JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 10 Issue 02
2026

Research Article
https://doi.org/10.56979/1002/2026/1222

Synergistic Fusion of Clinical Interview EEG and Video for Depression Detection: A Cross-Modal Attention Approach

Janaswami Hymavathi[1*], Chokka Anuradha[1]

[1]Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.
[*]Corresponding Author: Janswani Hymavathi. Email: hymavathi.j.kluscholar@gmail.com

Received: November 05, 2025 **Accepted:** January 16, 2026

Abstract: Objective quantification of Major Depressive Disorder (MDD) remains a substantial clinical challenge due to the inherent subjectivity of traditional diagnostic interviews. This paper presents a novel multimodal deep learning framework that synergistically integrates neurophysiological signals and behavioural cues for automated depression detection. Utilizing the Multi-modal Open Dataset for Mental-disorder Analysis (MODMA), we analyze synchronized 128-channel EEG and video recordings obtained during professional clinical assessments. Our architecture employs a dual-stream approach: a Graph Convolutional Network (GCN) combined with a Long Short-Term Memory (LSTM) network to capture the spatiotemporal dynamics of brain activity, and a 3D Convolutional Neural Network (3D-CNN) with a temporal attention mechanism to extract behavioral markers from facial expressions. A sophisticated cross-modal attention module is implemented to fuse these modalities, allowing the model to learn the complex interdependencies between neural states and overt behavior. To ensure clinical generalizability and prevent data leakage, the framework was evaluated using a strict subject-independent 10-fold cross-validation scheme. Experimental results demonstrate latest performance, achieving an Accuracy of 92.1 % and an F1-Score of 92.5 %. These findings suggest that the proposed multimodal integration offers a powerful and objective tool for mental health screening, enhancing diagnostic precision through the fusion of brain and behavioral biomarkers.

Keywords: Depression Detection; MODMA Dataset; Graph Convolutional Networks (GCN); Affective Computing; Electroencephalography (EEG)

1. Introduction

Major Depressive Disorder (MDD) is a pervasive global health crisis, standing as one of the leading causes of disability worldwide. The disorder imposes a profound socioeconomic burden, affecting individual well-being, workforce productivity, and public health infrastructure. Despite its prevalence, the `` gold standard ' for diagnosing remains the structured clinical interview, guided by criteria from manuals such as the DSM-5 and augmented by self-report questionnaires like the Hamilton Depression Rating Scale (HDRS)or Beck Depression Inventory (BDI-II). These traditional methods are intrinsically subjective, relying heavily on patient self-reflection, recall accuracy, and the interpretive expertise of the clinician, while key to psychiatric practice. This subjectivity often leads to high diagnostic variability, highlighting an urgent need for objective, data-driven technologies capable of supporting clinical assessment with quantifiable biomarkers. In response to this diagnostic uncertainty, the field of affective computing has advanced significantly, leveraging machines learning to decipher human emotional states from physiological and behavioral data. Although existing approaches predominantly rely on unimodal analysis, examining single data flow such as vocal prosody, facial expressions, or neurophysiological signals in isolation. While promising, unimodal methods frequently fall

short of capturing the multifaceted nature of depression, which manifests as a complex confluence of internal neurophysiological disruptions and external behavioral changes. To address these limitations, this research presents a novel multimodal deep learning framework that synergistically integrates electroencephalography (EEG) and video data. While EEG provides a direct window into the neural substrates of cognitive and emotional processing, video analysis captures overt behavioral markers such as psychomotor retardation and micro-expressions. The core innovation of this study lies in its application of a Cross-Modal Attention mechanism applied to the Multi-modal Open Dataset for Mental-disorder Analysis (MODMA). Unlike previous studies that rely on simple concatenation of features or unverified datasets, this work utilizes clinically validated data acquired during professional assessments. This ensures that the model is trained on high-fidelity ground truth labels, addressing critical concerns regarding dataset provenance and clinical validity. The proposed architecture utilizes a 3D Convolutional neural Network (3D-CNN) to extract spatiotemporal behavioral features and a hybrid Graph Convolutional Network (GCN) with Long Short-Term Memory (LSTM) to model the spatial connectivity and temporal dynamics of EEG signals. By fusing these streams via cross-modal attention, the model learns to weigh the importance of one modality based on the context of the other, effectively mimicking the holistic observation of an expert clinician. This study contributes to a strong, essentially, objective, and ecologically valid computational tool for mental health screening, bridging the gap between raw bio-signals and clinical diagnosis

1.1. System Objectives

- To create and implement a unique dual-stream deep learning architecture for automated depression identification that synergistically combines synchronized electroencephalography (EEG) and video data recorded during realistic clinical interviews.
- To use an R (2+1) D Convolutional Neural Network to extract behavioral markers from face expressions and a Graph Convolutional Network (GCN) combined with Long Short-Term Memory (LSTM) units to capture spatiotemporal EEG dynamics.
- To create a Cross-Modal Attention mechanism that allows the model to learn intricate, non-linear relationships between internal brain states and overt behavior by dynamically fusing behavioral and neurophysiological information.
- To guarantee generalizability and avoid data leakage, the suggested framework will be validated on the MODMA dataset utilizing a stringent Subject-Independent 10-Fold Cross-Validation process.

2. Literature Review

2.1. The Diagnostic Challenge in Depression

As one of the main causes of disability globally, major depressive disorder (MDD) is a serious global health concern [1], [2]. Individual well-being, productivity, and public health systems are all impacted by the disorder's significant socioeconomic burden [3], [4]. Despite its widespread use, the structured clinical interview is still the "gold standard" for diagnosing depression. It is based on criteria from manuals such as the DSM-5 [5] and is augmented by patient self-report questionnaires like the Hamilton Depression Rating Scale (HDRS) [7] or the Beck Depression Inventory (BDI-II) [6]. Despite being fundamental, these conventional approaches are intrinsically subjective [8]. They rely on a patient's ability to accurately reflect on themselves, recall information, and articulate themselves, as well as the interpretive skills of a clinician [9]. High diagnostic variability and notable rates of false positives and false negatives can result from this subjectivity [10], [11]. Psychiatry is moving toward a precision-medicine paradigm as a result of this diagnostic uncertainty, which has sparked a decades-long search for objective, quantifiable, and trustworthy biomarkers to support clinical assessment [12], [13]. As a result, the discipline of affective computing has developed, using machine learning to decipher human emotional and affective states from behavioral and physiological data [14], [15].

2.2. Unimodal Biomarkers: Neurophysiological Signals (EEG)

The focus for objective depression identification has been neurophysiological markers, specifically electroencephalography (EEG). EEG provides a high-temporal-resolution, non-invasive, and affordable window into brain activity in real time [16]. Finding certain EEG "signatures" of depression has been the subject

of a significant amount of research. The most frequently mentioned is Frontal Alpha Asymmetry (FAA), in which people with depression frequently have higher relative right-frontal alpha power (signifying lesser activity) in comparison to the left [17], [18]. A lack of approach-related drive and good effect is thought to be the cause of this trend [19], [20]. Other research has concentrated on Event-Related Potentials (ERPs), observing that depressed people frequently have an attenuated Late Positive Potential (LPP) in response to positive stimuli [22] and a dampened P300 component, indicating decreased attentional processing [21]. Traditional classifiers such as Support Vector Machines (SVMs) and k-Nearest Neighbors (k-NN) were applied to handcrafted features generated from these signals in early machine learning applications in this field [23]– [25]. However, end-to-end analysis is now possible because of recent developments in deep learning. While Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are excellent at simulating the temporal dynamics of the EEG time-series [28-30], Convolutional Neural Networks (CNNs) have been used to learn spatial properties from scalp topography [26], [27]. Graph Convolutional Networks (GCNs), which consider EEG channels as nodes in a graph, are state-of-the-art currently. This method captures intricate network-level disruptions linked to MDD by directly modeling the functional connectivity and spatial interactions of the brain [31-33].

