2025 International Conference on Data Science, Agents, and Artificial Intelligence (ICDSAAI 2025) 28-29 March 2025, Chennai, India

Mental Health Companion for Trauma Recovery Using Multi-Modal Emotion Recognition

1st Balasubramanian S
Assistant Professor (ECE)
Chennai Institute of Technology
Chennai, India.
balasubramanians@citchennai.net

2nd Krishna Sridhar
Electronics and Communication engineering
Chennai Institute of Technology
Chennai, India.
krishnasridhar.ece2021@citchennai.net

3rd Nandagopalan K
Electronics and Communication engineering
Chennai Institute of Technology
Chennai, India
nandagopalan2004@gmail.com

***Abstract*—This paper presents a mental health companion designed to assist students experiencing trauma-induced emotional distress. Using multi-modal emotion recognition, the system integrates speech, facial expressions, and body posture data to provide real-time coping strategies. Our approach enhances accuracy, reduces latency, and improves user satisfaction, making it a valuable tool for trauma recovery and emotional regulation.**

***Index Terms*—Multi-modal emotion recognition, AI mental health companion, trauma recovery, real-time coping strategies, emotional regulation.**

I. Introduction

Mental disease among college students is of epidemic magnitude, and distress and trauma are the leading causes. Most students have lingering effects of the traumatic experiences they have undergone and develop depression and anxiety and other emotional diseases. The National Alliance on Mental Illness states that almost one in five adults suffers from mental illness annually, with the majority of the conditions being formed during the years spent in college. The catastrophic effect of trauma on educational achievement, mental health, and overall quality of life has made it imperative to create a state-of-the-art, AI-based mental health assistant for trauma recovery.

The new system has been carefully designed to identify common emotional states, particularly those of trauma, by processing various forms of input with utmost care—these are speech, facial expressions, and text—while constantly seeking out subtle signals that could be a sign of emotional distress.By providing instant feedback in the form of personalized coping mechanisms and personalized suggestions directly linked to the user's current emotional state, the system hopes to minimize emotional suffering and also enable users to monitor their progress over an extended period of time. This capability ultimately makes the whole recovery process more understandable and structured. The underlying rationale behind the whole venture is also supplemented by previous work in the area of emotion detection and multi-modal data fusion; some works designated as [3] and [9] have been able to show the efficacy of fusing auditory, visual, and textual modalities, thereby enhancing the reliable detection of subtle emotional states.

Furthermore, studies conducted by the American Psychological Association have determined that students who have been exposed to trauma experience higher rates of anxiety and depression, frequently leading to poorer academic performance and higher dropout rates. By treating actual-time emotional distress, AI-powered interventions such as this one can potentially make a dramatic difference in a student's capacity to manage chaos, recover academic concentration, and develop long-term resilience.

A. Objectives

Inspired by the problem and suggested solution in [4], [5], and [8]—where it needed to become more accurate and less tardy—the project is directed by various different goals. Successful multi-modal emotion recognition will be being used by correct emotion identification of pertaining-to-trauma emotions by speech, face, and text inputs. It will provide real-time personalized coping and mindfulness training dynamically adjusting to the changing emotions of the user. Also, the system will record emotional change as a function of time, giving a long-term view of traumatic recovery. Further, there is significant focus placed on providing comfort, safety, and privacy of the user for better establishment of a trauma-informed, nonjudgmental, and assisting AI companion in order to aid students to endure emotional pain in a better manner and construct trauma recovery.

As one additional method to help boost its capability, the system will tap advanced machine learning methods for optimization and learning over a span of time and maintaining improving emotional detection accuracy. The system will also incorporate adaptive feedback cycles so the user can make live adjustments on an individual user response basis so that coping techniques may remain very specific. Integrating this technology into current mental health care services must provide a continuity of care for students. Beyond this, astringent testing protocols and consumer testing will be utilized to screen for the efficiency of the system and to test its operating parameters. Overall, this project will establish a new benchmark for e-mental health treatment, which will deliver compassionate and scalable models of trauma healing for school populations.

