SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I4P114

Volume 13 Issue 4, 183-193, April 2026

Original Article

MERMHS: A Multimodal Emotion Recognition Framework Using Probability- Based Late Fusion for Mental Health Monitoring

Yellamma Pachipala[1], Dhanush Vardhan Yalamati[2], Pavan Kumar Karubhuktha[3], Gayathri Jagarlamudi[4], Pavani Challa[5]

[1,2,3,4,5]*Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India.*

[1]*Corresponding Author : pachipala.yamuna@gmail.com*

Received: 11 January 2026 Revised: 12 February 2026 Accepted: 15 March 2026 Published: 30 April 2026

Abstract *- Mental health issues are now more common than ever, and a need arises to have strong, intelligent systems that can accurately identify human emotional states. Although Artificial intelligence provides advanced methodologies, many existing systems are facing challenges in achieving the best accuracy in human emotion detection in the real world. This is due to variations in facial expressions, background noise in speech signals, contextual ambiguity in textual inputs, and low-performance fusion techniques. The proposed work is a Multimodal Emotion Recognition Mental Health System (MERMHS) that aims to narrow this gap, which uses CNN for video-based face expression recognition, LSTM is applied for audio emotion recognition through Speech signals, and Bi-LSTM is utilised for text emotion recognition through textual inputs. The investigation shows that the proposed MERMHS approach by using the CMU-MOSEI dataset achieves an accuracy of 92.7%, a precision of 93.70%, a recall of 92.67%, and an F1-score of 93.10%. Compared with the existing approach, the proposed MERMHS is superior because of the probability-based late fusion technique.*

Keywords *- Facial Emotion Detection, Text Emotion Detection, Speech Emotion Detection, Multimodal Emotion Recognition, CNN, Bi-LSTM, LSTM.*

1. Introduction

Nowadays, the use of AI is rapidly increasing in the modern world, and it is also being applied in the healthcare sector [2]. The usage of AI in the healthcare sector has quickly increased during the COVID pandemic. Because of this, AI usage has also grown in healthcare. Additionally, in the field of mental health care, AI is being used to identify challenges such as anxiety, depression, and stress, which have increased in today's generation mainly due to social media [5]. Traditional methods of diagnostics often involve the use of subjective judgments and have been associated with delayed clinical interventions. This has limited the ability to offer support to those who need it in a timely manner. Advances in AI have tremendous potential for improving mental health analysis.

The emotional recognition system uses multimodal data, such as mainly facial expressions, to identify suitable emotional cues [12]. Behavioural analytics can capture patterns in daily activities like writing a diary or communication, and social interactions [1]. In this project, we also use speech, facial expressions, and text to identify emotional expressions and feelings. Additionally, behavioural analytics are derived from the text, where users can write whatever they like in their diary within our chatbot, interact with it, and allow the system to analyse their behaviour. Using both behavioural analytics and emotional recognition, we can assess a person's mental health, including stress and anxiety. We can then suggest possible treatments or recommend visiting a psychiatrist. The system can also provide basic remedies to help reduce mental health issues and determine whether someone truly needs to see a doctor. This approach can help save doctors' time by efficiently screening individuals [2].

Despite the rapid advances in artificial intelligence for healthcare applications, mental health monitoring remains a complex and underdeveloped area. Currently, AI-based systems primarily focus on either emotional recognition or behavioural analytics [12]. Coming to emotional recognition, they are using only one among these three, which are speech, text, and facial expressions. When it comes to behavioural analytics, they are using only text, which is also connected to one artificial intelligence bot [5]. Because of this, the

behaviour analytics are very complex to get accurate results because there is no communication between anyone to understand their feelings and emotions [10]. So, in our project, we are taking both behavioural analytics and emotional recognition, combining these two. This can lead to comprehensive and exact results, open individual mental wellbeing, and existing approaches often neglect personality, cultural sensitivity, and real-time adaptability, which are essential for empathetic AI-driven mental health support.

The growing global crisis of mental health care disorders highlights an urgent requirement for innovative, data-driven, and empathetic support systems [14]. Early identification of these problems is crucial to prevent psychological issues such as anxiety, depression, and stress. For this, the traditional methods are not suitable because they are time-consuming and not accurate, also. So, we are using modern methods by integrating AI and using some ML models and DL models. By integrating all this, we are creating an AI bot and a user interface with a dashboard, so that we can take care of the two methods we have used in this, which are behavioural analytics and emotional recognition.

The main goal of the proposed work is to develop a predictive and empathic approach for Mental health [1] that effectively integrates emotional recognition and behavioural analytics to enable prompt detection and prediction of psychosocial health problems. It also helps us to get accurate information from the patient; it saves the time and money of the patients, and it ensures that the treatment is for the right one only. At the same time, many existing models focus on one feature, like text or facial expressions. Our proposed model integrates both through a multimodal approach.

Problem Statement: The existing emotion recognition systems use a single mode of interaction with the environment, which affects their accuracy and reliability in real-world applications. Thus, to ensure accurate emotion detection in mental health monitoring systems, a strong multimodal framework that uses facial, textual, and vocal data is needed.

Key contributions of the proposed MERMHS framework:

1. Comprehensive Multimodal Emotion Recognition Framework:

A unified framework that uses different features, facial for CNN, textual for Bi-LSTM, and speech for LSTM with MFCC modalities to recognise emotions in an integrated manner.

2. Innovative Probability-Based Weighted Late Fusion:

Weighted averaging is a straightforward yet effective method that increases the accuracy of the suggested framework.

3. Increased Robustness to Noisy or Missing Modalities:

The proposed framework ensures high accuracy even in the presence of noisy or missing modalities.

