2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA) | 979-8-3315-3562-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/ACDSA65407.2025.11165999

# Personalized Mental Health Interventions Using Generative AI and Multimodal Data

Qianyu Yang
*School of Psychology*
*Central China Normal University*
Wuhan, China
yangqianyu@mails.ccnu.edu.cn

Zewen Xie
*School of Psychology*
*Central China Normal University*
Wuhan, China
17306304371@163.com

Gengfeng Niu*
*School of Psychology*
*Central China Normal University*
Wuhan, China
niugfpsy@ccnu.edu.cn

***Abstract*— Artificial Intelligence (AI) systems could provide customized mental health interventions based on multiple types of data. The study introduces dynamic graph convolutional neural network (DGCNN) which identifies emotions simultaneously from text signals, as well as voice and face data as a real-time recognition model. The research addresses the gap in effective multimodal integration by introducing a weighted decision-level fusion technique that compensates for modality-specific data degradation. This research implements one-frame-per-second constant video processing of IEMOCAP and MELD video clips through majority voting which serves as a reliable classification technique. By minimizing the differences between various input signals, the weighted decision-level fusion technique optimizes its accuracy performance. The assessment system combines dependable psychological measures which act as infrastructure to support AI-based customized therapeutic services. Key findings reveal that the proposed model achieves improved classification accuracy and supports early detection of emotional distress with minimal computational load. Such emotional evaluation immediately supports clinicians during early detection of psychological stress indicators. Advanced mental health care access is achieved through better patient reach and enhanced health achievement according to the implemented strategy.**

***Keywords*— *Dynamic graph convolutional neural network, emotion recognition, multimodal data fusion, generative AI and mental health assessment***

## I. INTRODUCTION

Multiple millions of people worldwide experience symptoms of anxiety and depression and post-traumatic stress disorder (PTSD). Although Cognitive Behavioral Therapy (CBT) and related interventions demonstrate efficacy in therapy delivery, the same benefits remain blocked by expensive healthcare costs and limited availability alongside insufficient personalized treatment [1]. Multiple factors including a lack of available professionals and public opinions about mental health create significant delays which prevent proper mental health care access. Specialist practitioners use combined information sources to develop mental health approaches which improve people's access to treatment. On the back of emerging AI, specifically [2], generative AI, there have come new solutions in the form of data led, scalable, personalized mental health support.

The therapeutic framework for generative AI includes augmenting it with multimodality of data sources and specific, customized recommendations [3]. The speech programs can be put on a stream of data that include not only speech patterns but also, how we interact with the medium, facial expressions, and a measure of heart rate variability. These programs can do some analysis for emotional health and prediction of the chances of being in a less-than-ideal mental state. Emotional Health Detection is an example of such an innovation which analyzes an individual's psychological state in real-time using data from multiple sources [4]. Users can use AI powered systems to be reinforced and be fed back to at regular intervals as well as making them more context aware and dynamic than ever with their mental health support. It therefore has the potential to personalize mental health care, close existing gaps, and enhance the ability to introduce interventions that are more accessible, more proactive, and perhaps more effective in text for people and places around the word [5]. The main contribution of the paper is presented as follows,

- Proposed a real-time multimodal emotion recognition framework that leverages visual, audio, and textual data through a DGCNN and employs a weighted decision-level fusion strategy to enhance classification robustness against modality-specific noise and degradation.
- Implemented an efficient processing pipeline using one-frame-per-second sampling and majority voting on IEMOCAP and MELD datasets, integrating clinically relevant psychological indicators to support early detection of emotional stress and enable scalable deployment in mental health monitoring systems.

## II. LITERATURE REVIEW

Advanced manipulation intelligence, which entails customizable help and assistance through real time analysis and adaptive learning together with solutions for mental health illness that use artificial intelligence, are supported. It makes the strategies effective and available but it also has down sides of data reliance and confidentiality concern. Based on the AI, merits and demerits of the works, the Table 1 summarizes the published works.

