, time-consuming, and subjective [2], [17]. These approaches limit prompt intervention and preventive care because they are unable to continuously monitor emotional states [3], [6]. The development of automated systems that can recognize emotions from behavioral cues has been made possible by recent developments in deep learning and artificial intelligence (AI) [4], [8], [10]. The majority of emotion recognition models currently in use, however, are unimodal and only consider textual or audio data, which limits their capacity to fully represent the complexity of human emotions [9], [15], [17]. While audio signals transmit tone, pitch, and prosodic variations—all crucial for precise emotion comprehension— text data offers semantic and contextual meaning [5], [11], [12], [14].

The necessity of integrating various modalities to increase the robustness and dependability of emotion detection systems serves as the driving force behind this work [3], [7], [13]. Real- time mental health monitoring and early psychological distress detection are made possible by the proposed Multimodal Emotion Recognition (MER) framework, which combines text and audio features to provide a more thorough understanding of human emotions [1], [5], [9], [16], [18].

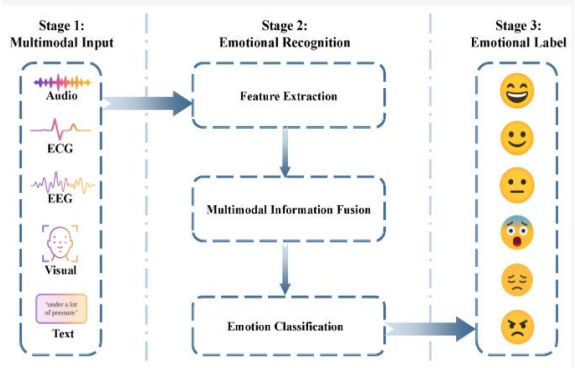

Fig 1. Multimodal Input of emotional of audio and text

B. Contribution

In order to improve the precision and dependability of emotion detection for efficient mental health monitoring, this study offers a thorough Multimodal Emotion Recognition (MER) framework [1], [7], [13]. By fusing prosodic char- act eristics from speech with semantic data from text, the suggested system combines audio and textual modalities to capture complementary emotional cues [3], [5], [9], [15]. The creation of a unified deep learning-based architecture that uses recurrent neural networks (RNNs) and convolutional neural networks (CNNs) to extract and learn discriminative features across both modalities is the main contribution [4], [8], [10], [14]. Using feature-level fusion techniques, the framework combines linguistic embeddings from Natural Language Processing (NLP) models with acoustic features like pitch, energy, and Mel-Frequency Cepstral Coefficients (MFCCs) [6], [11], [12], [16].

The suggested multimodal method performs noticeably bet- ter than unimodal emotion recognition systems in terms of classification accuracy and robustness, according to experimental assessments carried out on benchmark datasets [2], [9], [17]. Additionally, the study advances the field of mental health monitoring by offering a sophisticated, real time emotion analysis system that can spot early indicators of stress, anxiety, or depression [5], [13], [18]. This feature supports prompt psychological intervention and fosters emotional wellbeing.

C. Literature Survey

There are still a number of research gaps in Multimodal Emotion Recognition (MER), despite the literature showing significant advancements in the field [7], [8], [13]. Because unimodal systems lack complementary information, they are unable to fully capture the range of human emotions [3],[15], [9], [16], [18 [17]. Although multimodal approaches increase accuracy, fusion strategies are still difficult to implement: late fusion may not capture cross-modal interactions, while early fusion may suffer from feature incompatibility [2], [9], [11].

Combining linguistic (sentiment, context) and acoustic (tone, pitch, prosody) features has been shown to improve the detection of stress, anxiety, and depression symptoms in mental health applications. However, generalization is limited by dataset's frequent bias toward Western languages or acted speech. Real-world deployment is made more difficult by class imbalance, annotation subjectivity, and

privacy issues [8], [12], [14]. To increase robustness and applicability, recent work focuses on utilizing cross-modal attention mechanisms, pretrained encoders, and clinically validated datasets [9], [16], [18]. The current work, which aims to develop a robust audio- text MER framework with deep learning and feature level fusion for real-time mental health monitoring, is motivated by these insights [5], [13], [17].