4. Better Performance on Benchmark Dataset:

The proposed framework achieves an accuracy 92.7%, a precision 93.70 %, a recall 92.67%, and an F1-score 93.10%. on the CMU-MOSEI benchmark dataset.

5. Application-Focused Design for Mental Health Monitoring:

The proposed MERMHS is designed to recognise accurate and reliable emotion detection in real-world environments, correlated to MHS.

The information in the article is organized as follows. Section 2 offers a detailed review of the literature on ML and DL models. Section 3 showcases the current strategy for increasing performance metrics with various parameters using the proposed methodology, which is a multi-model for identifying emotions and behaviours at a better level. We discuss the machine learning models in detail. Section 4 displays the results and a comparison of our proposed approaches. Section 5 specifies a comprehensive overview of the proposed work, key findings, and closing remarks.

2. Literature Survey

The literature overview on current approaches to improve performance and accuracy is provided below. It offers a comparison of several methods, like CNN, LSTM, and Bi-LSTM.

In 2025, Qianhe Ouyang investigated audio-based emotion classification using CNN-LSTM on SAVEE and RAVDESS datasets [1]. They used only audio input, and the accuracy of the disgust emotion is only 38%, with incorrect classification between Fear and Sad emotions. Late fusion probability weighting can reduce this type of confusion.

A multimodal speech emotion detection was proposed by Shamin Bin Habib Avro, Taieba Taher, and Nursadul Mamun in 2025 [2]. The proposed research work uses BiLSTM and CNN models on the IEMOCAP dataset. BiLSTM and CNN are integrated in early fusion; however, this model can be less sensitive to small noises in imbalanced classes.

Guowei Zhong, Junjie Li, Huaiyu Zhu, Ruohong Huan, and Yun Pan studied the topic of cross-modal semantic alignment in the year 2025 [3]. In the study, the researchers used PLGM, PFM, and MCR to work with the CMU dataset. Although the PLGM-PFM-MCR framework successfully handles inconsistencies through a balance, the study only used basic preprocessing overheads.

Shuo Zhang, Jinsong Zhang, Zhejun Zhang, and Lei Li (2025) carried out a review of multi-task affect processing [4]. In their study, the authors developed the Mixture of Low-Rank Experts with UniTSE fusion model. Although the Low-Rank Experts model optimises parameters, its scalability limitations are evident.

Chengyan Wu, Yiqiang Cai, Yang Liu, Pengxu Zhu, Yun Xue, Ziwei Gong, Julia Hirschberg, and Bolei Ma conducted a survey in 2025 on various Multimodal Emotion Recognition techniques in Conversations (MERC) [5].

They discussed preprocessing and multimodal large language models (MLLMs) on the IEMOCAP and MELD datasets [5], and the accuracy is moderate due to the low complexity of the fusion methods.

Prashant Umar Nag, Bhagat, and Vishnu Priya in 2025 investigated the task of AI in Healthcare, a detailed review of facial emotional detection [6]. The authors worked on the DAIC-WOZ dataset using different models like LSTM, CNN, Bi-LSTM, and CNN-LSTM, which claimed to be emotion recognition, but the results showed the best accuracy and F1-Score, which supports speech and text only.

Puneet Kumar and Xiaobai Li (2025) proposed interpretable multimodal emotion recognition (MER) systems [7] by using the VISTANet-KAAP technique on the IIT-R MMEmoRec database, which supports the flexible adaptation of the hybrid fusion strategy through the VISTANet technique. Although the proposed technique results in higher accuracy, emotions are limited to four types: angry, happy, hate, and sad.

In 2025, Yehun Song and Sunyoung Cho assessed the performance of pre-trained vision-language effect models [8]. The authors created a MER-CLIP model by adding a label encoder and a cross-modal decoder, which was applied to the MOSI-MOSEI dataset. Although MER-CLIP provides semantic alignment, there are possible neutral effects due to the uncertainties in the labels.

An investigation of the potential for improved focus in Attention-Enhanced Speech Emotion Recognition was conducted in 2023 [9]. The models DBN, SDNN, LSTM-ATN, and CNN-ATN were tested with the datasets SAVEE-RAVDESS. And the model LSTM-AT.

In a review, Sachin and Prakash Mohan (2025) discussed the developments in the application of AI in the domain of mental health care [10]. They suggested the combination of NLP, ML, and DL architectures, and technology-based cognitive-behavioural therapy and virtual/augmented reality therapy. These methods hold promise in providing better access to therapy, but the issue of bias and privacy in AI remains a neutral ethical challenge.

Feiyang Chen, Ziqian Luo, Yanyan Xu, and Dengfeng Ke studied audio-text sentiment fusion using the DFF-ATMF on CMU-MOSI and IEMOCAP datasets [11]. DFF-ATMF framework is effective in fusing complementary features; however, neutral adjustments might be necessary to achieve modality alignment.

Sanmay Kotkar proposed a study on real-time multimodal affect detection in 2025 [12]. They used only two CNN and Bi-LSTM models with hybrid fusion here. The proposed model achieves the best accuracy on clear faces, such as happiness. Also, the model has a lot of confusion about these disgust and fear emotions.

Fazliddin Makhmudov et al proposed a study on cross-modal affect recognition using a deep learning model [13]. The study used a CNN-BERT model on the CMU-MOSEI-MELD dataset; however, overfitting in the attention mechanism remains a problem.

Dilnoza Mamieva, Akmalbek Bobomirzaevich, and Abdusalomov proposed a study on facial-speech multimodal MER in 2023 [14]. Using the IEMOCAP-MOSEI dataset, the study employed a CNN model and a RepVGG model. However, the study found that the attention mechanism is effective.