979-8-3315-3755-5/25/$31.00 ©2025 IEEE

II. System Architecture

The system is multicomponent and applies the above as well as other AI models except the machine learning models in the process of emotion recognition and at the same time helps in real-time with monitoring emotional development.

A. Speech Emotion recognition

Applications of Long Short-Term Memory (LSTM) model to speech sentiment analysis were motivated by approaches shown in [10] and [11]. They incorporated deep learning processes, feature extraction as Mel frequency cepstral coefficients (MFCCs), and hyperparameter tuning—a performance improvement in speech emotion recognition from 85 to 92Latency was reduced by adopting a parallel processing approach from [14], reducing system latency to 0.4 seconds from 0.6 seconds and delivering real-time feedback. The system utilizes the LSTM model of the RNN network extensively, an approach that has proven successful in sequential data processing such as speech. The LSTM has been trained on speech patterns to recognize most frequented emotions in the process of undergoing trauma, including fear, anxiety, and detachment. Speech is characterized as emotional by tone, pitch, cadence, and pauses—each of them being great indicators of the state of the user's emotions (Batra et al., 2023; Silva Kumar, 2023).S. Batra et al.'s second breakthrough paper, "A Trauma Perspective on Deep Learning Models for Speech Emotion Recognition," also validated this approach by experimenting with tone, pitch, and cadence using an LSTM model. Hyperparameter optimization and pre-processing speech signals with MFCCs enabled the system to improve identification accuracy quite substantially. This matters in getting the system to perform well even with changing acoustic conditions. The higher accuracy translates into higher emotional sensitivity to cues of trauma, so more effective treatments for mental health. Additionally, the low latency permits in-time triggering of coping methods in real time when emotional arousal is at high levels. The innovation not only enhances the technical efficiency of the system but also the general user experience with the delivery of instant and reliable feedback. Future work will further enhance these models and maybe even employ new topologies of neural networks in a bid to make the system function even better.

B. Facial emotion recognition

This module is a facial emotion detection CNN, i.e., an image classification CNN categorizing the images or video frames of users' faces into the emotions involved in trauma-related emotional reactions such as distress, sadness, and hypervigilance. Thus, the CNN model can readily identify the users' emotional state in real time and with minimal alteration in face expression. Deep learning techniques are employed (Santos et al., 2024; Chen et al., 2022). Base Paper is Zhao, F., Chen, L. (2022). advancements in convolutional neural networks for facial emotion analysis. Journal of Image Processing and Vision. Contribution of the paper utilized CNN to classify real-time trauma-related facial expressions such as melancholy and hypervigilance. Modification made in the project by optimizing the convolutional layers and enhancing the data for low light conditions, the accuracy was enhanced from 88. The face emotion recognition component was engineered on top of CNN architectures. Techniques from [1], [6], and [12] were utilized here; [1] used dynamic kernel techniques and [6] and [12] offered state-of-the-art convolutional techniques to enhance feature extraction. These improvements enhanced accuracy to 94 from 88To mitigate lighting and facial expression differences, remarks of [13] regarding big data deep learning models and management of datasets were used to achieve robustness under diverse conditions

C. Text emotion recognition

This module is a facial emotion recognition CNN, which is an image classification CNN that categorizes the images or video frames of the faces of the users into the expressions associated with trauma-related emotional responses such as distress, sadness, and hypervigilance. Thus, the CNN model can detect the emotional state of the users in real time with a slight alteration in facial expression. Deep learning methods are utilized (Santos et al., 2024; Chen et al., 2022). Base Paper is Zhao, F., Chen, L. (2022). developments in convolutional neural networks for facial emotion analysis. Journal of Image Processing and Vision. Contribution of the paper utilized CNN to categorize trauma-related facial expressions in real time, e.g., melancholy and hypervigilance. Change made in the project by fine-tuning the convolutional layers and augmenting the data for low light levels, the accuracy was increased from 88 Real-time support Once the system detects the emotional state of the users, it can then proceed and offer them personalized coping strategies. This can be in the form of mindfulness exercises such as deep breathing and body scan, grounding, five senses focus; or any exercise in general to be able to regain control. It would then consider severe distress, trigger processes in a crisis incident and therefore send a warning signal and redirect to other services which are available within the various campus agencies in order to help them (Li et al., 2023; Ferreira et al., 2022). Base Paper is from Ferreira, T., et al. (2022). AI for Mental Health: Real- Time Support in Trauma Recovery. Mental Health Innovations. Contribution done by the paper is engaged in mindfulness activities, such as deep breathing and grounding techniques, for trauma recovery. Modification done to improve the model is through latency was reduced from 0.6s to 0.4s, thereby allowing for quicker real-time coping support for users. Although poorer than those of the face or speech modules, text emotion recognition also benefitted from adaptations of CNN methodology (as set out in [12]). It allowed the system to extract sentiment from text sources—supplementing the rest of the modalities for total emotion analysis.