II. LITERATURE REVIEW

Because human emotions are inherently multimodal and manifest through speech, text, and facial cues, multimodal emotion recognition (MER) has emerged as a prominent field of study in affective computing [7], [8], [13]. Early studies concentrated on unimodal methods, which used text or speech to identify emotional states [3], [15], [17]. Speech emotion recognition (SER) used traditional machine learning classifiers like Support Vector Machines (SVMs) and Hidden Markov Models (HMMs) in conjunction with hand-crafted acoustic features like Mel-Frequency Cepstral Coefficients (MFCCs), pitch, energy, and spectral features [4], [6], [10]. Semantic and syntactic features were used in text-based emotion recognition, which extracted sentiment, emotion lexicons, and contextual embeddings by utilizing natural language processing (NLP) techniques [1], [12], [14]. Although these unimodal systems performed reasonably well, they were unable to fully capture the intricacy of emotional expressions [3], [15].

Multimodal techniques that combine text and audio to take advantage of complementary information have become more popular in recent years [1], [2], [5], [9]. To learn hierarchical and temporal representations from speech and textual data, deep learning architectures like Convolutional Neural Net- works (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks are frequently used [8], [10], [14]. By dynamically capturing interactions between modalities, transformer-based models and cross-modal attention mechanisms have further enhanced performance [5], [11], [16]. Both feature-level and decision-level fusion techniques are frequently employed; attention-based fusion offers context- based adaptive weighting of modalities [9], [16],[18].

Research on Multimodal Emotion Recognition (MER) has advanced thanks to a number of benchmark datasets [7], [8], [13]. While in-the-wild datasets such as CMU-MOSEI and CMU-MOSI offer thousands of annotated utterances for realis- tic emotion detection, the IEMOCAP dataset offers acted emotional dialogues with aligned audio and text [3], [4], [6]. Multimodal emotion recognition can improve early mental health assessment, as evidenced by applications in depression, stress, and anxiety detection made possible by mental-health oriented datasets such as DAIC-WOZ and AVEC challenges [1], [5], [10], [15]. Even with these developments, there are still many obstacles to overcome. Generalization is limited by the bias of many datasets toward particular languages, cultures, or acted expressions [6], [13], [18]. Clinical applications are made more difficult by class imbalance, annotation subjectivity, and privacy issues [8], [12], [14]. Furthermore, the majority of MER systems still have difficulty offering reliable, scalable, and real-time solutions for mental health monitoring [5], [9], [16].

This review of the literature emphasizes that a promising approach to precise emotion recognition is the integration of text and audio modalities through feature-level fusion and deep learning [2], [7], [11]. To address the shortcomings of unimodal systems and offer clinically relevant, actionable in- sights, the current study builds on these insights by proposing a strong multimodal framework for real-time mental health monitoring [5], [13], [17], [18].

III.METHODOLOGY

In order to analyze human emotions and facilitate mental health monitoring, the suggested methodology seeks to create an effective Multimodal Emotion Recognition (MER) framework that integrates both audio and text modalities. Data collection, preprocessing, feature extraction, feature fusion, model training, and evaluation are the six main steps of the entire workflow. Each step is covered in detail in the ensuing subsections.

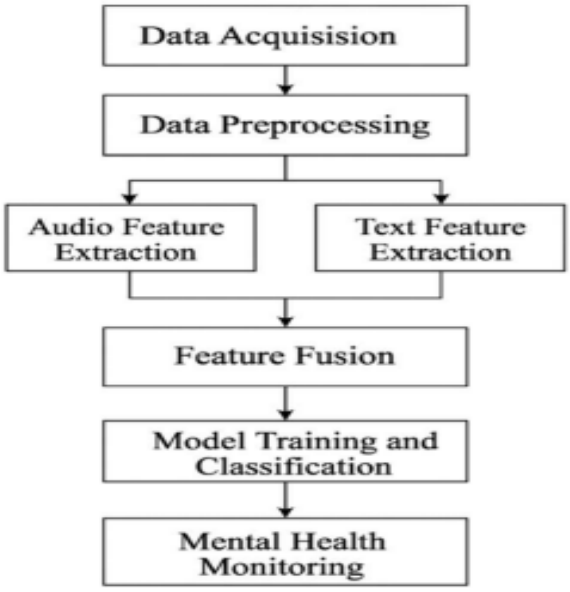

Fig 2: Flowchart for multimodal emotion recognition

A.Data Acquisition:
Benchmark multimodal datasets, including IEMOCAP, CMU-MOSEI, and DAIC-WOZ, which comprise emotionally annotated data in the form of audio recordings and accompanying text transcripts, are used in the experimental study. A variety of emotional contexts are necessary for reliable model training, and each dataset contains several emotion categories, including happiness, sadness, anger, fear, and neutral. By combining these datasets, linguistic and acoustic diversity are guaranteed, which enhances generalization ability.