Mittal, Bhattacharya, Chandra, Bera, and Manoch presented a robust framework for multimodal emotion recognition fusion. The authors proposed a multiplicative model, M3ER, and evaluated it on the IEMOCAP-MOSEI datasets [15]. Although the multiplicative model is effective in attenuating noise, sensor issues remain a problem in achieving balanced robustness.

Xu, Zhijing, and Gao, Yang researched cross-modal emotion recognition based on multi-layer semantic fusion [16]. They developed a model called CM-MSF-BiLSTM-Conv1D-MGF using the CMU-MOSI-MOSEI dataset. The model is evaluated on Noisy real-world datasets, which also have many trainable parameters. Not suitable for real-time.

Iyortsuun, Ngumimikaren, Hyung Kim, John, Hyung, and Sudarshan Pant (2023) used ML and DL models for mental health care [17]. In the proposed model called “Tri-modal system,” the features used are “speech features” (MFCC, prosody), “text features” (BERT, sentiment analysis), and “facial video features” (OpenFace, Optical Flow). These are incorporated using “CNN-LSTM” (BiLSTM, 3D-CNN). This model is applicable to small data sets and has a high computational cost.

Ansari, Kashif, and Mansoor (2024) used the idea of employing AI to identify mental health signs early on in social media. [18]. In the proposed model, “CNN,” “3D CNN,” and “CNN-LSTM” are used. These are validated using “accuracy” and “F1 Score.” This model is applicable to small data sets.

Jinghui, Changsong Liu, Tianchi, Dahuang Liu, and Team (2025) used the concept of mental health detection using audio and text-based multi-models in the assessment of mental health care disorders [19]. In the proposed methodologies, “speech preprocessing” (noise reduction, “MFCC,” “pitch”)

and "text preprocessing" (BERT, Word2Vec) are used. Also, "CNN," "LSTM," and "BiLSTM" are used along with "attention." "Late fusion" is used in the multimodal model. Some of the drawbacks of the model are the use of small data sets, "speech noise variance," and "cultural and linguistic bias."

Nafiseh Ghaffar Nia, Kaplanoglu, and Ahad Nasab (2022) used the concept of health forecasting using AI [20]. In the proposed model, "hybrid models" are used. These include "ANN," "CNN," "RNN," "DNN," "DBN," "Autoencoder," "SVM," "DT," "KNN," "Fuzzy," and "IoT." These models increase the efficiency of diagnostics, which is a major advantage. However, there are data requirements and the lack of "explainability."

3. Materials and Methods

The proposed system is a multimodal framework for emotion detection that can identify human emotions via facial images, textual expressions, and speech signals. Each of the modalities will be analysed by a specific deep learning model to achieve the highest classification accuracy possible.