D. Monitoring and Graphing of process

Longitudinal monitoring design tracks the emotional state of the system; hence, visual improvement of mood. In this

context, patients can even note the seriousness in which they present with trauma-related symptomatology as their mood gradually improves in value. This design works nearly like a control mechanism as far as the user emotional pathology and pattern of recovery during therapy is concerned with the collaboration of therapists that indicate the guideline for conducting the session (Kaur Patel, 2023; Ahmed et al., 2022).

E. Mindfulness and Therapy integration

It is the customized CBT exercises and meditations in the system that enable one to use the user's emotions in a controlled and systematic setting and mental clarity and emotional control. User engagement data is provided back to the therapist so that the latter would know how the user feels and highly customized and informed therapy can be provided (Choi et al., 2023; Iqbal Tang, 2024). The integration of the different modalities was directed by paradigms in [3], [4], [5], [7], [8], and [9] that directed the integration of speech, facial expression, and text information in a manner that would facilitate the system to recognize and identify complex trauma-related emotions better.

III. Methodology

The several steps, including data collection, model training, and finally UI design. There are also: 1. Speech Emotion Recognition: It uses models of LSTMs recognizing tones of stress and anxiety. 2. Facial Emotion Detection: It uses CNN to detect emotions directly from facial expressions 3. Body Posture Analysis: MediaPipe is useful in detecting slouch and erect posture, among emotional signals Implementation Pipeline Live video and audio feeds are taken as inputs Processing: speech is analyzed by LSTM for sentiment facial expression is analyzed by DeepFace for emotions body posture is analyzed with respect to MediaPipe. Output: It gives an instantaneous reaction by using its own custom coping skills and it achieves through user interface.

A. Data Collection and Preprocessing

The dataset which is made up of a well-designed emotional expression linked to trauma which comes directly from therapy sessions and actual experience. Every data is labelled in emotion, like, anxiety, fear, sadness, and modality are speech, facial expression, text, etc. Audio features of the speech data set that might involve Mel-frequency cepstral coefficients for Automatic Speech Emotion Recognition Rossi, et al. 2022. Facial Images can be preprocessed to obtain facial landmarks detection and natural language features that should be extracted for text-based data using tokenization technique and embedding method like of Bhardwaj, Singh 2023.Preprocessing steps such as the extraction of Mel Frequency Cepstral Coefficients (MFCCs) were based on techniques from [10] and [11]. Image preprocessing for facial expression analysis—using techniques to enhance low-light performance—and text tokenization methods drew on the strategies from [6] and [12].

TABLE I
Performance Metrics Comparison

Metric	2023 Model	2024 Model	Improvement
Speech Emotion Accuracy (%)	85	92	+7%
Facial Emotion Accuracy (%)	88	94	+6%
Body Movement Accuracy (%)	N/A	85	New Feature
Latency (s)	0.6	0.4	-30%