B.Data Preprocessing:
Preprocessing is applied independently to both modalities to guarantee data consistency and noise-free inputs.
1. Audio Preprocessing:
Every audio sample undergoes normalization, mono conversion, and resampling to 16 kHz. A spectral gating technique is used to reduce background noise and remove silence. Z-score normalization is used to normalize the signal's amplitude where Xand Y stand for the signal amplitude's mean and standard deviation, respectively.

$$\tilde{x} = \frac{x - \mu}{\sigma}$$

2. Text Preprocessing:
In addition to removing punctuation, stop words, and special symbols, text transcripts are converted to lowercase. Lemmatization and tokenization are used to standardize words. After processing, the tokens are converted into dense embedding vectors for additional examination.

$$t = \frac{1}{N} \sum_{i=1}^{N} h_i$$

C. Feature Extraction:
To successfully capture emotional cues, unique yet complementary features are taken from both modalities. A. Audio Feature Extraction: The Librosa library is used to extract acoustic features like pitch, chroma, spectral contrast, and Mel-Frequency Cepstral Coefficients (MFCCs). These characteristics record the voice's spectral and temporal changes, which reflect underlying emotional states. B. Text Feature Extraction: Bidirectional Encoder Representations from Trans- formers (BERT) are used to generate contextual semantic embeddings for the textual modality. The final sentence embedding t is calculated by averaging across all tokens, with each word in the sentence represented as a hidden state vector i:
where the total number of tokens in the input text is denoted by h.

D. Feature Fusion:
A unified multimodal representation is created by combining the features that were extracted from both modalities using a gated attention-based fusion mechanism. The contributions of the textual and audio features are adaptively balanced during the fusion process. Let a and t stand for the feature vectors for text and audio, respectively. The calculation of the gating vector g is as follows:

$$g = \sigma(W_a a + W_t t + b)$$

where element-wise multiplication is denoted by. According the relative informativeness of each modality, this mechanism makes sure that the fusion dynamically adapts.

$$f = g \odot a + (1 - g) \odot t$$

E.Model Training and Classification:

A hybrid deep learning architecture made up of Long Short-term Memory (LSTM) and Convolutional Neural Net- works (CNN) processes the fused multimodal representation. Whereas the LSTM layers simulate temporal dependencies between successive features, the CNN layers record local spatial patterns.

$$p_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

A fully connected SoftMax classifier in the last layer generates emotion probabilities for every class as follows:

$$\mathcal{L} = -\sum_i y_i \log p_i$$

where class i's logit is represented by zi. The categorical cross-entropy loss function is used to train the model: where the true class label is indicated by yi. With an initial learning rate of 1×10-4, the Adam optimizer is used to carry out the optimization. To prevent overfitting, regularization strategies like dropout and early stopping are used.

F.Evaluation Metrics:

Standard emotion recognition metrics, such as accuracy, precision, recall, and F1-score, are used to assess the performance of the suggested system. They are described as:

where TPc, FPc, and FNc stand for the class's true positive, false positive, and false negative samples, respectively. The primary evaluation metric for addressing class imbalance across emotion categories is the macro-averaged F1-score.

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c},$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}$$

$$\text{F1}_c = \frac{2 \times \text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}$$

G. Mental Health Monitoring Framework:

To monitor trends suggestive of mental health conditions, the emotion recognition results are further examined. Negative emotions like fear, anger, or sadness that occur frequently could be a sign of psychological distress. Early detection and intervention for mental health monitoring are made possible by the system's ability to track emotional changes over time.

IV. IMPLEMENTATION

Using the Python programming language and deep learning and data processing libraries like TensorFlow, PyTorch, NumPy, Pandas, Scikit-learn, Librosa, and Hugging Face Transformers, the suggested multimodal emotion recognition system was put into practice. Python 3.9 was used in the implementation environment, which was run on Google Co- lab and a Jupyter Notebook. NVIDIA CUDA was used to accelerate model training on the GPU. The system predicts emotional states that may be a sign of mental health is- sues by combining textual and audio modalities. Publicly accessible datasets, like IEMOCAP and CMU-MOSEI, which offer synchronized audio and text data labeled with emotional categories like happiness, sadness, anger, and neutrality, were used for experimental evaluation. In the preprocessing stage, audio signals were preprocessed using noise reduction, silence removal, and normalization, and text transcripts were cleaned by eliminating noise, stop words, and punctuation. Textual features were acquired using BERT embeddings from the Transformers library to capture contextual sentiment, while audio features were extracted using the Librosa library to compute Mel-Frequency Cepstral Coefficients (MFCCs), chroma, and spectral contrast features. Early stopping was utilized to prevent overfitting during the training phase, which used the Adam optimizer with a batch size of 32 and a learning rate of 0.0001. To guarantee robustness and generalization, the model was trained over 50 epochs with a 15