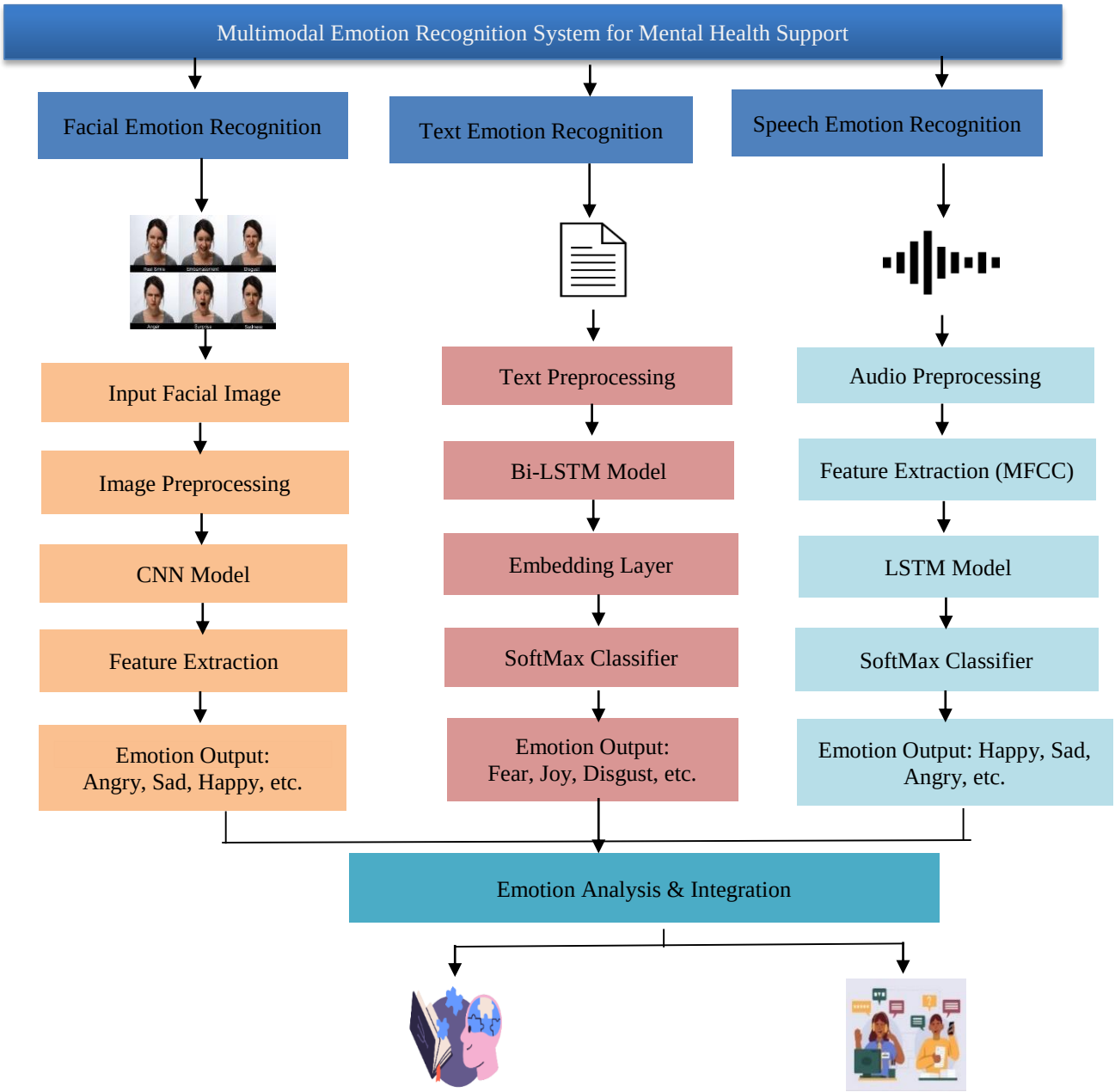

Fig. 1 Proposed MERMHS Architecture

In Figure 1, the three pipelines are represented.

- Convolutional Neural Networks (CNN) for Facial Emotion Recognition.
- Bidirectional Long Short-Term Memory (Bi-LSTM) for Text Emotion Recognition.
- Recognizing emotion speech using Long Short-Term Memory (LSTM).

Predicted probabilities generated by each model are integrated using a probability-based late fusion strategy to recognise the final emotional state.

3.1. Dataset Description

With more than 23,500 annotated video clips from more than 1,000 YouTube speakers on 250 various themes, including politics and entertainment, CMU-MOSEI is the largest dataset for multimodal sentiment analysis and emotion recognition. It was collected by the MultiComp Lab at Carnegie Mellon University and includes 65+ hours of data with a well-balanced gender distribution, representing three modalities: language (transcripts), visual (facial expressions), and acoustic (audio).

Each sentence-level speech has fine-grained annotations for sentiment intensity (-3 to +3 scale) and six fundamental emotions like “happy”, “sad”, “angry”, “fearful”, “disgusted”, and " surprised”, allowing for effective training of complex models such as transformers for opinion mining. It has more scale, speakers, and natural "in-the-wild" diversity than its predecessors, such as CMU-MOSI, with more varied cameras and microphones. Use it with CMU's Multimodal SDK for emotion AI research, but in this case, the research has been formulated as multi-class emotion classification, where each utterance is categorised into six discrete emotion classes: ‘anger’, ‘happiness’, ‘sadness’, ‘fear’, ‘disgust’, and ‘surprise’.

There is no cross-validation performed in this experiment. To ensure that the results are reproducible and consistent with other studies, we chose to use the predefined splits for train, validation, and test sets provided by the CMU-MOSEI official implementation. This means that the Training split contains roughly 15,891 utterances, the Validation split contains roughly 1,267 utterances, and the Test split contains roughly 4,168 utterances.

The evaluation protocol is strictly speaker-independent, meaning that no speaker is present in more than one split. This ensures that there is no speaker or utterance leakage.

Proposed Algorithm: Probability-Based Late Fusion Model

Algorithm: Probability_Based_Late_Fusion_Model

Input:

X → Facial video frames

T_u→ Text sequence

A_u→ Audio signal

θ_cnn, θ_bilstm, θ_lstm → Model parameters

w1, w2, w3 → Modality weights (w1 + w2 + w3 = 1)

Facial Emotion Recognition (CNN)

I_frames = ExtractFrames(X, fps = 30)# Framing

I_resized = Resize(I_frames, 224, 224)# Resizing

I_norm = (I_resized - 0) / (255 - 0)# Normalisation

F_conv = Convolution(I_norm, K, b)# Convolution

F_relu = max(0, F_conv)# Activation

F_pool = MaxPool(F_relu)# Pooling

F_img= Flatten(F_pool)# Flattening

$Z\_img = W\_img^{T} \cdot F\_img + b\_img$# Dense

P_image = Softmax(Z_img)# Probability

Text Emotion Recognition (Bi-LSTM)

T_tok = Tokenize(T_u)# Tokenization

T_low = Lowercase(T_tok)# Lowercasing

T_pad = Pad(T_low)# Padding

T_seq = IntegerEncode(T_pad)# Encoding

E= Embedding(T_seq)# Embedding

h_fwd= LSTM_forward(E, θ_bilstm)# Forward

h_bwd= LSTM_backward(E, θ_bilstm)# Backward

F_txt= Concatenate(h_fwd, h_bwd)# Concatenation

$Z\_txt = W\_txt^{T} \cdot F\_txt + b\_txt$# Dense

P_text = Softmax(Z_txt)# Probability

Speech Emotion Recognition (LSTM)

A_mono= ConvertToMono(A_u)# Mono

A_resample = Resample(A_mono)# Resampling

A_frames= FrameSignal(A_resample)# Framing

MFCC_feat = DCT(log(|FFT(A_frames)|))# MFCC

MFCC_norm = Normalize(MFCC_feat)# Normalization

F_audio = LSTM(MFCC_norm, θ_lstm)# Temporal

$Z\_audio = W\_audio^{T} \cdot F\_audio + b\_audio$# Dense

P_audio = Softmax(Z_audio)# Probability

Fusion

For each emotion class e:

P_final(e) = w1 * P_image(e)+ w2 * P_text(e) + w3 *

P_audio(e)# Fusion

k = argmax(P_final)# Prediction

End

Output:

k → Final predicted emotion label

3.2. Facial Emotion Recognition Using CNN

Video Preprocessing: Each video is first divided into individual frames at a fixed frame rate of 30 frames per second because it captures facial changes clearly while keeping computation efficient. From each frame, the face region is detected and extracted.

$$x_{norm} = \frac{(x - x_{min})}{(x_{max} - x_{min})} \quad (1)$$

Before being passed to the CNN model, all extracted face frames are converted into a fixed resolution of 224 × 224. This is done so that every frame has the same shape, because CNN models require a consistent input size. After resizing, the pixel values are scaled down into the range of 0 to 1, which helps the model train faster and makes the learning process more stable.