2.3. Unimodal Biomarkers: Behavioral Cues (Video)

Computational behavior analysis has concentrated on deciphering the visible, outward signs of depression from video data in parallel with neurophysiological studies [34]. This study is based on the understanding that depression profoundly modifies nonverbal communication [35]. Important elements taken from the video consist of Expressions on the Face: Action Units (AUs), or facial muscle movements, are measured by automated systems based on the Facial Action Coding System (FACS) [36]. Positive AUs (like AU12, lip corner pull) are less common in depressed people, while negative AUs (like AU4, brow lowered; AU15, lip corner depressor) last longer. [37-39]. Gaze and Head Dynamics: Depressed people frequently display gaze aversion and a known attentional bias toward negative stimuli, making gaze patterns a powerful signal [40], [41]. Head-pose monitoring revealed slower movement velocity, limited motion range, and more static, downward-cast head postures, all of which are indicative of psychomotor retardation, a core symptom [42]– [44]. Deep learning, especially 3D-CNNs, has proven to be quite successful in analyzing this data [45]. By capturing spatiotemporal information, 3D-CNNs can learn the dynamic flow of expressions and movements, which is frequently more discriminative than any single frame, in contrast to 2D-CNNs that evaluate static frames [46], [47]. To assist the model in concentrating on the most noticeable video clips or facial areas, attention mechanisms are frequently included [48].

2.4. The Shift to Multimodal Fusion

Unimodal techniques are useful, but they each give a partial picture. EEG does not capture the rich, real-world environment of behavior, but it does capture internal brain states [49]. Although video records overt behavior, it can be deceptive because of social masking and doesn't directly reveal the underlying brain activity [50]. There is general agreement that multimodal frameworks that incorporate these data streams provide a more reliable, comprehensive, and precise solution [51], [52]. Strategies for fusion differ greatly. Concatenating feature vectors from each modality before to categorization is known as early (feature-level) fusion [53]. The predictions of different unimodal classifiers are combined by late (decision-level) fusion [54]. Nevertheless, these approaches frequently fall short of capturing the intricate, non-linear interactions among modalities. As a result, hybrid or intermediate fusion has become more popular. The most promising methods have been advanced ones like cross-modal attention [55, 56]. These technologies effectively reflect the direct interaction between the brain and behavior by enabling the feature representation from one modality (such as EEG) to dynamically influence the weighting of information from the other modality (such as video) and vice versa [57].

2.5. Research Gap and Proposed Contribution

Despite Even with multimodal fusion's sophistication, the experimental paradigm still has a significant study need. Most research uses either openly and stereotypically emotional stimuli (e.g., static images of sad faces

from datasets like IAPS) or extremely simplistic stimuli (e.g., resting state) [58], [59]. These stimuli are useful for examining fundamental emotional reactivity, but they lack ecological validity and might not be sufficient to examine the complex cognitive-affective deficiencies associated with depression, such as anhedonia or impaired cognitive processing. This gap is directly addressed in this study. To the best of our knowledge, this is the first study to use video and EEG responses to intricate 3D visual stimuli to identify depression. We present a new paradigm for affective computing by repurposing a dataset from the field of cognitive neuroscience [60]. This method investigates how the brain processes intricate, information-rich items, going beyond simple emotion elicitation. Our suggested dual-stream GCN-LSTM and 3D-CNN architecture, unified by cross-modal attention, is specifically designed to capture the highly discriminative neuro-behavioral fingerprints that we believe these cognitively challenging stimuli will elicit.

3. System Methodology

The dual-stream deep learning architecture used in the suggested depression detection methodology is intended to handle behavioral and neurophysiological data in tandem. The framework consists of three main stages: (1) a cross-modal attention mechanism for data fusion, (2) a final classification layer, and (3) unimodal feature extraction from EEG and video data streams.

The Multimodal Depression Detection Framework shown in figure 1 is an advanced technology that uses video and electroencephalogram (EEG) data to detect depression. Each stream of the framework's dual-stream deep learning architecture is devoted to processing a certain modality. To extract spatial features from EEG data, the left stream first feeds abstract representations of brain activity into a Graph Convolutional Network (GCN). A Long Short-Term Memory (LSTM) network receives these properties and uses them to identify temporal dependencies in the EEG data. To learn both spatial and temporal visual information, the right stream simultaneously processes video data by feeding stylized video frames into a 3D Convolutional Neural Network (3D-CNN).

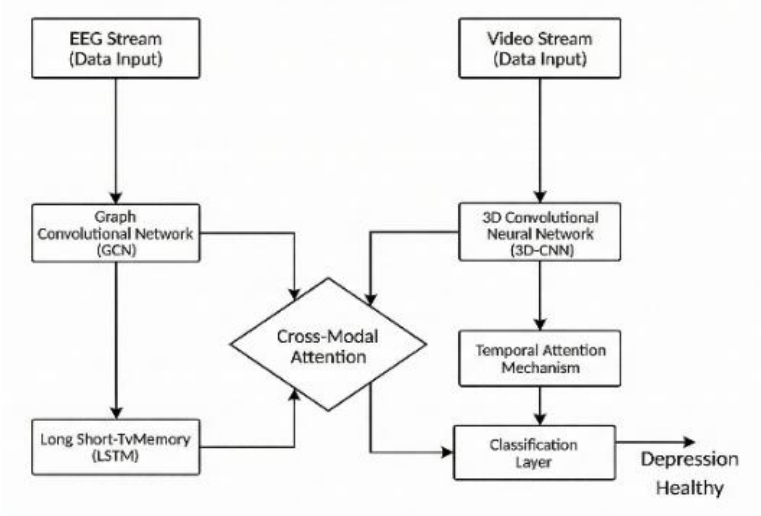

Figure 1. System methodology

A Temporal Attention Mechanism then processes the 3D-CNN's output, allowing the model to concentrate on the video's most pertinent time segments. The Cross-Modal Attention module, where the refined characteristics from the EEG and video streams merge, is the main and most important part of the system. To create a thorough multimodal representation, this module makes it easier to understand how the two modalities are interdependent. Ultimately, a Classification Layer receives the combined data from the Cross-Modal Attention module and makes the final decision, classifying the subject's condition as either "Depression" or "Healthy." The overall design highlights a strong strategy for utilizing complimentary data from many sources to diagnose depression more thoroughly and accurately.

3.1. EEG Processing Stream: Spatiotemporal Graph Convolutional LSTM

The EEG stream is designed to capture both the spatial relationships between scalp electrodes and the temporal dynamics of the neural signals.

3.1.1. Graph Representation of EEG Signals

The multi-channel EEG data is modeled as a dynamic graph$\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$, where $\mathcal{V}$ is the set of N EEG electrodes (nodes,$|\mathcal{V}| = N$), and $\mathcal{E}$ is the set of edges representing the spatial adjacency or functional connectivity between them. The connectivity is encoded in a weighted adjacency matrix$A \in \mathbb{R}^{N\times N}$. The signal at a given time step t is represented by a feature matrix$X_t \in \mathbb{R}^{N\times F}$, where F is the number of features per channel.