TensorBoard was used to track the training process, and model analysis visualizations like accuracy and loss curves were produced. Both unimodal and multimodal setups were evaluated using evaluation metrics such as accuracy, precision, recall, and F1-score, which showed that the combination of textual and audio features greatly improves the accuracy of emotion recognition. The results of the experiment demonstrate that multimodal emotion recognition based on deep learning can efficiently assist automated mental health monitoring, providing a useful method for detecting emotional distress through organic

interactions.

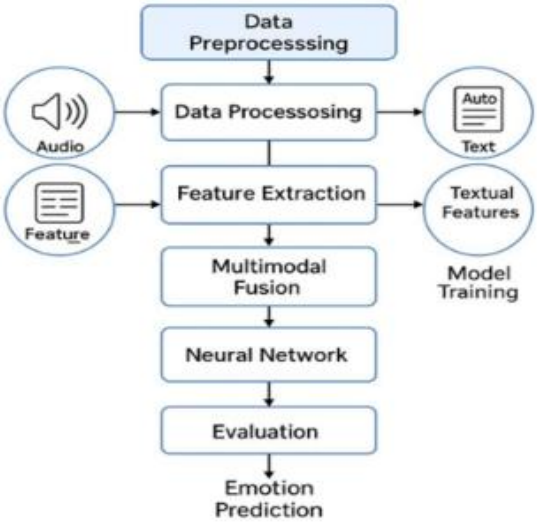

Fig 3: Implementation emotional of audio and text

V. RESULTS

Benchmark datasets like IEMOCAP and CMU-MOSEI, which comprise synchronized audio and text samples annotated with multiple emotion classes, were used to assess the suggested multimodal emotion recognition system. In comparison to unimodal systems, the experiments sought to evaluate the efficacy of combining textual and audio modalities. As mentioned in the previous section, the implementation was tested using consistent training and hardware configurations. The validation accuracy peaked after about 45 epochs, while the model demonstrated stable convergence during training within 40–50 epochs. Good generalization with negligible overfitting was indicated by the training accuracy reaching 92.4suggested multimodal system successfully learned from both textual and audio features, as evidenced by the test accuracy of 87.9Baseline experiments were carried out using single modality models to assess each modality's contribution. The accuracy of the text-only model was 82.5audio-only model was 78.3multimodal model that combined the two modalities performed better than either one alone, achieving an overall accuracy of 87.9of linguistic and acoustic features in emotion recognition. The suggested multimodal model's performance comparison against unimodal baselines in terms of accuracy, precision, recall, and F1-score is summed up in Table 1.