Normalisation is performed using Eq. (1). Here, x is the original pixel value, x_min is the minimum intensity (typically 0), and x_max is the maximum intensity (typically 255). After applying this, the output pixel value (x_norm) is between 0 and 1.

Convolutional Neural Network Architecture: CNN is best for video because it learns patterns automatically, like Eye shapes, smiles, Mouth Curves, and eyebrow movements, so it is used.Convolution Operation

$$Y(i,j) = \sum_{m} \sum_{n} X(i+m, j+n) \cdot K(m,n) + b \quad (2)$$

In Eq. (2), X is the input frame, K is the convolution kernel, b is the bias term, and Y (i, j) is the output feature map value, where i and j are the current output position, m and n are the kernel index positions. This process enables the model to learn emotion-specific patterns while preserving spatial relationships in the facial structure. Here, CNN takes a small patch from the frame equal to the kernel size. This operation enables CNN to detect emotion-specific facial configurations such as frowns, smiles, and eyebrow contractions.

Activation Function (ReLU): After convolution, the Rectified Linear Unit (ReLU) activation function is applied because it introduces non-linearity, avoids the vanishing gradient problem, and improves learning speed.

In Eq. (3), the output remains unchanged if x is positive, and it becomes 0 if x is negative; if x is 0, the output is 0

$$f(x) = \max(0, x) \quad (3)$$

Max Pooling: It is applied after convolutional blocks to minimize dimensionality. During spatial down-sampling, it helps the model by making the data smaller and faster to process.

$$Y = \max(X_{region}) \quad (4)$$

In Eq. (4), Xregion is a small block taken from the feature map 2×2, max () is the largest number inside the block, and Y is the output value after pooling for that block.

Fully Connected Layer: After convolution and pooling, CNN has already learned about these features, which are eye shape changes, mouth curvature, eyebrow movement, and facial muscle patterns. After making the final decision, a fully connected layer is used.

$$z = W^T x + b \quad (5)$$

In Eq. (5), x contains information like eye patterns and mouth curves, and W contains learnable values, meaning which feature is important for which emotion. WT means in this, it performs the Transpose of weights, and b is the bias, which is used to adjust the output by adding extra values. Z is the output of a denser layer.

Softmax Output Layer: Obtained Z values are like logits, like Z1, Z2, Z3… these values can be negative, positive, and vary greatly. To convert these logits into probabilities, softmax is used, and Eq. (6) represents it.

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \quad (6)$$

- $z_i \rightarrow$ score (logit) for class i
- $e^{z_i} \rightarrow$ exponential of that score (makes all values positive)
- $C \rightarrow$ total number of classes (in your case, 7 emotions)
- $\sum_{j=1}^{C} e^{z_j} \rightarrow$ sum of exponentials of all class scores
- $P(y_i) \rightarrow$ probability of class i(between 0 and 1)

Loss Function (Categorical Cross-Entropy): To measure prediction error for multi-class classification, penalise confident wrong predictions, guide weight updates, and the loss Function is used.

$$L = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \quad (7)$$

- $L \rightarrow$ loss value (error)
- $C \rightarrow$ total number of emotion classes (6)

- $y_i \rightarrow$ true label (one-hot encoded: 0 or 1)
- $\hat{y}_i \rightarrow$ predicted probability from Softmax (0 to 1)
- $\log(\hat{y}_i) \rightarrow$ log of predicted probability

Eq (7) This loss function penalises incorrect predictions and guides weight updates so that the model improves classification performance.

3.3. Text Emotion Recognition Using Bi-LSTM

Text Preprocessing: The Bi-LSTM model accepts text as input and first undergoes text preprocessing.

Steps include:

Tokenisation is nothing but separating the text into words, where all the words are converted into lowercase. Due to the different lengths of each sentence, we use padding to correct the lengths equally.

Then each word is converted into an integer corresponding to the index. The embedding layer is mapped to a dense vector $E(w_i) \in \mathbb{R}^d$. Here, embedding is used for capturing semantic meaning to preserve contextual similarities.

Word Embedding Layer: This layer converts each word ID into a dense vector representation of that word, which helps the model learn the semantic and emotional associations of words. During training, the embedding vectors are updated using backpropagation. The embedding vectors are then passed to the Bi-LSTM layer for contextualised emotion learning.

Long Short-Term Memory: LSTM handles long-term dependencies using gates. LSTM networks process sequential data by using gated mechanisms to control the information flow. The forget gate is defined as follows: $f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$ Here, h_{t-1} represents the previous hidden state, x_t indicates current input W_f, represents the weighted matrix b_f, represents the bias, and is the sigmoid function. Here, the gate shows the output values between 0 and 1, indicating the information from the previous cell state.

The Input gate combines the present word embedding along with the previous hidden state to decide the importance of the current word and how the information is stored in the memory. $i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$ Here, the words strongly express emotions such as happiness, sadness, and anger to produce higher input gate values to pass the information into the memory.

The candidate formula $\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c)$ helps in generating the candidate information from the current word to its context. This cell state represents the emotional and semantic information of each word. Here, the values of candidates are ranged between -1 and 1 to maintain stable memory.

Cell State Update is computed as ($C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$). where the C_{t-1} is the previous cell state, which contains information about the past context, and the C_t is determined as the updated cell state forget gate decides how much context from the previous state is retained. The current state is represented by i_t where the new information is added to the existing word by the candidate state $\tilde{C}_t$. Hence, this update allows LSTM to store the long-term emotional context throughout the sentences.