TABLE I. PROBLEM FORMULATION

| Author(s) | Techniques Involved | Advantages | Disadvantages |
|---|---|---|---|
| H. Hadjar et al.,[6] | Deep learning for facial emotion analysis | Real-time diagnosis, enhances teleconsultation | Needs high-quality data, privacy concerns |
| G. Yadav etal.,[7] | Multi-modal data, Ensemble learning | High accuracy, diverse data sources | Complex, computationally intensive |
| S. Zamani et al.,[8] | IoT, Machine learning | Considers environmental factors | Sensor dependency, security issues |
| N. P. Shetty et al.,[9] | NLP, Sentiment analysis | Extracts large-scale emotional patterns | Bias, context challenges |
| D. Kodati et al.,[10] | Soft-parameter sharing transformers | Better generalization | Needs large datasets, high training time |

Different AI assessment methods for mental health evaluations are examined throughout the reviewed works which use speech patterns alongside facial expressions and user interactions and body signals. All methods used for AI evaluation have both favourable aspects and disadvantages regarding precision accuracy but struggle with data confidentiality and lack customization capabilities and real-time adaptability. To improve flexibility as much as possible, DGCNN is used to learn intricate correlations between the multimodal input for real time emotional analysis. Given all of this, our method is a success for AI driven mental health support as it is scalable, adaptable to learning, and delivers strong performance.

## III. PROPOSED SYSTEM MODEL

Human behaviour greatly relies on emotions as they exist simultaneously in positive and negative categories. The development of mental health concerns including stress and depression and anxiety arises from negative emotions. Positive emotions enhance a person's lifestyle quality and overall wellness but negative emotions have negative impacts on reasoning abilities and health status. Real-time video surveillance required three operational modes to develop a model to detect human emotions expressively. Video surveillance data obtains two different information streams that consist of audio and visual data. When the subject appears in the camera view yet the audio recordings remain unclear or when the opposite happens, it produces a pronounced negative impact on the quality of either visual or audio data. The emotional state evaluation includes measuring the actual contents discussed by the subject. The audio file gains transcription through database-related text to allow instant real-time calculations. The emotional state evaluation of the subject derives from a weighted union between visual data and text materials and audio sensory input. Researchers chose decision level fusion after showing that regulating all three elements until reaching a choice point makes them free from depending on data quality from different modalities.

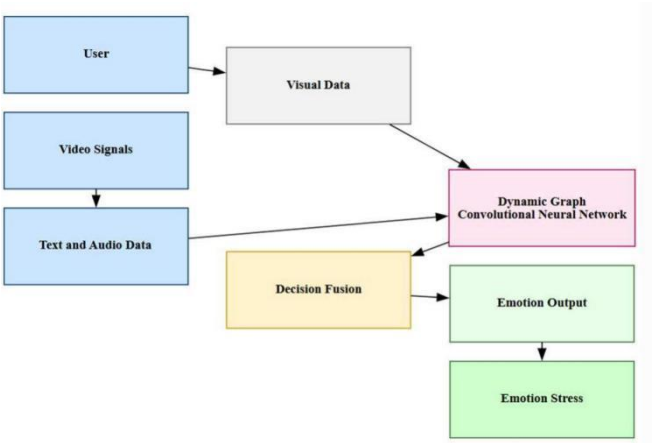


Fig. 1. Proposed flow architecture

A video represents the temporally and spatially related collection of pictures. The recommended objective for real-time computation involves this data which includes brief speech. Facial expressions in this analysis are studied through selected photographs extracted from the video clips. Temporal features are not an element in this process which would provide benefits while analysing extended videos [11].

### *A. Process of the proposed method*

The sampling rate selection for both video clips in the IEMOCAP and MELD sub-databases is established at one frame per second. Utterances remain so short that peak frames should be avoided for validation purposes. Video clips need to undergo emotion classification on their individual frames then use a majority voting approach for the analysis.