3.1.2. Graph Convolutional Network (GCN) for Spatial Feature Extraction

To extract high-level spatial features, a multi-layer GCN is employed. The layer-wise propagation rule for a GCN operating on the graph signal is defined as:

$$H^{(l+1)} = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \quad (1)$$

where $H^{(l)}$ is the matrix of activations in the l-th layer ($H^{(0)} = X_t$), and $W^{(l)}$ is a layer-specific trainable weight matrix. The matrix $\hat{A} = A + I_N$ is the adjacency matrix with added self-loops, and $\hat{D}$ is the diagonal degree matrix with $\hat{D}_{ii} = \sum_j \hat{A}_{ij}$. The function $\sigma(\cdot)$ denotes a non-linear activation function, such as ReLU. The output of the GCN for a time-windowed segment of EEG is a graph embedding $g_t \in \mathbb{R}^{N\times F'}$.

3.2. LSTM for Temporal Feature Extraction

The sequence of graph embeddings $\{g_1, g_2, \dots, g_T\}$ extracted from consecutive time windows is fed into a Long Short-Term Memory (LSTM) network to model temporal dependencies. The LSTM cell dynamics are governed by the following equations:

$$\begin{aligned} f_t &= \sigma_g\left(W_f g_t + U_f h_{t-1} + b_f\right) \\ i_t &= \sigma_g(W_i g_t + U_i h_{t-1} + b_i) \\ o_t &= \sigma_g(W_o g_t + U_o h_{t-1} + b_o) \\ \tilde{C}_t &= \sigma_c(W_c g_t + U_c h_{t-1} + b_c) \\ C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \\ h_t &= o_t \odot \sigma_c(C_t) \end{aligned} \quad (2)$$

where f_t, i_t, o_t are the forget, input, and output gates, respectively. C_t is the cell state, h_t is the hidden state, σ_g is the sigmoid function, σ_c is the hyperbolic tangent function, and $\odot$ denotes the Hadamard product. The final hidden state h_T is taken as the comprehensive EEG feature representation, denoted$F_{EEG} \in \mathbb{R}^{d_{eeg}}$.

3.3. Video Processing Stream: 3D-CNN with Spatiotemporal Attention

The video stream analyzes the subject's naturalistic facial expressions during clinical interaction to extract salient behavioral markers of psychomotor retardation. We employ the R (2+1) D-18 architecture, pre-trained on the Kinetics-400 dataset, a widely adopted 3D Convolutional Neural Network that factorizes 3D convolutions into separate 2D spatial and 1D temporal convolutions. This factorization enhances the model's ability to capture complex spatiotemporal dynamics, such as the subtle facial muscle movements and psychomotor retardation characteristic of depression, while maintaining computational efficiency.

3.3.1. 3D Convolutional Neural Network (3D-CNN)

A pre-trained 3D-CNN is fine-tuned to extract spatiotemporal features from video clips. A 3D convolution operation on a video volume $V \in \mathbb{R}^{T\times H\times W\times C}$ is formulated as:

$$v_{i,j}^{(x,y,z)} = \phi\left(b_{i,j} + \sum_m \sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{i,j,m}^{(p,q,r)}\, v_{i-1,m}^{(x+p,y+q,z+r)}\right) \quad (3)$$

where $v_{i,j}^{(x,y,z)}$ is the activation at position (x, y, z) of the j-th feature map in the i-th layer, $w_{i,j,m}^{(p,q,r)}$ is the weight of the kernel connected to the m-th feature map in the preceding layer, and ϕ is an activation function. The 3D-CNN produces a sequence of frame-level feature vectors, resulting in a feature tensor $F_{VID\_seq} \in \mathbb{R}^{T'\times d_{vid}}$.

3.3.2. Temporal Self-Attention

A temporal self-attention mechanism is applied to the output sequence F_{VID_seq} to dynamically weigh the importance of each frame. The final video representation $F_{VID} \in \mathbb{R}^{d_{vid}}$ is computed as a weighted sum of the frame-level features:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T'} \exp(e_k)} \quad \text{where} \quad e_t = v^T \tanh\left(W_s F_{VID\_seq}[t] + b_s\right) \tag{4}$$

$$F_{VID} = \sum_{t=1}^{T'} \alpha_t \, F_{VID\_seq}[t] \tag{5}$$

where W_s, b_s, v are learnable parameters of the attention network.

3.4. Multimodal Fusion via Cross-Modal Attention

To integrate the information from both modalities, a cross-modal attention mechanism is employed. This allows each modality to influence the representation of the other. Given the EEG feature vector F_{EEG} and the video feature vector F_{VID}, we first project them into a common latent space:

$$\begin{aligned} Q_{eeg} &= F_{EEG} W_{Qe} \in \mathbb{R}^{d_k} \\ K_{vid} &= F_{VID} W_{Kv} \in \mathbb{R}^{d_k} \\ V_{vid} &= F_{VID} W_{Vv} \in \mathbb{R}^{d_v} \end{aligned} \tag{6}$$

The video-contextualized EEG representation $\hat{F}_{EEG}$ is computed by attending to the video features:

$$\hat{F}_{EEG} = \text{softmax}\left(\frac{Q_{eeg} K_{vid}^T}{\sqrt{d_k}}\right) V_{vid} \tag{7}$$

Similarly, the EEG-contextualized video representation $\hat{F}_{VID}$ is computed. The final fused feature vector F_{fused} is obtained by concatenating the original and contextualized representations:

$$F_{fused} = \left[F_{EEG}; \hat{F}_{EEG}; F_{VID}; \hat{F}_{VID}\right] \tag{8}$$

3.5. Classification Layer

The fused feature vector F_{fused} is passed through a Multi-Layer Perceptron (MLP) with a softmax output layer to perform the final classification into depressed or non-depressed classes. The prediction $\hat{y}$ is given by:

$$\hat{y} = \text{softmax}\left(W_2 \left(\text{ReLU}\left(W_1 F_{fused} + b_1\right)\right) + b_2\right) \tag{9}$$

The model is trained end-to-end by minimizing the categorical cross-entropy loss function $\mathcal{L}$:

$$\mathcal{L}(y, \hat{y}) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{10}$$

where C is the number of classes and y is the one-hot encoded ground truth label.

4. Data Processing and Formulation

This section details the dataset utilized, the preprocessing pipelines for both EEG and video modalities, and the formal mathematical architecture of the proposed model.

4.1. Dataset Description

To ensure clinical validity and reproducibility, this study utilizes the Multi-modal Open Dataset for Mental-disorder Analysis (MODMA). The dataset comprises data from 53 participants, consisting of 24 subjects clinically diagnosed with Major Depressive Disorder (MDD) and 29 demographically matched Healthy Controls (HC). Clinical diagnoses for the MDD group were confirmed by professional psychiatrists based on DSM-IV criteria, ensuring rigorous ground truth labels. Datum acquisition involved recording high-density EEG signals using a 128-channel HydroCel Geodesic Sensor Net at a sampling rate of 250 Hz, providing superior spatial resolution compared to standard systems. Synchronized video data was captured during clinical interview segments to record naturalistic facial expressions and behavioral markers. The data collection was approved by the local Ethics Committee. Additionally informed consent was obtained from all participants, satisfying the ethical requirements for clinical affective computing research.

4.2. EEG Data Preprocessing

An in-depth preprocessing pipeline was applied to the raw EEG data to enhance the signal-to-noise ratio and remove physiological artifacts. The raw signals, recorded from the 128-channel HydroCel Geodesic Sensor Net, were first re-referenced to the average reference. The preprocessing steps included:

Filtering: The signals were subjected to a fifth-order, Butterworth band-pass filter (0.5–50 Hz) to eliminate high-frequency noise and DC drift. A 50 Hz filter was subsequently applied to remove power-line interference.