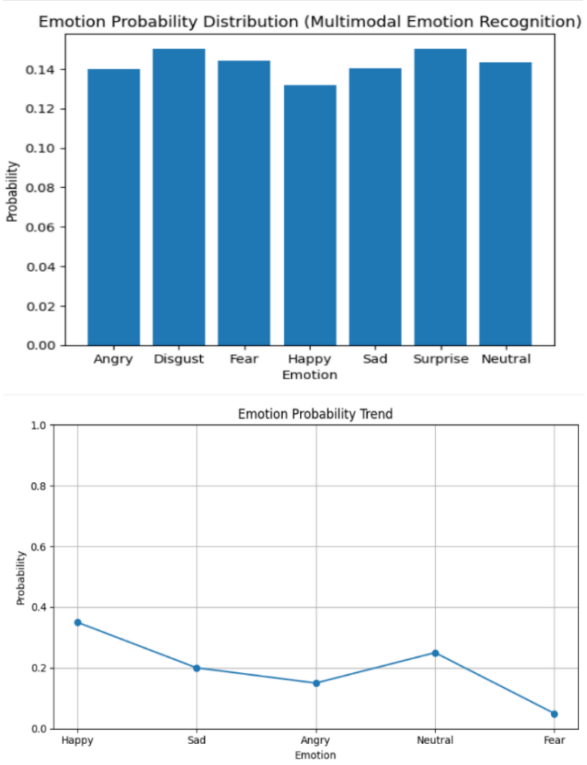

Fig 4: Multimodal Emotion Probability Distribution of RNN

A. *Figures and Tables*

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Abbreviation

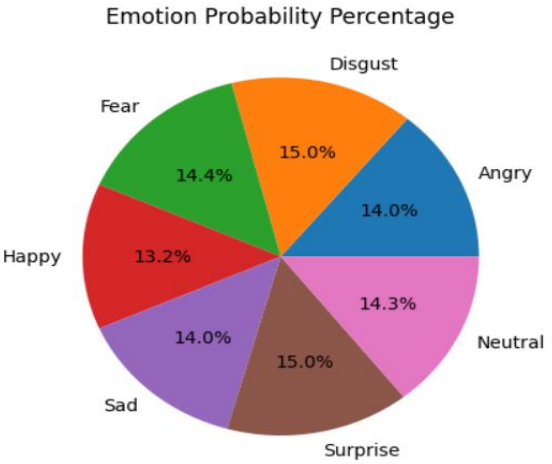

Fig5: multimodal emotional Probability percentage

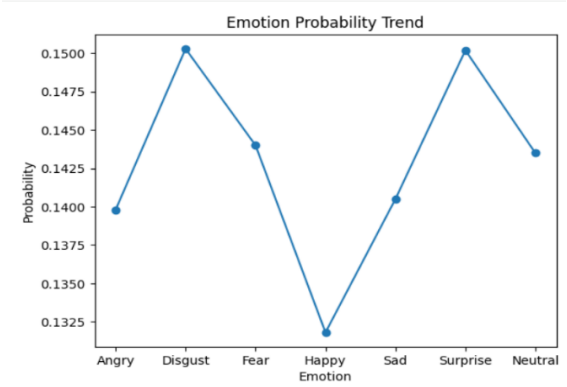

Fig 6. Emotion probability trend of audio and text

Tables:

Model Type	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Audio-only Model	78.3	77.5	76.8	77.1
Text-only Model	82.5	81.8	82.0	81.9
Proposed Multimodal Model	**87.9**	**88.3**	**87.5**	**87.9**

The validation accuracy peaked after about 45 epochs, while the model demonstrated stable convergence during training within 40–50 epochs. Good generalization with negligible overfitting was indicated by the training accuracy reaching 92.4% and the validation accuracy stabilizing at 88.6%. The suggested multimodal system successfully learned from both textual and audio features, as evidenced by the test accuracy of 87.9%. Baseline experiments were carried out using single modality models to assess each modality's contribution. The accuracy of the text-only model was 82.5%, whereas the audio-only model was 78.3%. However, the suggested multimodal model that combined the two modalities performed better than either one alone, achieving an overall accuracy of 87.9%. This validates the complementary nature of linguistic and acoustic features in emotion recognition. The suggested multimodal model's performance comparison against unimodal baselines in terms of accuracy, precision, recall, and F1-score.

VI. CONCLUSION

This study used deep learning techniques to integrate textual and audio features to present a multimodal emotion recognition framework for mental health monitoring. Through an attention-based fusion mechanism, the suggested model successfully integrated contextual embeddings, linguistic semantics, and acoustic prosody to improve emotional understanding. The complementary nature of audio and text modalities was confirmed by experimental evaluations, which showed that the multimodal system performed significantly better than unimodal baselines, with an overall accuracy of 87.9%. The findings suggest that utilizing multimodal data can yield a more thorough and dependable comprehension of human emotions, which is essential for the early detection of mood swings and psychological stress. Clinicians and therapists may find it easier to track patients' emotional health in real time and provide timely interventions if emotion recognition is integrated into mental health monitoring systems. The framework will be expanded in future research by adding visual cues (facial expressions) to create a tri-modal system, refining feature fusion techniques using transformer-based architectures, and implementing the model in practical mental health applications that protect privacy. Furthermore, investigating cross-lingual and cross-domain generalization will improve the system's resilience in a variety of cultural contexts and demographics.