The DGCNN utilizes the selected scene frames to extract necessary facial information through its main focus on facial expressions [12]. Faceless frames need to be excluded through analysis before the pre-trained architectural model accepts valid face images in the enhancement process.

The application of multimodal data fusion divides into three primary strategies which are early fusion and joint fusion and late fusion. The researchers in this study opted for decision-level fusion as their last step after running simultaneous computations on all components to prevent degradation of performance resulting from diverse data quality across modalities. The system achieves continuous results by making use of data obtained from multiple modalities.

Simulation results demonstrate that observer can identify subjects while revealing poor or distorted audio quality. Subject emotional assessment requires a thorough analysis of their verbal communication speech throughout the interaction.

The system implements weighted decision fusion strategies to unite results generated through text models as well as audio components and graphical models. The results from specific custom model versions adjust automatically to data changes resulting from various operational conditions in each modality. The signals enter individual learning procedures which detect their distinctive roles in emotional recognition.

The research evaluates the performance impact from different fusion methods on the system as a whole. Weighted late fusion forms the basis of the proposed technique since

detection ratios define the weight assignment process. The ratios established in these scores surpass the F-scores specified by the architecture to support final decision processes. The calculation for Detection Ratio (DR) works as shown below: The formula for Detection Ratio (DR) appears as follows:

$$DR = \frac{TP}{(TP+TN+FP+FN)} \quad (1)$$

Here, TP is a True Positive, FP is a false positive, FN is a False Negative and TN is a True Negative. The calculation of DR takes place for each class because the model detects seven distinct emotional states which results in a total of seven classes. The calculation of output class weight W occurs according to the following rule [13]:

$$W = 1 - DR \quad (2)$$

Model predictive scores are obtained by multiplying class weights with probability vectors from each model that are associated with the classes.

$$S = W * P \quad (3)$$

Where, W is a weight, S is the score and P is the probability score. To compute the model, score the program performs addition between every class weight. The output stage selects the prediction from the maximum function since this model contains the highest score for the test case. This proposed design contains 3 modalities for 3 models. The system provides the final output in this class at termination.

$$\text{output class} = \text{Max}(\text{visual}, \text{audio}, \text{text}) \quad (4)$$

Where, "isual is the visual data, audio is the audio data and text are the text data. A late weighted fusion method selects the output class from the most suitable architecture for the target output class.

*B. Dynamic graph convolutional neural network*

Two parallel verification approaches for graph creation with functional time series exist which lead to topic level graph construction and node level graph construction independently. The subjects function as nodes in this node-level construction while brain functional connectome and PCD information from each participant becomes the basis to create vectorized node features. The adjacency matrix of this graph demonstrates different similarity measurement techniques to compute the node's features distances. The proposed method uses node-level categorization to connect all nodes that belong to the same class within a single community. The training of representative characteristics for every graph node occurs through Graph Convolutional Network (GCN) by analysing neighbouring vertex features. A fundamental convolutional layer for spectral or spatial GCN contains two primary functions which are updated followed by aggregation processes. The fifth aggregating procedure operates in the following manner.

$$\alpha_I^F = \text{Aggregate}\left(\left\{H_J^{F-1}, W_J^{F-1} e(I)2,\ J \in N(I)5,\ H_I^F = \text{update}\ (\alpha_I^F, H_J^{F-1}, e(I)5\right.\right. \quad (5)$$

Where, $9_{:}^{F}$ is defined as attention coefficient, $<_J^{F-1}$ is defined as the hidden feature representation, $W_{:}^{F-1}$ is defined as the hidden feature representation, $<_{:}^{F}$ is defined as the feature vector of the node within the fth layer, e(I) is defined as the pair of connected layers and $>(:)$ is defined as the two neighbouring nodes connected to the central node with edges.