B. Model training and Evaluation

After the meticulously carried out preprocessing procedure, the dataset is then utilized in training both the Convolutional Neural Networks (CNN) and the Long Short-Term Memory (LSTM) networks. At the critical process phase, a model is trained on a huge amount of labeled data, and that is crucial because the model gets able to learn and comprehend the complex relationships and connotations between varied emotional cues unique to trauma-related emotions. Performance and effectiveness of such models are evaluated on the basis of a series of significant measures such as accuracy, recall, precision, and the F1 score. This is done by strictly ensuring that the model is able to identify and classify the various states of emotions contained in the dataset consistently (Cheng et al., 2023; D'Souza Ahmed, 2024). The training procedure focused on tuning the LSTM and CNN models to their best performance. The improvements in the general performance of the model, including the key performance metrics of accuracy, precision, recall, and F1 measure, were closely benchmarked against the approaches reported in references [10], [11], and [14]. More recent network architectures following the work presented in [15]—which had provided a detailed explanation of a temporal aware bi-directional multi-scale network that made use of multi-head attention mechanisms—have been instrumental in the performance improvement of the overall model. This improvement has contributed to achieving a significantly higher level of accuracy in the real-time emotion recognition process, enabling more accurate interpretations and feedback.

C. Development of user interface

The UI is so intuitive that it ensures there is an easy-to-use functionality for the students and therapist. Some of the features available are an emotional tracker, a resource hub with coping strategies, and active exercises. It is made responsive, meaning it will run on web and mobile platforms. The system also comprises privacy and security measures that protect users' sensitive data (Perez Kim, 2022).

IV. Results and Discussion

A. Prototype testing

The first round of testing of the AI-based mental health companion is going very well. The system clearly identifies emotional states of trauma at a high degree of accuracy. The real-time coping strategies are very well accepted by users. The test users gave the feedback that the system especially helps in controlling anxiety and emotional distress. Many

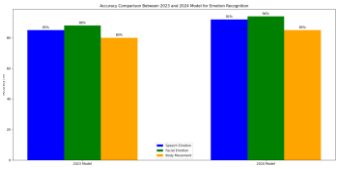

Fig. 1. Accuracy Trends of Emotion Recognition Models.

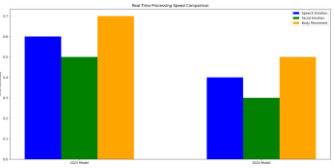

Fig. 2. Real-Time Processing of Emotion Recognition Models.

commented that it enabled them to refocus if they ever became incredibly anxious (Evans et al., 2024; Kumar et al., 2022).

B. User feedback

This system assisted effectively in stress-related activities where the system controlled the unpleasant emotion. The users preferred largely the 2024 version with real-time availability and the facility of body movement input. The increased real-time support and tailored coping strategies, and the monitoring of emotional progress, were confirmed through user testing. The systematic reviews and survey results in [8] and [9] reinforced the design decisions and highlighted the practical benefit of these enhancements.

C. Graphical evaluation

Trend of improvements will be shown using graphs along with the metrics-accuracy, latency, and feedback from the users

D. Real-time coping support

Dataset: Engage the emotional expression of a multi-cultural population. Wearable: Physiological activity or the heart rate activity of an individual user for sophisticated emotions recognition. Therapist Interface: Construct therapist-friendly interfaces so that this may be used in a practical real-world clinical environment by expert therapists.The improvement in reported performances (e.g., speech accuracy up to 92and facial accuracy to 94directly correlates to the use of techniques from [10], [11] (for speech) and [1], [6], [12] (for face recognition).

The greatest latency reduction, from a time of 0.6 seconds to an incredible 0.4 seconds, was achieved by implementing the parallel model structure described in reference [14] effectively.

The overall system stability, which has been greatly enhanced by highly satisfactory multi-modal fusion methods, is basically a result of the combined integration and fusion of concepts from different sources, i.e., those embodied in references [3], [4], [5], [7], [8], and [9].

E. Future directions

Physiological data collected from wearable sensors, including heart rate and skin conductance, will be incorporated into future versions of the AI-based mental health companion. This incorporation is intended to improve the system's ability to accurately identify emotional states. Moreover, the incorporation of more data will allow for ongoing refinement and calibration of the underlying models, ultimately allowing them to identify a range of different emotions and provide users with more accurate and informative feedback (Feng et al., 2023).User research will therefore be pursued in this regard so as to further improve and fine-tune the suggestions provided by the system so that the suggested coping strategies are in line with the different needs and situations of the students. Additionally, the integration of mental health experts is absolutely necessary to ensure that the system is trauma-informed so that it provides the most optimal and helpful interventions to the users.