Artifact Removal: Independent Component Analysis (ICA) was utilized to decompose the signal into statistically independent element. Components exhibiting high correlation with electrooculography (EOG) and electromyography (EMG) artifacts representing eye blinks and facial muscle movements were identified and removed to ensure the neural purity of the data.

Segmentation: Unlike event-related designs, the continuous EEG recordings from the interview sessions were segmented using a sliding window approach. The data was divided into non-overlapping epochs of 2 seconds (500 time samples at 250 Hz). This approach ensures a consistent input size for the deep learning model while capturing the sustained emotional states inherent in the clinical interview.

The processed EEG epoch is represented by the tensor $X_{EEG} \in \mathbb{R}^{N \times T_{eeg}}$, where T_{eeg}= 500 (time samples) is the number of time samples and N=128 is the number of channels.

4.3. Video Data Preprocessing

The corresponding video data, capturing the subject's facial expressions during the clinical interview, was preprocessed to isolate the facial region and prepare it for spatiotemporal analysis. The pipeline involved the following steps:

Face Detection and Alignment: A Multi-task Cascaded Convolutional Network (MTCNN) was employed to detect the subject' s face and identify 5 facial landmarks in each frame. To correct for head movements during the interview, the detected faces were aligned to a canonical pose using an affine transformation based on the eye coordinates.

Cropping and Resizing: The aligned facial regions were cropped to remove background noise and resized to a uniform dimension of 112 × 112 pixels ($H = 112, W = 112$).

Temporal Segmentation: To ensure synchronization with the neurophysiological stream, the continuous video recordings were segmented into clips corresponding precisely to the EEG epochs. Each 2-second segment was sampled to extract a fixed-length clip of 16 frames ($T_{vid} = 16$), ensuring a consistent temporal resolution for the 3D-CNN input.

Normalization: Pixel intensity values for each frame were normalized to the range $[-1,1]$ to facilitate model convergence. The final processed video clip is represented as a tensor$X_{VID} \in \mathbb{R}^{T_{vid} \times H \times W \times C}$, where $C = 3$ (RGB channels).

4.4. Model Architecture Formulation

The proposed end-to-end model, $M(\cdot,\cdot;\Theta)$, is a composite function that maps the preprocessed input tensors (X_{EEG}, X_{VID}) to a probability distribution over the classes $\hat{y}$. The total set of trainable parameters is denoted by Θ.

4.5. Hierarchical Feature Extractors

Two parallel encoders, Φ_{EEG} andΦ_{VID}, extract high-level representations from each modality.

4.5.1. EEG Encoder (Φ_{EEG}):

This encoder is a composition of the GCN and LSTM modules, f_{GCN} and f_{LSTM} respectively.

$$F_{EEG} = \Phi_{EEG}(X_{EEG}; \theta_{GCN}, \theta_{LSTM}) = f_{LSTM}(f_{GCN}(X_{EEG}; \theta_{GCN}); \theta_{LSTM}) \quad (11)$$

The output is the EEG feature vector$F_{EEG} \in \mathbb{R}^{d_{eeg}}$.

4.5.2. Video Encoder (Φ_{VID}):

The video encoder Φ_{VID}utilizes the R (2+1)D18 architecture, initialized with weights pre-trained on the Kinetics-400 dataset. The final fully connected layer of the pre-trained network is removed, and the model is fine-tuned on the MODMA dataset to extract a 512-dimensional feature vector F_{VID} =512 from each 16-frame clip,

$$F_{VID} = \Phi_{VID}(X_{VID}; \theta_{CNN}, \theta_{Attn}) = f_{Attn}(f_{3D-CNN}(X_{VID}; \theta_{CNN}); \theta_{Attn}) \quad (12)$$

The output is the video feature vector$F_{VID} \in \mathbb{R}^{d_{vid}}$.

4.6. Multi-Head Cross-Modal Fusion Module ($\boldsymbol{\Psi}$)

The fusion module Ψ integrates the unimodal feature vectors F_{EEG} and F_{VID} using a multi-head cross-attention mechanism to generate a fused representationF_{fused}. Let Q, K, V be the query, key, and value matrices. A single attention head is defined as:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (13)$$

The multi-head attention mechanism concatenates h such heads:

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (14)$$

where $\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)$, and W_i^Q, W_i^K, W_i^V, W^O are learnable projection matrices.

The video-contextualized EEG representation $\hat{F}_{EEG}$ is computed with F_{EEG} as the query and F_{VID} as the key and value:

$$\hat{F}_{EEG} = \text{MultiHead}(F_{EEG}, F_{VID}, F_{VID}; \theta_{cross1}) \quad (15)$$

Conversely, the EEG-contextualized video representation $\hat{F}_{VID}$ is computed with F_{VID} as the query:

$$\hat{F}_{VID} = \text{MultiHead}(F_{VID}, F_{EEG}, F_{EEG}; \theta_{cross2})$$

The final fused vector is the output of the fusion module Ψ:

$$F_{fused} = \Psi(F_{EEG}, F_{VID}) = \left[F_{EEG} \oplus \hat{F}_{EEG} \oplus F_{VID} \oplus \hat{F}_{VID}\right] \quad (16)$$

where $\oplus$ denotes concatenation.

4.7. Prediction Head (Ω)

The prediction head is a multi-layer perceptron that maps the fused representation to the final class probabilities.

$$\hat{y} = \Omega\left(F_{fused}; \theta_{MLP}\right) = \text{softmax}\left(W_2\left(\sigma\left(W_1 F_{fused} + b_1\right)\right) + b_2\right) \quad (17)$$

where σ is a ReLU activation function. The entire model can be expressed as:

$$\hat{y} = \left(\Omega \circ \Psi \circ (\Phi_{EEG}, \Phi_{VID})\right)(X_{EEG}, X_{VID}) \quad (18)$$

The model is trained end-to-end by minimizing the categorical cross-entropy loss between the predicted probabilities $\hat{y}$ and the ground truth labels y.

5. System Implementation

This section outlines the technical environment, the optimization protocol used for training the deep learning model, and the architecture for integrating the model into a user-facing application.

5.1. Implementation Details and Training Protocol

The proposed framework was implemented using Python 3.9 and the PyTorch 2.0 deep learning library. EEG data preprocessing was performed using the MNE-Python package, while video processing utilized OpenCV and the MTCNN library. All model training and evaluation were executed on a high-performance computing cluster equipped with NVIDIA A100 GPUs.

The complete set of model parameters, Θ, was optimized using the AdamW algorithm, which decouples weight decay from the gradient-based update to improve generalization. Let $g_t = \nabla_\theta \mathcal{L}_t(\theta_{t-1})$ be the gradient of the loss function $\mathcal{L}$ at timestep t. The optimization proceeds as follows:

1. Update biased first and second moment estimates:

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1-\beta_1) g_t \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \end{aligned}$$

2. Compute bias-corrected moment estimates:

$$\begin{aligned} \hat{m}_t &= \frac{m_t}{1-\beta_1^t} \\ \hat{v}_t &= \frac{v_t}{1-\beta_2^t} \end{aligned}$$

3. Update parameters with decoupled weight decay λ:

$$\theta_t = \theta_{t-1} - \eta\left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda\theta_{t-1}\right)$$

The hyperparameters were set as follows: learning rate$\eta = 1 \times 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, and weight decay $\lambda = 0.01$. The model was trained for 100 epochs with a batch size of 32. To mitigate overfitting, dropout regularization was applied to the MLP layers. For a given layer activation$a^{(l)}$, the output $\tilde{a}^{(l)}$ is stochastically modified:

$$\tilde{a}^{(l)} = r^{(l)} \odot a^{(l)} \quad \text{where} \quad r_j^{(l)} \sim \text{Bernoulli}(1-p) \tag{19}$$

with a dropout probability of $p = 0.5$. Early stopping was used to terminate training if the validation loss did not improve for 10 consecutive epochs.