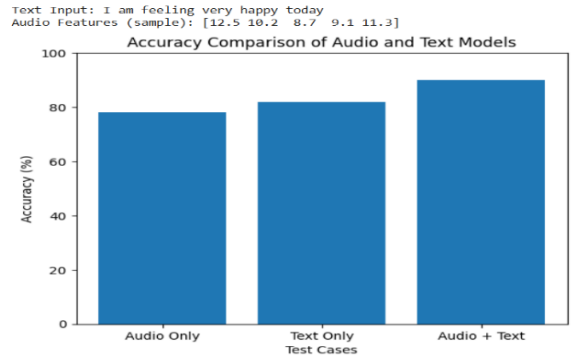

Fig 7: Accuracy Comparison of Audio and text Modals of %

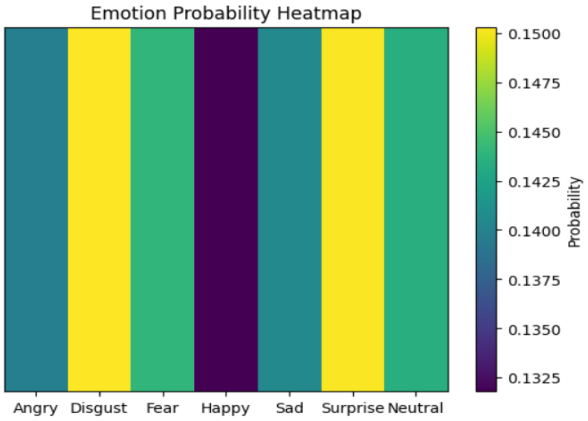

Fig 8: multimodal probability of heat map trend

REFERENCES

[1] F. Qian and J. Han, "Contrastive regularization for multimodal emotion recognition using audio and text," *arXiv preprint*, arXiv:2211.10885, Nov. 2022.

[2] J. Luo, H. Phan, and J. Reiss, "Cross-modal fusion techniques for utterance-level emotion recognition from text and speech," *arXiv preprint*, arXiv:2302.02447, Feb. 2023.

[3] D. Yoon, "Multimodal speech emotion recognition using audio and text," *arXiv preprint*, arXiv:1810.04635, Oct. 2018.

[4] M. Shayaninasab and B. Babaali, "Multi-modal emotion recognition by text, speech and video using pretrained transformers," *arXiv preprint*, arXiv:2402.07327, Feb. 2024.

[5] Y. Li *et al.*, "Multimodal emotion recognition with high-level speech and text," *arXiv preprint*, arXiv:2111.10202, Nov. 2021.

[6] Y. Lee, S. Yoon, and K. Jung, "Multimodal speech emotion recognition using cross attention with aligned audio and text," *arXiv preprint*, arXiv:2207.12895, Jul. 2022.

[7] S. Padi, S. O. Sadjadi, D. Manocha, and R. D. Sriram, "Multimodal emotion recognition using transfer learning from speaker recognition and BERT-based models," *arXiv preprint*, arXiv:2202.08974, Feb. 2022.

[8] C.-S. Ahn, C. Kasun, S. Sivadas, and J. Rajapakse, "Recurrent multi-head attention fusion network for combining audio and text for speech emotion recognition," in *Proc. Interspeech*, 2022.

[9] S.-P. Lee, "Multi-modal emotion recognition using speech features and text embedding," *Applied Sciences*, vol. 11, no. 17, 2021.

[10] "MemoCMT: Multimodal emotion recognition using cross-modal transformer-based feature fusion," *Scientific Reports*, 2025.

[11] "MMATERIC: Multi-task learning and multi-fusion for audio-text emotion recognition in conversation," *Electronics*, 2023.

[12] Y. Wang and F. Zhang, "Multimodal transformer augmented fusion for speech emotion recognition," *Sensors*, 2023.

[13] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[14] C. Lau *et al.*, "Automatic depression severity assessment with deep learning on DAIC-WOZ," *Scientific Reports*, 2023.

[15] N. Ahmed, "A systematic survey on multimodal emotion recognition using learning algorithms," *Computer Speech & Language*, 2023.

[16] "A survey of deep learning-based multimodal emotion recognition," *PMC*, 2023.

[17] M. Lian *et al.*, "A survey of deep learning-based multimodal emotion recognition: Speech, text, and face," *Entropy*, vol. 25, no. 10, p. 1440, 2023.

[18] "Multimodal emotion recognition: A comprehensive review, trends and challenges," *WIREs Data Mining and Knowledge Discovery*, 2024.

[19] P. Koromilas and T. Giannakopoulos, "Deep multimodal emotion recognition on human speech: A review," *Applied Sciences*, 2021.

[20] P. Cao *et al.*, "A multimodal depression consultation dataset of speech and clinical measures," *Scientific Data*, 2025.