The output gate is demonstrated as a formula. $o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$ and $h_t = o_t \cdot \tanh(C_t)$. Here, o_t it is determined by how much information is exposed from the previous cell state to the next layer h_t. The hidden state states the processed meaning of the current word in the context of surrounding words. As in the text, emotion recognition, hidden states carry over the emotional cues forward and are later used by the classifier to predict the final label of emotion.

3.3.1. Bidirectional LSTM (Bi-LSTM)

In this process, Bi-LSTM helps in the processing of input in two directions. The forward LSTM examines the sequence from left to right; meanwhile, the backward LSTM operates from right to left. Hidden states are then joined over $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$. Here, $\overrightarrow{h_t}$ shows the information from previous words and $\overleftarrow{h_t}$ determines the information from upcoming words. Connecting both hidden states enables the models to understand the emotions accurately, particularly when the sentiment depends on the future context.

3.3.2. SoftMax Output Layer

The last hidden states obtained from the Bi-LSTM are passed through a dense layer, and then the SoftMax function converts raw scores into probabilities.

3.4. Speech Emotion Recognition Using LSTM

Speech is a major modality for emotional expression, since human emotions are conveyed by pitch, intensity, rhythm, and speaking rate. In the proposed multimodal emotion recognition system, speech modality is intended to take advantage of certain properties in the audio signal. Because speech signals are essentially sequential, a LSTM network is utilised as a model to learn the emotional patterns in the signals, intended to exploit these characteristics from the audio signal. Since the speech signal is inherently a sequential process, a LSTM network is used to learn the emotional patterns in the signal.

Audio Preprocessing: The raw speech signals are first preprocessed to ensure uniformity and robustness for all audio samples. To achieve equal temporal resolution, the audio signals are resampled at a consistent sampling rate after being transformed to mono to remove channel disparities. The resampled audio signals are then segmented into short overlapping frames because the emotional content of speech is better represented in localised temporal windows.

The preprocessing steps eliminate noise and normalise the signal characteristics to prepare the audio data for robust feature extraction.

Feature Extraction Using MFCC: To compactly and meaningfully express the speech signal, "Mel-Frequency Cepstral Coefficients (MFCCs)" are extracted from preprocessed audio frames. MFCCs are calculated by executing a "Fast Fourier Transform (FFT)" on frames, followed by Mel-filtering, log compression, and Discrete Cosine Transform (DCT), as

$$\text{MFCC} = \text{DCT}\big(\log\big(\text{Mel}(|\text{FFT(x)}|)\big)\big) \quad (8)$$

MFCCs are preferred in speech emotion recognition tasks because they can approximate human auditory perception and well represent the spectral properties associated with emotional expression, pitch, and energy. The effectiveness of MFCCs in previous speech and multimodal emotion recognition tasks makes them a good candidate for this research.

Temporal Modelling Using LSTM: The obtained MFCC features are a sequence that represents the temporal dynamics of speech. To represent these sequential patterns, an LSTM network is employed because of its capability to handle long-term dependencies without being affected by the vanishing gradient problem.

The LSTM network can learn the emotionally significant temporal dynamics of speech through gated memory mechanisms, such as pitch modulation, energy, and speech rate variations. The input gate, forget gate, and output gate are the fundamental equations of LSTM gates that regulate the information flow over time.

Speech Emotion Classification: A high-level summary of the speech signal over time is stored in the LSTM's final hidden state. A Softmax function and a dense layer are applied to this summary in order to generate a probability distribution over the predetermined classifications of emotion. An emotion class corresponding to the highest probability is selected.

Model Training and Optimisation: The Adam optimisation technique is utilised to train the speech emotion recognition model, which leads to an adaptive learning rate for each parameter and consistent convergence during training. The parameter update rule for Adam is as follows:

$$\theta_{t+1} = \theta_t - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} \quad (9)$$

In Eq. (9), $\hat{m}_t$ denote the bias-corrected first and $\hat{v}_t$ denote the bias-corrected second moment estimates, where is the learning rate. Adam is selected due to its fast convergence, robustness to noisy gradients, and effectiveness in training deep recurrent networks.

3.5. Probability-Based Late Fusion

In the proposed multimodal emotion recognition system, the Probability-Based Late Fusion approach combines the outputs of three independent models: CNN (image modality), Bi-LSTM (text modality), and LSTM (audio modality). Each model generates a probability distribution over predefined emotion classes. For a given emotion class, each modality predicts a probability value representing its confidence that the input belongs to that emotion. These probabilities are combined using the weighted fusion formula:

$$P_{final}(e) = w_1 P_{image}(e) + w_2 P_{text}(e) + w_3 P_{audio}(e) \quad (10)$$

Eq. (10) Here, $P_{image}(e)$, $P_{text}(e)$, and $P_{audio}(e)$ are the predicted probabilities from the respective modalities, w_1, w_2, w_3 are the modality weights such that.

$$w_1 + w_2 + w_3 = 1 \quad (11)$$

The Eq. (11) fusion formula is applied to each emotion class to compute a combined probability using the outputs from image, text, and audio models. This produces final probability values for all emotions.

The system then compares these values and selects the emotion with the highest combined probability as the final predicted emotion.

4. Results and Discussion

The proposed MERMHS model was evaluated on the dataset CMU-MOSEI, and this dataset contains multimodal data, which includes facial expressions data, Speech signals, and Text transcripts.