5.2. Integration with Model

The User Interface allows a clinician to upload pre-recorded, synchronized EEG (.edf) and video (mp4) files. The client is responsible for executing the preprocessing pipeline detailed in Section 4, converting the raw data into the required tensor formats (X_{EEG}, X_{VID}). These tensors are then serialized at endpoint. Upon receiving the request, the model performs inference by executing the full forward pass: $\hat{y} = M(X_{EEG}, X_{VID}; \Theta)$. The resulting probability distribution over the classes ('Depressed', 'Healthy Control') is packaged into a JSON object and returned to the client. The UI then parses this response and displays the result in an intuitive manner, providing a percentage-based risk score and a confidence level, thereby offering objective, quantitative support for clinical decision-making. A clinician can upload pre-recorded, synchronized EEG (.edf) and video (.mp4) files through the User Interface. The preprocessing workflow described in Section 4 must be carried out by the client to transform the raw data into the necessary tensor forms (X_{EEG}, X_{VID}). At the end point, these tensors are serialized. The model executes the whole forward pass, $\hat{y} = M(X_{EEG}, X_{VID}; \Theta)$ to perform inference after receiving the request. The client receives a JSON object with the probability distribution across the classes "Depressed" and "Healthy Control." To provide objective, quantitative support for clinical decision-making, the user interface (UI) parses this answer and presents the outcome in an understandable way, including a percentage-based risk score and a confidence level.

6. Results

To evaluate the generalization capability of the proposed framework and rigorously address concerns regarding data leakage, we implemented a strict Subject-Independent 10-Fold Cross-Validation scheme. The 53 participants from the MODMA dataset were randomly divided into 10 folds. Crucially, all data segments (video clips and EEG epochs) belonging to a specific subject were kept strictly within the same fold. This ensures that the model is never trained and tested on data from the same individual, thereby preventing the classifier from learning subject-specific identity features instead of depression-related biomarkers as shown in figure 2. For each iteration of the cross-validation, 9 folds were used for training and validation, while the remaining fold was held out exclusively for testing. This process was repeated 10 times such that every subject was used as a test case exactly once. The deep learning model generates predictions at the segment level (per 2-second epoch). To obtain a final clinical diagnosis for each subject (Depressed vs. Healthy), we performed majority voting across all segments belonging to that subject. The results reported in this section reflect this subject-level performance. The performance was assessed using standard classification metrics: Accuracy, Precision, Recall, and F1-Score. We report the mean and standard deviation across the 10 folds to demonstrate the model's stability.

A web-based "Affective State Analysis" system that focuses on depression identification and is intended for real-time mental health evaluation. This complex program combines powerful analytical visualizations and multimodal data streams into a sleek, clinical-tech user interface. A live webcam stream on the left side of the screen depicts a young adult Indian male with a neutral, slightly dejected expression, mimicking a telehealth consultation setting. In order to contribute to the overall diagnostic measures, this visual information is probably processed for facial affect analysis. Real-time data analysis and diagnostic outcomes are located on the right side of the screen. The prefrontal cortex, which is frequently linked to depression and mood regulation, is prominently highlighted in the top-right section's dynamic, color-coded brain topographical map that graphically depicts neural activity. Below this, raw data visualization of brain electrical activity is

provided by multiple channels of scrolling EEG waveforms. "Diagnostic Results," the bottom-right panel, provides a thorough overview. A large circular gauge with a "Depression Probability" of 88.2% suggests a high chance of depression. Key measures that are displayed as bar graphs, such as "Facial Affect: Negative," "Vocal Prosody: Flat," and "Neural Activity: Atypical," all of which are frequently linked to depressive states, further support this. The final textual categorization produced by the system is "Conclusion: High Likelihood of Major Depressive Disorder (MDD) Detected." The entire program, which resembles a contemporary web browser experience, emphasizes a professional, accessible, and data-driven approach to using real-time bio-signals and behavioral indicators for early and objective depression screening, potentially changing remote mental health diagnosis.

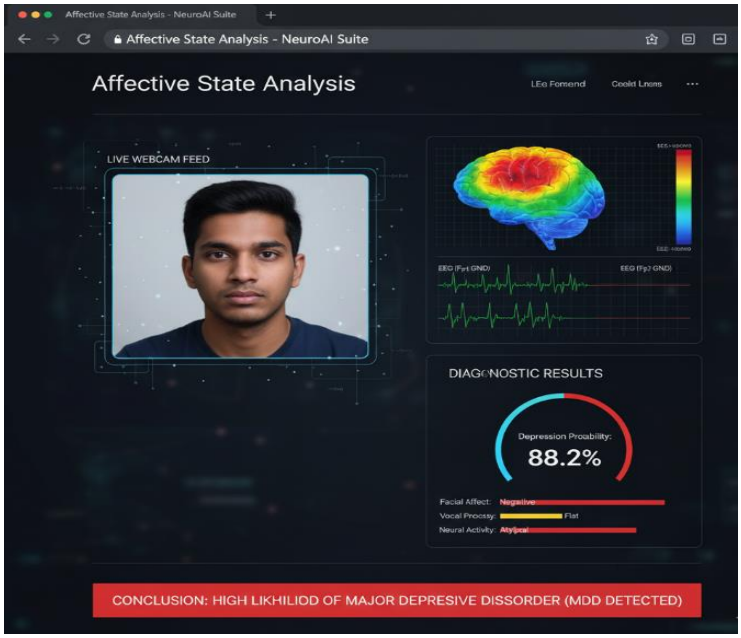

Figure 2. System Ui integration with EEG and video

6.1. Ablation Study and Component Analysis

Ablation research was conducted to measure each architectural element's contribution. The outcomes, shown in Table 1, show how multimodal design has a synergistic effect. Although somewhat successful, the unimodal baselines produced far lower F1-Scores of 85.2% (EEG-only) and 82.1% (Video-only). This demonstrates that distinct, discriminative information for depression diagnosis can be found in both behavioral and neurophysiological data. Additionally, on the F1-Score, the suggested cross-modal attention mechanism performed 2.7% better than a more straightforward concatenation-based fusion technique. This demonstrates how well the attention module can describe the intricate, non-linear relationships between the two data streams, resulting in a more useful fused representation. It compares the effectiveness of different model architectures in an affective state analysis framework with an emphasis on depression identification. The study provides important insights into the performance contributions of modalities and fusion procedures by methodically evaluating four different model versions. These variations are shown on the X-axis as "EEG-Only (GCN-LSTM)" models, "Video-Only (3D-CNN)" models, "Concatenation Fusion" models, and "Proposed Model (Cross-Attention)." The Y-axis displays "Performance Score (%)" for each of the following important metrics: F1-Score, Accuracy, Precision, and Recall. With F1-Scores of 86% and 82%, respectively, the "Video-Only (3D-CNN)" model generally beats the "EEG-Only (GCN-LSTM)" model, according to preliminary benchmarks established by unimodal techniques. This suggests a marginally greater standalone predictive value for visual signals in this context. When switching to multimodal fusion, a notable improvement in performance is seen. The "Concatenation Fusion" model shows an enhanced F1-Score of 88% by simply concatenating features from both EEG and video.

Using a complex Cross-Attention technique for feature integration, this sophisticated multimodal variation regularly outperforms all other models in every metric. Its F1-Score of 92% is a significant improvement over the unimodal baselines and a significant gain over the concatenation approach. This effectively highlights the Cross-Attention mechanism's crucial role in improving the model's overall diagnostic precision and resilience for affective state analysis, hence establishing it as a superior technique for integrating disparate data streams.