The permutation invariant features gathered from neighbourhood nodes features is processed by the aggregation operator which forms aggregated features and the upgrade operator utilizes an MLP with gated architecture or convolutional layer. The standard receptive field of graph attention network and other conventional GCN layers stops at 1-hop boundaries restricting their ability to identify complex functional connections from diverse node characteristics.

Data movement to 1-hop regions yields results regarding the central ROI when combined with two-hop or multi-hop procedures. The standard implementation of deep GCN layers enables users to obtain two-hop or multi-hop data connections. During deep-layer operation over smoothing occurs alongside weight convergence in stationary points which makes nodes represent features differently from input characteristics until gradient values reach zero. The planned construction of deeper layers does not occur because of this reason. In each convolutional layer, the suggested method aggregates the 1-hop and 2-hop neighbor data. The regular operation function for graph convolution takes the following form:

$$<_{?}^{F} = M@A>BF_{CDPE}(<_G^{F-1} II <_G^{F-1} - <_{?}^{F-1}), JK>(:)L = \Sigma_{GRN(?)} B\theta_{0,} <_{?}^{F-1} IIP_N. (<_G^{F-1} - <_{?}^{F-1})L \quad (6)$$

The data encoding process for 2-hop neighbourhood information utilizes $\delta_N \cdot \left(H_J^{F-1} - H_I^{F-1}5\right.$ which subtracts 1-hop node data from the whole adjacent data of 1-hop neighbouring nodes while encoding 1-hop neighbourhood information as $\left(\theta_n . H_I^{F-1}5\right.$ and implementing the concatenation operation as $\|$ with $\Theta = U\theta_1, \cdots \theta_S; P_1, \cdots, P_> X.$

In the proposed framework, facial expression features are extracted using a CNN based on a pre-trained ResNet-50 architecture, selected for its proven efficiency and accuracy in visual recognition tasks. The CNN accepts input images resized to 224×224 pixels and processes them through a series of convolutional layers with 3×3 kernels, batch normalization, and ReLU activation functions. These layers are followed by residual identity blocks and max pooling layers to progressively capture complex spatial patterns while reducing dimensionality. A global average pooling layer and a fully connected layer generate the final feature vector representing the facial expression. This feature vector is subsequently passed into the DGCNN for emotion classification within the multimodal fusion setup. The choice of CNN over alternative architectures such as LSTM, GRU, or 3D CNNs is motivated by its superior performance in spatial feature extraction from individual video frames, its ability to operate efficiently in real-time settings, and the simplicity of integration into decision-level fusion pipelines. Since the system focuses on short utterance video clips where temporal dynamics are minimal, CNN offers a practical and computationally efficient solution that balances accuracy with speed, ensuring robust facial analysis without the overhead of sequence modelling.

## IV. Performance Evaluation

Clinical mental health assessment benefits from using the proposed generative AI-based multimodal framework due to the advantages observed during CNN, LSTM, and RNN model comparisons. Throughout the evaluation process the proposed system outperforms other methods based on accuracy, precision, recall, F1-score, false positive rate (FPR) and false negative rate (FNR). To validate the proposed methodology, 8-fold cross validation is considered. Failing to store sequences across time represents a drawback of specialized LSTM systems while requiring substantial processing training. The advantage of CNN models to extract features intuitively brings difficulties to maintain temporal relationships within the data. RNNs provide excellent capabilities for processing sequential data but their ability to learn from distant dependencies is negatively affected by gradient vanishing problems. The proposed system combats detection obstacles through multidimensional data processing and advanced deep learning systems which leads to better mental condition identification. The lowered percentage of incorrect positive and negative predictions improves assessment reliability which helps prevent diagnostic mistakes and increases accuracy. The new method proves to be an optimistic technique which enabling precise mental health treatment through early identification along with customized therapeutic plans.