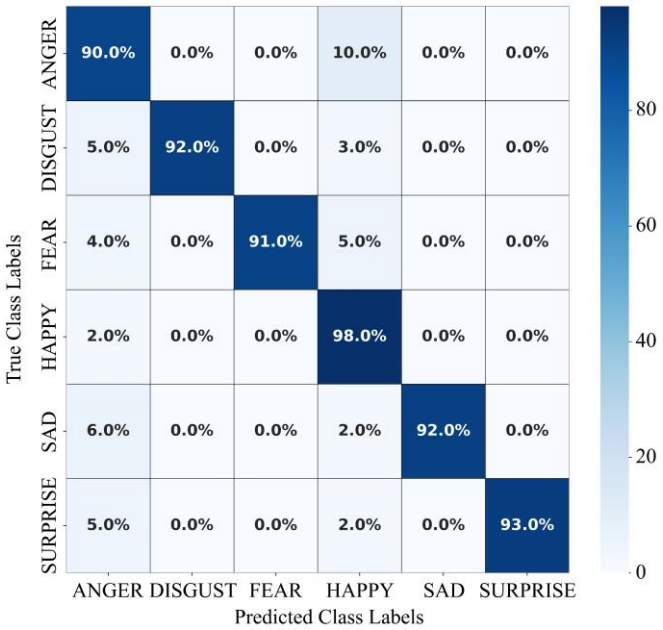

Fig. 2 Confusion matrix of the proposed MERMHS on the CMU-MOSEI dataset

The confusion matrix of the proposed MERMHS on the CMU-MOSEI dataset is presented in Figure 2. From these confusion matrices, True Negative (T_{NE}), False Positive (F_{PO}), False Negative (F_{NE}), and True Positive (T_{PO}) values are obtained at the entity level. For each emotion class, performance metrics such as accuracy, precision, recall, and F1-score are calculated individually for each emotion class using this evaluation matrix.

The total number of samples is calculated as Eq. (12)

$$NUM = T_{NE} + F_{PO} + F_{NE} + T_{PO} \quad (12)$$

Here
NUM – Total number of samples.
T_{NE} – Samples belonging to a particular emotion class that are correctly predicted as the same class.
F_{PO} – Samples belonging to a particular emotion class that are incorrectly predicted as one of the other emotion classes.
F_{NE} – Samples belonging to other emotion classes that are incorrectly predicted as the target emotion class.
T_{PO} – Samples that neither belong to the target emotion class nor are predicted as that class.

Accuracy(A) is the percentage of total samples properly identified (both true positives and true negatives).

$$A = \frac{T_{PO} + T_{NE}}{NUM} \quad (13)$$

Precision(P): How many of the samples predicted by the model turned out to be positive?

$$P = \frac{T_{PO}}{T_{PO} + F_{PO}} \quad (14)$$

Recall(R): How many of the positive samples (True Positive + False Negative) did the model correctly identify?

$$R = \frac{T_{PO}}{T_{PO} + F_{NE}} \quad (15)$$

F1-Score: It is a single metric that counts both precision and recall. It is functional when the class distribution is asymmetrical.

$$f1 - score = 2 * \frac{(P*R)}{P+R} \quad (16)$$

Specificity (S): How many negative samples did the model accurately identify? It measures the model's ability to recognize the negative class exactly.

$$S = \frac{T_{NE}}{T_{NE} + F_{PO}} \quad (17)$$

The False Discovery Rate (FDR): What percentage of the samples that the model predicted as Positive were Negative (i.e., wrong predictions)?

$$FDR = \frac{F_{PO}}{F_{PO} + T_{PO}} \quad (18)$$

After calculating the evaluation metrics separately for every emotion class (happy, sad, angry, fearful, disgusted, surprised), the overall performance is reported as the average of these class-wise results, which are presented in Table 1.

Table 1. Performance of existing CNN, LSTM, Bi-LSTM, and the proposed multi-model on the CMU-MOSEI dataset

Features	CNN	LSTM	Bi-LSTM	MERMHS (proposed)
Accuracy (%)	77.96	78.23	80.92	92.7
Precision (%)	76.57	77.92	81.01	93.70
Recall (%)	75.12	77.56	80.23	92.67
F1-score	74.53	75.89	79.82	93.10

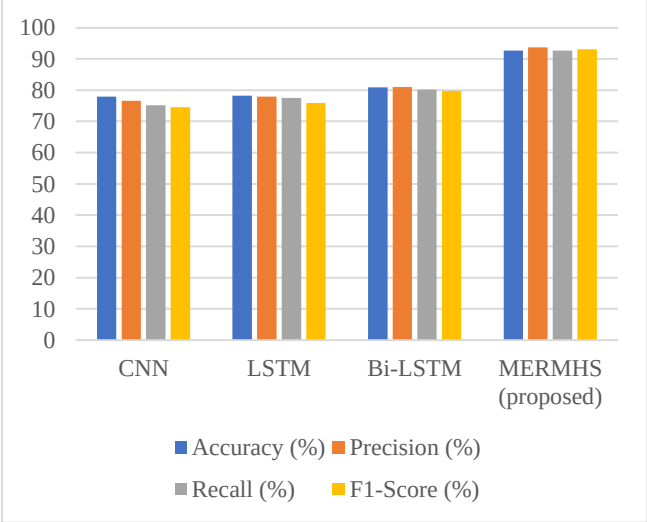

Fig. 3 Comparison of CNN, LSTM, Bi-LSTM, and MERMHS

Table 1 and Figure 3 provide a detailed overview of the accuracy and performance metrics corresponding to each model class. The proposed MERMHS model performs well compared to single models. MERMHS achieves 92.7% accuracy, which is 14.74% higher than CNN, 14.47% higher than LSTM, and 11.78% higher than Bi-LSTM. The Precision and Recall values are well-balanced, and the F1-score of 93.10% confirms stable and consistent classification across all emotion classes. This is all because of the effectiveness of the probability-based late fusion strategy. Robust Feature Extraction and Improved Preprocessing Strategies are the primary reasons for the proposed MERMHS model's excellent and well-balanced performance across all metrics (Accuracy, Precision, Recall, and F1 Score).