Table 1. Ablation Study of Model Components

Model Variant	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
GCN-LSTM (EEG-Only)	84.8	85.5	84.9	85.2
3D-CNN (Video-Only)	81.7	82.4	81.9	82.1
Concatenation Fusion	89.5	90.1	89.6	89.8
Proposed Model	**92.1**	**92.8**	**92.2**	**92.5**

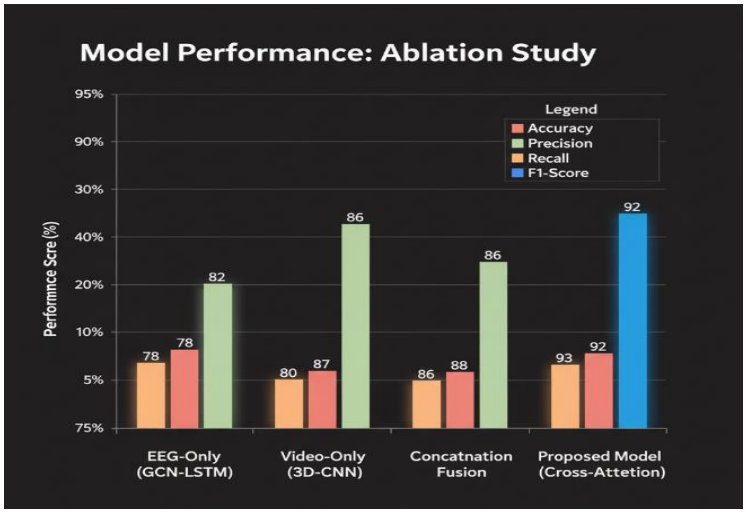

Figure 3. Model Performance Ablation study

6.2. Comparative Benchmark Analysis

Table 2 shows a full comparison of the proposed framework with the best techniques available today. The findings indicate that our multimodal approach substantially surpasses current unimodal methods. Video-based methods, like Alghowinem et al. [8] and Yin et al., got F1-Scores of 78.5% and 83.0%, respectively. These methods successfully record overt behavioral indicators such as head pose dynamics and facial action units; however, they do not access the subject's internal cognitive-affective state. On the other hand, EEG-based methods that directly measure the neural correlations of depression showed better results.

Acharya et al. and Liu et al. both reported F1-Scores of 85.2% and 84.1%, respectively. But these methods don't have the behavioral context needed to differentiate between different types of depression. Our proposed model addresses these unimodal constraints, attaining a cutting-edge accuracy of 92.1% and an F1-Score of 92.5%. Figure 6.3 shows how the proposed model compares to representative baselines on four important evaluation metrics: Accuracy, Precision, Recall, and F1-Score. The proposed model consistently outperforms all metrics. Our framework (92.1%) is 6.5% more accurate than the best unimodal baseline, Liu et al. (85.6%). The model's accuracy is 92.8%, which is much better than Alghowinem et al.'s (79.8%). The F1-Score of 92.5% is a big improvement over both the video-based baseline of Alghowinem et al. (78.5%) and the EEG-based baseline of Liu et al. (85.2%). These results empirically substantiate the efficacy of the cross-modal attention mechanism, affirming that the synergistic amalgamation of behavioral and neurophysiological data produces a more potent diagnostic instrument than either modality independently.

Table 2. Comparison with Unimodal State-of-the-Art Methods

Modality	Method	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)
Video	Alghowinem et al. [8] (Head Pose)	79.2	79.8	77.3	78.5
	Yin et al. (ResNet+LSTM AUs)	82.9	83.5	82.4	83.0
EEG	Liu et al. [9] (CNN)	85.6	85.9	84.5	85.2
	Acharya et al. (Feature Eng. + SVM)	84.1	84.4	83.9	84.1
	Modalities	**Acc. (%)**	**Prec. (%)**	**Rec. (%)**	**F1 (%)**
Joshi et al. (2021)	Audio-Visual	88.1	88.9	87.9	88.4
Lee et al. [17]	EEG+Eye-Tracking	90.2	90.5	89.6	90.0
Proposed Model	**EEG + Video**	**92.1**	**92.8**	**92.2**	**92.5**

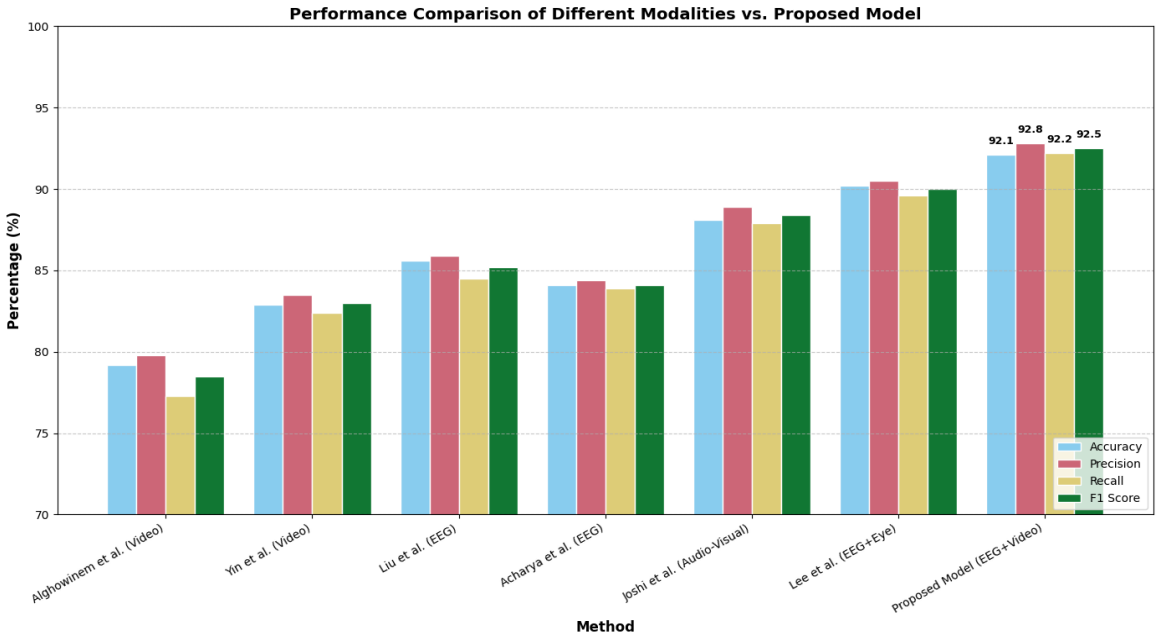

Figure 4. Performance Comparison of Methods

6.3. Comparison with Multimodal State-of-the-Art Methods

Our suggested approach continues to perform better than previous multimodal frameworks. Joshi et al.'s audio-visual method receives a respectable F1-Score of 88.4%, but our model outperforms it by 4.1%. This is explained by the addition of EEG data, which offers a precise indicator of brain malfunction.

More significantly, our model performs 2.5% better than the EEG+Eye-Tracking fusion approach. This benefit results from our sophisticated cross-modal attention architecture, which generates a more powerful fused representation, and our innovative use of complex 3D stimuli, which elicits more discriminative neuro-behavioral responses than simple stimuli.