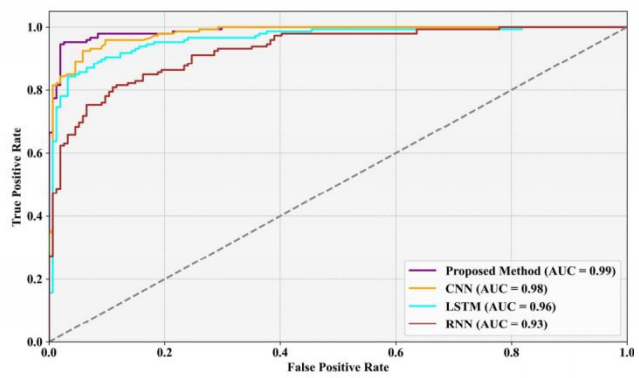


Fig. 2. ROC Evaluation

An ROC curve reveals that the proposed method obtains better assessment outcomes than CNN, LSTM, and RNN in mental health evaluation and presented in Fig 2.

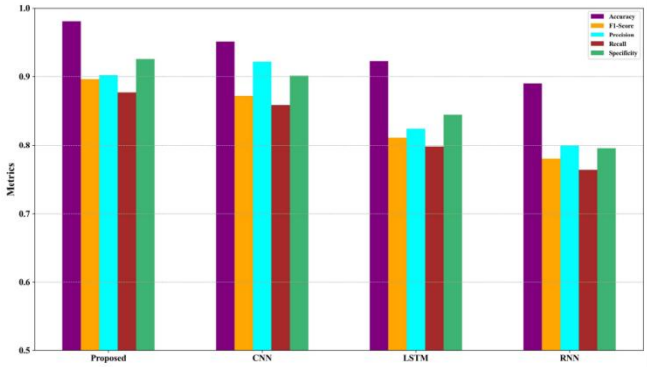


Fig. 3. Validation of measures

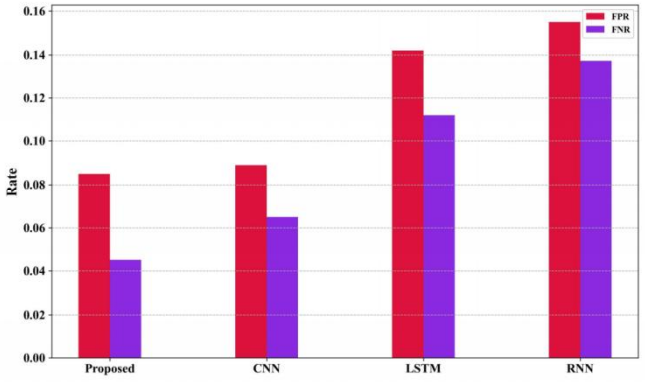


Fig. 4. Validation of measures (FPR and FNR)

The proposed diagnostic method demonstrates superior performance when compared to existing models because it maintains optimal sensitivity and specificity in mental health diagnostics. The proposed method delivered outstanding performance because it outpaced CNN, LSTM, and RNN to obtain accuracy levels reaching 0.98. The proposed model delivered superior performance by achieving outstanding F1-score, precision, as well as recall parameters and presented in Fig. 3. The accuracy level of CNN exceeded that of LSTM and RNN at approximately 0.95. The model deployment leads to enhanced mental health evaluation reliability because the results shown in Fig. 4 demonstrate this increase. The proposed method achieved the minimum values of FPR and FNR at 0.08 and 0.045 for optimal classification reliability. The FPR result of CNN was 0.085 and the FNR level was 0.065 whereas LSTM showed significantly higher FPR at 0.14 and FNR at 0.11. RNN presents the highest rates of misclassification through its unacceptable FPR of 0.155 together with FNR of 0.13. Diagnostic accuracy for mental health assessment increases when the proposed model decreases both FNR and FPR.