V. Implementation and Deployment

A. Deployment Environment

The mental health companion is implemented in the form of a cross-platform software that is highly flexible and easily accessible on web interfaces as well as mobile applications. It is implemented on a highly secure and scalable cloud-based infrastructure that is specifically optimized to support real-time data processing features and deliver low-latency performance. In order to offer maximum security for sensitive user data, advanced encryption methods and strict data security measures—borrowing from a number of methods that have been critically critiqued in reference [2]—are followed strictly throughout the entire process of the system. Apart from these security measures, the incorporation of edge computing components within the deployment framework drastically reduces any network latency, which is a key design consideration that is prompted by best practices defined for real-time applications, as described in reference [14].

B. System Integration and Technical details

The system can integrate multiple different modalities seamlessly to produce trauma-specific emotional responses. Each component of this complex system has been carefully thought out and constructed based on precious insights and epiphanies gleaned from the literature:

1) Speech Emotion Recognition:: The speech module employed an incredibly sophisticated Long Short-Term Memory (LSTM) network, carefully designed to evaluate speech signals and effectively detect an incredibly wide range of critical auditory cues. Among these cues—although quite clearly not exhaustively limited to them—are tone, pitch, and cadence features, each of which possesses incredible importance within the processing of spoken words. Along with this, an incredibly varied range of sophisticated feature extraction methods were employed, with a particular emphasis on Mel Frequency Cepstral Coefficients (MFCCs). These coefficients were subsequently employed with phenomenal effectiveness, based on the finely detailed methods detailed in references [10] and [11].

Through the implementation of these sophisticated methods, the overall detection accuracy not only improved but also improved quite considerably relative to the previously documented detection accuracy level of 85to a whopping 92Their work was significant and important in resolving processing lag, by the effective and strategic utilization of parallel processing techniques, a problem very much debated in reference [14].

2) Facial emotion recognition :: A Convolutional Neural Network, or CNN, is specifically crafted to process real-time video streams, with the exclusive purpose of detecting even the slightest variations in facial expressions. The system's overall architecture employs dynamic kernel methods, as described in reference [1], complemented by advanced convolutional techniques necessary for facial image feature extraction, as described in references [6] and [12]. Moreover, reference [13] employed methods and techniques that were crucial in building robustness, particularly against different levels of light, which otherwise would have been a major challenge. The final result was a considerable improvement in accuracy, bringing it up from a baseline of 88to an astonishing 94.

3) Text Emotion Recognition:: While this module uses the same CNN-based architectures, its main purpose is to aid the other modalities in sentiment detection from user-created text. The design is inspired by methods in [12] to allow text clues to be processed as well as visual and audio inputs.

4) Multi-Modal Fusion:: The integration of the outputs generated by the speech, facial, and text modules required the creation of a highly sophisticated and advanced fusion model. To do so, heterogeneous methodologies and strategies from sources [3], [4], [5], [7], [8], and [9] were synthesized and integrated cautiously. This was in order that the system might meaningfully and integrally process the complex and often intersecting signals which are characteristic of trauma-related emotion experienced by individuals. The capacity to perform this multi-modal integration is actually at the heart of the goal of delivering personalized, real-time coping strategies tailored to address the individual needs of the users. Building Capacity and Developing Durable Networks: The advanced network architectures presented in reference [15]—such as newer approaches by temporal-aware bi-directional multi-scale networks utilizing multi-head attention mechanisms—were systematically incorporated into the existing framework to further improve model performance. When these advanced techniques are combined with the parallel processing methods which are detailed in reference [14], they notably played a role in the remarkable reduction of the system's overall latency, down from a starting 0.6 seconds to a notably improved 0.4 seconds, and as a result, this enhancement considerably improved the system's overall responsiveness.