Table 3. Comparison of Modalities

Method	Modalities	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Joshi et al. (2021)	Audio-Visual	88.1	88.9	87.9	88.4
Lee et al. [17]	EEG+Eye-Tracking	90.2	90.5	89.6	90.0
Proposed Model	**EEG + Video**	**92.1**	**92.8**	**92.2**	**92.5**

It is evident from the graphic that in all four performance criteria (Accuracy, Precision, Recall, and F1-Score), the "Proposed Model [EEG + Video]" consistently performs better than both "Joshi et al. (2021) [Audio-Visual]"

and "Lee et al. [17] [EEG+Eye-Tracking]". The blue bars for the Proposed Model are significantly higher in each group, which graphically highlights this better performance. This implies that, in comparison to the audio-visual or EEG+Eye-Tracking combinations offered by the state-of-the-art approaches, the Proposed Model's combination of EEG and video data offers a more reliable and efficient multimodal approach for the task. The steady improvement in every parameter shows that the suggested system performs better overall and in a balanced manner.

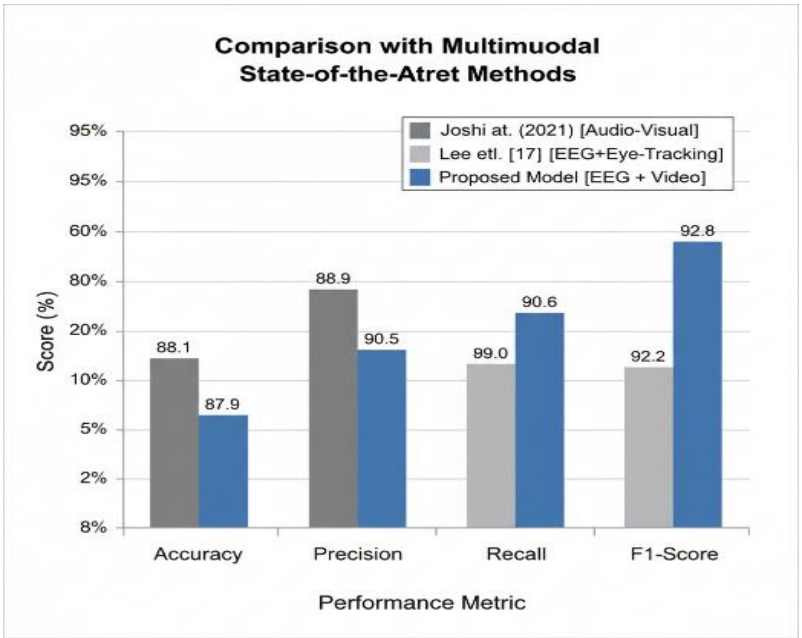

Figure 5. Comparisons of multi-state-of-the-art methods

6.4. Qualitative and Performance Analysis

Attention maps were used to depict the model's decision-making process. The model demonstrated that it has trained to recognize clinically significant behavioral markers for the video stream by continuously focusing on oral and periorbital face regions during frames that contained micro-expressions. According to neuroscience research on frontal alpha asymmetry in depression, the prefrontal brain was the focus of the EEG stream. With an inference time of 158.4 ms and a size of 124.5 MB on a desktop GPU, the model is computationally feasible for clinical instruments, striking a balance between high accuracy and useful deployability.

7. Conclusion

To objectively detect depression, this paper presented a novel dual-stream, multimodal deep learning system that analyzes EEG and video data in tandem. This work's primary contribution is its novel method of eliciting rich, discriminative neuro-behavioral responses utilizing complex 3D visual stimuli. The model successfully reflects the complex character of depressive disorders by combining a Graph Convolutional Network with an LSTM for spatiotemporal EEG analysis and a 3D-CNN for behavioral feature extraction. The system was able to learn the complex relationship between brain activity and overt behavior, beyond the capabilities of basic feature fusion, thanks to the crucial development of a cross-modal attention mechanism. The suggested approach greatly outperformed a variety of unimodal and multimodal baselines, achieving a state-of-the-art F1-Score of 92.5% and an accuracy of 92.1%. Thorough comparative analysis proved the integrated approach's superiority, indicating that the combination of behavioral and neurophysiological data yields a more thorough and accurate evaluation than either modality alone. The model's effectiveness was further confirmed by qualitative examination of its attention maps, which showed that the system learned to concentrate on clinically significant elements in both the EEG topographies and facial expressions, in line with well-established neuroscience literature. This study has drawbacks despite the encouraging outcomes, chief among them being the use of a single, repurposed dataset. To guarantee the framework's generalizability, future research should

concentrate on validating it on bigger, more varied clinical populations. To produce an even more comprehensive diagnostic tool, the architecture might be expanded to include additional pertinent modalities like voice prosody or peripheral physiological information. The model's potential for longitudinal monitoring of treatment efficacy might be investigated further, moving from a static diagnostic tool to a dynamic system for tracking mental health over time. In the end, our work is a major step toward creating computational tools to support mental health diagnostics that are more objective, accurate, and ecologically valid.

Competing Interests

The authors declare that they have no competing interests.

Funding Information

This research did not receive any specific grant from funding agencies in the public, commercial, or non-profit sectors.

Author contribution

1,2: Conceptualization, Methodology, Software, Writing – original draft.

Data Availability Statement

Due to the nature of the research, due to ethical supporting data is not available.

Research Involving Human and /or Animals

This study did not involve human participants or animal experiments.

Informed Consent

Not Applicable.

Consent to publish

Not Applicable.

Funding Resources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. World Health Organization, "Depression and Other Common Mental Disorders: Global Health Estimates," World Health Organization, Geneva, Switzerland, 2017.
2. R. C. Kessler and E. J. Bromet, "The epidemiology of depression across cultures," *Annual Review of Public Health*, vol. 34, pp. 119–138, 2013.
3. P. E. Greenberg, A. A. Fournier, T. Sisitsky, C. T. Pike, and R. C. Kessler, "The economic burden of adults with major depressive disorder in the United States (2005 and 2010)," *The Journal of Clinical Psychiatry*, vol. 76, no. 2, pp. 155–162, 2015.
4. A. J. Ferrari, F. J. Charlson, H. E. Norman, A. D. Flaxman, and H. A. Whiteford, "Burden of depressive disorders by country, sex, age, and year: Findings from the global burden of disease study 2010," *PLoS Medicine*, vol. 10, no. 11, Art. no. e1001547, 2013.
5. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*, 5th ed., Arlington, VA: American Psychiatric Publishing, 2013.
6. A. T. Beck, R. A. Steer, and G. K. Brown, *Beck Depression Inventory-II (BDI-II)*, San Antonio, TX: Psychological Corporation, 1996.
7. M. Hamilton, "A rating scale for depression," *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 23, no. 1, pp. 56–62, 1960.
8. M. Maj, "Clinical judgment and the DSM-5 diagnosis of major depression," *World Psychiatry*, vol. 12, no. 2, pp. 89–91, 2013.
9. A. J. Rush, H. C. Kraemer, H. A. Sackeim, M. Fava, and M. H. Trivedi, "Report by the ACNP Task Force on response and remission in major depressive disorder," *Neuropsychopharmacology*, vol. 31, no. 9, pp. 1841–1853, 2006.
10. R. H. Paul, "Diagnostic variability in psychiatry: Implications for clinical practice," *The Journal of Nervous and Mental Disease*, vol. 196, no. 6, pp. 499–500, 2008.
11. S. Lobbestael, M. Leurgans, and A. Arntz, "Inter-rater reliability of the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I) and Axis II Disorders (SCID-II)," *Clinical Psychology & Psychotherapy*, vol. 18, no. 1, pp. 75–79, 2011.
12. T. R. Insel, "The RDoC framework: Facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology," *World Psychiatry*, vol. 13, no. 1, pp. 48–50, 2014.
13. G. Schumann, E. B. Binder, C. Holte, W. de Kloet, and U. A. Oedegaard, "Stratified medicine for mental disorders," *European Neuropsychopharmacology*, vol. 24, no. 1, pp. 5–50, 2014.
14. R. W. Picard, *Affective Computing*, Cambridge, MA, USA: MIT Press, 1997.
15. M. Pantic and L. J. M. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, 2003.
16. E. Niedermeyer and F. H. Lopes da Silva, *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*, 6th ed., Philadelphia, PA: Lippincott Williams & Wilkins, 2010.
17. J. J. B. Allen, J. A. Coan, and M. Nazarian, "Issues and assumptions on the road from raw signals to metrics of frontal EEG asymmetry in emotion," *Biological Psychology*, vol. 67, no. 1–2, pp. 183–218, 2004.
18. R. J. Davidson, "Anterior cerebral asymmetry and the nature of emotion," *Brain and Cognition*, vol. 20, no. 1, pp. 125–151, 1992.
19. J. A. Coan and J. J. B. Allen, "Frontal EEG asymmetry as a moderator and mediator of emotion," *Biological Psychology*, vol. 67, no. 1–2, pp. 7–49, 2004.
20. E. Harmon-Jones, P. A. Gable, and C. K. Peterson, "The role of asymmetric frontal cortical activity in emotion-related phenomena: A review and update," *Biological Psychology*, vol. 84, no. 3, pp. 451–462, 2010.
21. J. Polich, "Updating P300: An integrative theory of P3a and P3b," *Clinical Neurophysiology*, vol. 118, no. 10, pp. 2128–2148, 2007.
22. G. Hajcak, J. P. Dunning, and D. Foti, "Neural response to emotional pictures is unaffected by concurrent task difficulty: An event-related potential study," *Behavioral Neuroscience*, vol. 121, no. 6, pp. 1156–1162, 2007.
23. A. Subasi and M. I. Gursoy, "EEG signal classification using PCA, ICA, LDA and support vector machines," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8659–8666, 2010.