## V. CONCLUSION

This paper presents a generative AI working with multimodal information fusion approaches manages to compute mental health by the validation of voice patterns combined with facial expressions and textual signals. This method delivers dependable emotional state analysis through DGCNN alongside weighted decision-level fusion techniques to help medical professionals discover mental health problems in advance and design specific clinical treatment methods. The method faces several critical weaknesses related to its susceptibility to changes in background noise, illumination, and video footages that may get obstructed while the pre-trained models maintain potential biases stemming from unbalanced dataset collections. The implementation of real-time deployment demands extensive computational power making it impossible to achieve through limited resource settings. Research focus will move towards implementing performance improvements of models together with scalability advancements while making improvements to real-time implementation capabilities.

To address potential privacy and data confidentiality concerns associated with using multimodal data (video, audio, and text) in mental health assessment, the proposed system employs anonymized datasets where personally identifiable information is removed prior to processing. Future implementations will integrate privacy-preserving techniques

such as secure data encryption, on-device processing, and federated learning to prevent data leakage and ensure compliance with ethical standards. All data handling follows strict confidentiality protocols aligned with relevant data protection regulations. These measures help build trust and safeguard user information throughout the assessment process.

## REFERENCES


[1] M. Nykoniuk, O. Basystiuk, N. Shakhovska, and N. Melnyk "Multimodal data fusion for depression detection approach," International Journal of Computation, vol. 13, no. 1, 2025.

[2] Z. Chen, D. Wang, L. Lou, S. Zhang, X. Zhao, S. Jiang, J. Yu, a Xiao, "Text-guided multimodal depression detection via cross-n feature reconstruction and decomposition," Information Fusion, 117, 2025.

[3] R. Flores, M. L. Tlachac, A. Shrestha, and E. A. Rundenste "WavFace: A multimodal transformer-based model for depre screening," IEEE Journal of Biomedical and Health Informatics, 2(

[4] S. Hu, W. Fang, H. Bao, and T. Zhang, "Non-contact detectic mental fatigue from facial expressions and heart signals: A supervised-based multimodal fusion method," Biomedical S Processing and Control, vol. 105, 2025.

[5] E. Abdelfattah, S. Joshi, and S. Tiwari, "Machine and deep lea models for stress detection using multimodal physiological data," Access, 2025.

[6] H. Hadjar, B. Vu, and M. Hemmje, "TheraSense: Deep learnin facial emotion analysis in mental health teleconsultation," International Journal of Electronics, vol. 14, no. 3, 2025.

[7] G. Yadav, M. U. Bokhari, S. I. Alzahrani, S. Alam, and M. Sh "Emotion-aware ensemble learning (EAEL): Revolutionizing m health diagnosis of corporate professionals via intelligent integrati multi-modal data sources and ensemble techniques," IEEE Ac 2025.

[8] S. Zamani, M. Nguyen, and R. Sinha, "Integrating environmental for mental health monitoring: A data-driven IoT-based approach, International Journal of Applied Sciences, vol. 15, no. 2, 2025.

[9] N. P. Shetty, Y. Singh, V. Hegde, D. Cenitta, and D. K. Dh "Exploring emotional patterns in social media through NLP mod unravel mental health insights," Healthcare Technology Letters, 2(

[10] D. Kodati and R. Tene, "Advancing mental health detection in tex multi-task learning with soft-parameter sharing transformers," N Computing and Applications, vol. 37, no. 5, pp. 3077–3110, 2025.

[11] Kaggle, "MELD dataset." [Online]. Avail https://www.kaggle.com/datasets/zaber666/meld-dataset. [Acce Apr. 15, 2025].

[12] F. Zhu, J. Zhang, R. Dang, B. Hu, and Q. Wang, "MTNet: Multin transformer network for mild depression detection through fusi EEG and eye tracking," Biomedical Signal Processing and Contro 100, 2025.

[13] J. Kim, O. Park, and E. Park, "MOSS-6: A multi-label dataset and learning model for detecting diverse social support-seeking behav in online mental health communities," Information, Communicati Society, pp. 1–35, 2025.