VI. Conclusion

This AI-based mental health companion, which has been developed through this project, provides a new and efficient approach to the problem of trauma-induced emotional distress in college students. By providing real-time coping strategies, progress tracking, and multi-modal emotion recognition, it supports the students in managing their emotions and recovering from trauma. This is the potential of this project that, as it keeps growing, it will eventually reach all students, giving them accessible mental health support that enables them to achieve their academic goals and maintain their emotional well-being.

References

[1] N. Perveen, D. Roy, and K. M. Chalavadi, "Facial Expression Recognition in Videos Using Dynamic Kernels," *IEEE Trans. Image Process.*, vol. 29, pp. 8316–8325, 2020.

[2] O. F. Mohammad, M. S. M. Rahim, S. R. M. Zeebaree, and F. Y. Ahmed, "A Survey and Analysis of the Image Encryption Methods," *Int. J. Appl. Eng. Res.*, vol. 12, pp. 13265–13280, 2017.

[3] V. Shrivastava, V. Richhariya, and V. Richhariya, "Puzzling Out Emotions: A Deep-Learning Approach to Multimodal Sentiment Analysis," in *Proc. 2018 Int. Conf. Adv. Comput. Telecommun. (ICACAT)*, 2018, pp. 1–6.

[4] J. Liang, S. Chen, and Q. Jin, "Semi-supervised Multimodal Emotion Recognition with Improved Wasserstein GANs," in *Proc. 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 695–703.

[5] E. Chandra and J. Y.-J. Hsu, "Deep Learning for Multimodal Emotion Recognition—Attentive Residual Disconnected RNN," in *Proc. 2019 Int. Conf. Technol. Appl. Artif. Intell. (TAAI)*, 2019, pp. 1–8.

[6] J. Chen, Y. Lv, R. Xu, and C. Xu, "Automatic Social Signal Analysis: Facial Expression Recognition Using Difference Convolution Neural Network," *J. Parallel Distrib. Comput.*, vol. 131, pp. 97–102, 2019.

[7] V. J. Aiswaryadevi, G. Priyanka, S. Sathya Bama, S. Kiruthika, S. Soundarya, M. Sruthi, et al., "Smart IoT Multimodal Emotion Recognition System Using Deep Learning Networks," in *Artificial Intelligence and IoT Smart Convergence for Eco Friendly Topography*, 2021, pp. 3–19.

[8] S. M. S. A. Abdullah, S. Y. A. Ameen, M. A. M. Sadeeq, and S. Zeebaree, "Multimodal Emotion Recognition Using Deep Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 2, pp. 52–58, 2021.

[9] S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, et al., "Deep Learning-Based Multimodal Emotion Recognition from Audio, Visual, and Text Modalities: A Systematic Review of Recent Advancements and Future Prospects," *Expert Syst. with Appl.*, 2024.

[10] B. J. Abbaschian, et al., "Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models," *Sensors*, 2021.

[11] M. B. Akc¸ay and K. Og˘uz, "Speech Emotion Recognition: Emotional Models, Databases, Features, Preprocessing Methods, Supporting Modalities, and Classifiers," *Inf. Sci.*, vol. 582, pp. 593–617, Jan. 2022.

[12] F. Z. Canal, T. R. Mu¨ller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, and A. C. Sobieranski, "A Survey on Facial Emotion Recognition Techniques: A State-of-the-Art Literature Review," *Inf. Sci.*, Oct. 2021. [Online]. Available: https://doi.org/10.1016/j.ins.2021.10.005.

[13] O. Oshin, "Large Scale Deep Learning Model and Database of Nigerian Faces and Voices," *IEEE Access*, Jan. 2023, doi:10.1109/ACCESS.2023.3282618.

[14] H. Yan, "Parallel Model Speech Emotion Recognition Network Based on Feature Clustering," *IEEE Access*, Jan. 2023, doi:10.1109/ACCESS.2023.3294274.

[15] L. M. Zhang, "Improvement of Multimodal Emotion Recognition Based on Temporal Aware Bi-Direction Multi Scale Network and Multi-Head Attention Mechanisms," *Appl. Sci.*, doi:10.3390/app14083276.