The proposed MERMHS architecture is clearly the most robust and dependable option for this classification problem.

Our proposed MERMHS demonstrated optimised overall accuracy, and it shows that the proposed MERMHS gives better accuracy than the existing methods.

5. Conclusion

The primary goal was to develop a predictive MERMHS AI system for mental health by integrating emotional detection. The proposed work is best suited for an early detection approach, either using text, face expressions, or audio. Our proposed MERMHS model proved highly effective for early detection, achieving an overall accuracy of 92.7%. This architecture demonstrated a robust advantage, providing the best accuracy, precision, recall, and F1 score improvement over the existing model, positioning it as the superior choice for high-precision emotion recognition for mental health assignments.

Conflicts of Interest

The authors declare no conflict of interest.

Funding Statement

This research did not receive any specific funding.

Acknowledgements

The authors would like to express a very great appreciation to the co-authors of this manuscript for their valuable and constructive suggestions during the planning and development of this research work.

References

[1] Qianhe Ouyang, "Speech Emotion Detection based on MFCC and CNN-LSTM Architecture," *Applied and Computational Engineering*, vol. 5, pp. 243-249, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Shamin Bin Habib Avro, Taieba Taher, and Nursadul Mamun, "EmoTech: A Multi-Modal Speech Emotion Recognition Using Multi-Source Low-Level Information with Hybrid Recurrent Network," *2024 IEEE International Conference on Signal Processing, Information, Communication and Systems*, Khulna, Bangladesh, pp. 1-5, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[3] Guowei Zhong, et al., "Calibrating Multimodal Consensus for Emotion Recognition," *arXiv preprint*, vol. 14, no. 8, pp. 1-13, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[4] Shuo Zhang et al., "Multimodal Mixture of Low-Rank Experts for Sentiment Analysis and Emotion Recognition," *2025 IEEE International Conference on Multimedia and Expo (ICME)*, Nantes, France, pp. 1-6, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[5] Chengyan Wu et al., "Multimodal Emotion Recognition in Conversations: A Survey of Methods, Trends, Challenges and Prospects," *arXiv preprint*, pp. 1-18, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[6] Prashant Kumar Nag, Amit Bhagat, and R. Vishnu Priya, "Expanding AI's Role in Healthcare Applications: A Systematic Review of Emotional and Cognitive Analysis Techniques," *IEEE Access*, vol. 13, pp. 69129-69160, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[7] Puneet Kumar et al., "VISTANet: VIsual Spoken Textual Additive Net for Interpretable Multimodal Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 16, no. 4, pp. 2881-2893, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[8] Yehun Song, and Sunyoung Cho, "Leveraging CLIP Encoder for Multimodal Emotion Recognition," *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson*, AZ, USA, pp. 6115-6124, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[9] Konstantinos Mountzouris, Isidoros Perikos, and Ioannis Hatzilygeroudis, "Speech Emotion Recognition Using Convolutional Neural Networks with Attention Mechanism," *Electronics*, vol. 12, no. 20, pp. 1-31, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Utsav Poudel et al., "AI in Mental Health: A Review of Technological Advancements and Ethical Issues in Psychiatry," *Issues in Mental Health Nursing*, vol. 46, no. 7, pp. 693-701, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[11] Feiyang Chen et al., "Complementary Fusion of Multi-Features and Multi-Modalities in Sentiment Analysis," *arXiv preprint*, pp. 1-9, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[12] Sanmay Kotkar, "Real-Time Emotion Recognition with CNN and LSTM," *Preprints*, pp. 1-8, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[13] Fazliddin Makhmudov, Alpamis Kultimuratov, and Young-Im Cho, "Enhancing Multimodal Emotion Recognition through Attention Mechanisms in BERT and CNN Architectures," *Applied Sciences*, vol. 14, no. 10, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] Dilnoza Mamieva et al., "Multimodal Emotion Detection via Attention-Based Fusion of Extracted Facial and Speech Features," *Sensors*, vol. 23, no. 12, pp. 1-19, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Trisha Mittal et al., "M3ER: Multiplicative Multimodal Emotion Recognition Using Facial, Textual, and Speech Cues," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 2, pp. 1359-1367, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[16] Zhijing Xu, and Yang Gao, "Research on Cross-Modal Emotion Recognition based on Multi-Layer Semantic Fusion," *Mathematical Biosciences and Engineering*, vol. 21, no. 2, pp. 2488-2514, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[17] Ngumimi Karen Iyortsuun et al., "A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis," *Healthcare*, vol. 11, no. 3, pp. 1-27, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[18] Masab A. Mansoor, and Kashif H. Ansari, "Early Detection of Mental Health Crises through Artificial-Intelligence-Powered Social Media Analysis: A Prospective Observational Study," *Journal of Personalized Medicine*, vol. 14, no. 9, pp. 1-11, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[19] Jinghui Qin et al., "Mental-Perceiver: Audio-Textual Multi-Modal Learning for Estimating Mental Disorders," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 23, pp. 1-9, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[20] Nafiseh Ghaffar Nia, Erkan Kaplanoglu, and Ahad Nasab, "Evaluation of Artificial Intelligence Techniques in Disease Diagnosis and Prediction," *Discover Artificial Intelligence*, vol. 3, pp. 1-14, 2023. [CrossRef] [Google Scholar] [Publisher Link]