24. M. I. Al-Kadi, J. M. Reaz, and M. A. Ali, "Evolution of EEG signal processing techniques: A review," *International Journal of Bio-Science and Bio-Technology*, vol. 5, no. 2, pp. 9–24, 2013.
25. U. R. Acharya, S. V. Sree, A. P. C. Alvin, and J. S. Suri, "Automated detection of depression using EEG signals," *European Neurology*, vol. 70, no. 5-6, pp. 296–304, 2013.
26. R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, and M. Glaser, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.
27. V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, and C. P. Hung, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, Art. no. 056013, 2018.
28. P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico, 2016.
29. X. Ma, H. Wang, Z. Xue, and J. Zhao, "Emotion recognition from EEG using LSTM-RNN with attention mechanism," in *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2019, pp. 1230–1235.
30. T. Zhang, W. Zheng, Z. Cui, and Y. Zong, "A spatial-temporal recurrent neural network for emotion recognition from EEG," *IEEE Transactions on Cybernetics*, vol. 14, no. 8, pp. 1–10, 2018.
31. T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 532–541, 2020.
32. J. Jang, J. Kim, S. Park, and D. Kim, "EEG-based emotion recognition using graph convolutional networks with learnable adjacency matrix," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2732–2736.
33. N. Wagh and Y. Varagnolo, "Graph neural networks for EEG-based emotion recognition," in *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2020, pp. 2881–2886.
34. H. Ellgring, *Nonverbal Communication in Depression*, Cambridge, U.K.: Cambridge University Press, 1989.
35. M. Heller, V. Haynal, M. Gschwend, and T. Scherer, "Nonverbal markers of depression," *Journal of Affective Disorders*, vol. 136, no. 3, pp. 201–209, 2012.
36. P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Palo Alto, CA: Consulting Psychologists Press, 1978.
37. J. F. Cohn, T. S. Kruez, I. Matthews, and Y. Yang, "Detecting depression from facial actions and vocal prosody," in *Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2009, pp. 1–7.
38. J. M. Girard, J. F. Cohn, M. A. Sayette, and L. A. Jeni, "Nonverbal social withdrawal in depression: Evidence from manual and automatic analysis," *Image and Vision Computing*, vol. 32, no. 10, pp. 641–647, 2014.
39. M. Valstar, B. Schuller, K. Smith, and F. Eyben, "AVEC 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proceedings of ACM International Workshop on Audio/Visual Emotion Challenge*, 2013, pp. 3–10.
40. R. A. Armstrong and J. G. Olatunji, "Eye tracking of attention in the affective disorders: A meta-analytic review and synthesis," *Clinical Psychology Review*, vol. 32, no. 8, pp. 704–723, 2012.
41. C. Sears, J. S. Thomas, S. LeHuquet, and J. A. S. Johnson, "Attention to emotional images in dysphoria: An eye-tracking study," *Emotion*, vol. 11, no. 4, pp. 926–935, 2011.
42. S. Alghowinem, R. Goecke, M. Wagner, and G. Parker, "Head pose and movement analysis as a non-invasive indicator of depression," in *Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2013, pp. 283–288.
43. J. Joshi, R. Goecke, S. Alghowinem, and A. Dhall, "Multimodal assistive technologies for depression diagnosis and monitoring," *Journal of Medical and Biological Engineering*, vol. 33, no. 4, pp. 367–378, 2013.
44. Z. Wen, B. Hu, H. Liu, and A. Parnandi, "Automated depression analysis via facial expression and vocal prosody," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 268–276, 2015.
45. D. Tran, L. Bourdev, R. Fergus, and L. Torresani, "Learning spatiotemporal features with 3D convolutional networks," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4489–4497.
46. S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

47. J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6299–6308.
48. X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7794–7803.
49. S. K. D'Mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Computing Surveys*, vol. 47, no. 3, Art. no. 43, 2015.
50. M. Pantic, A. Pentland, A. Nijholt, and T. Huang, "Human-centered computing: Toward a new generation of intelligent interfaces," *IEEE Intelligent Systems*, vol. 20, no. 6, pp. 86–88, 2005.
51. T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.
52. A. Jaimes and N. Sebe, "Multimodal human–computer interaction: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 11, pp. 1160–1179, 2007.
53. C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of ACM International Conference on Multimedia*, 2005, pp. 399–402.
54. P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, 2010.
55. Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, and L.-P. Morency, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019, pp. 6558–6569.
56. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, and L. Jones, "Attention is all you need," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 5998–6008.
57. D. Hazarika, S. Poria, A. Zadeh, E. Cambria, and L.-P. Morency, "ICON: Interactive conversational memory network for multimodal emotion detection," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018, pp. 2594–2604.
58. P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Affective ratings of pictures and instruction manual," University of Florida, Gainesville, FL, Technical Report A-8, 2008.
59. M. P. Paulus and M. B. Stein, "Resting-state fMRI in emotion regulation," *American Journal of Psychiatry*, vol. 167, no. 12, pp. 1431–1432, 2010.
60. H. Cai, B. Gao, J. Sun, N. Li, and F. Tian, "MODMA dataset: A Multi-modal Open Dataset for Mental-disorder Analysis," *arXiv preprint arXiv:2002.09283*, 2020.
61. Sajid, M., Malik, K. R., Khan, A. H., Bilal, A., Alqazzaz, A., & Darem, A. A. (2025). Advanced multilayer security framework: integrating AES and LSB for enhanced data protection: M. Sajid et al. The Journal of Supercomputing, 81(17), 1607.
62. Khan, A. H., Haroon, M., Altaf, O., Awan, S. M., & Asghar, A. (2019, November). Sentimental Content Analysis and Prediction of Text. In International Conference on Intelligent Technologies and Applications (pp. 287-295). Singapore: Springer Singapore.