Multimodal transformer for depression detection based on EEG and interview data

Esmi, Nima; Shahbahrami, Asadollah; Gaydadjiev, Georgi; de Jonge, Peter

DOI
10.1016/j.bspc.2025.109039

Publication date
2026

Document Version
Final published version

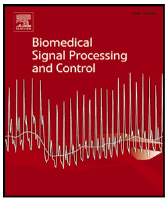
Published in
Biomedical Signal Processing and Control

Citation (APA)
Esmi, N., Shahbahrami, A., Gaydadjiev, G., & de Jonge, P. (2026). Multimodal transformer for depression detection based on EEG and interview data. *Biomedical Signal Processing and Control*, *113*, Article 109039. https://doi.org/10.1016/j.bspc.2025.109039

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



Multimodal transformer for depression detection based on EEG and interview data

Nima Esmi [a,b,*], Asadollah Shahbahrami [c,b], Georgi Gaydadjiev [d], Peter de Jonge [e]

[a] Bernoulli Institute, University of Groningen, Groningen, The Netherlands
[b] ISRC, Khazar University, Baku, Azerbaijan
[c] Department of Computer Engineering, University of Guilan, Guilan, Iran
[d] Computer Engineering Laboratory, Delft University of Technology, Delft, The Netherlands
[e] Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, The Netherlands

ARTICLE INFO

Keywords:
Depression detection
EEG
Multimodal transformer
Modality synchronization
Flexible temporal sequence matching

ABSTRACT

Depression detection benefits from combining neurological and behavioral indicators, yet integrating heterogeneous modalities such as EEG and interview audio remains challenging. We propose a transformer-based multimodal framework that jointly models spectral, spatial, and temporal EEG features alongside linguistic and paralinguistic cues from interviews. By employing synchronized multi-head cross-attention and self-attention mechanisms, the model effectively captures intra- and inter-modal correlations. In addition, a flexible temporal sequence matching strategy reduces EEG channel requirements, enhancing device portability. Evaluated on the MODMA and DAIC-WOZ datasets, our approach achieves superior performance compared to state-of-the-art models, with a 4.7% improvement in accuracy and a 10% increase in precision. These results demonstrate the potential of the proposed framework for accurate, scalable, and cost-effective depression detection in both clinical and real-world settings.

1. Introduction

Depression is a prevalent mental health condition with significant personal and societal impacts, including disability, diminished quality of life, and an increased suicide risk [1]. Traditional diagnostic methods for depression rely on a combination of objective biomedical evaluations, such as electroencephalography (EEG), and subjective behavioral assessments, like clinical interviews. Among biomedical tools, EEG is prominent due to its ability to directly measure neurological brain activity, crucial for detecting depression [2–5]. Research has explored various EEG signal features, such as power spectral density and connectivity [6–9]. Traditional EEG-based approaches involve two steps: feature extraction and classification, using machine learning classifiers like support vector machines (SVMs) and decision trees [10–13]. However, these methods depend heavily on handcrafted features and often fail to generalize across diverse populations.

To address these limitations, deep learning methods such as convolutional neural network (CNNs) and long short-term memory (LSTM) have been increasingly applied to automate EEG feature extraction, demonstrating notable improvements in accuracy [14–17]. Parallel to neurological signals, behavioral indicators derived from clinical interviews have also proven valuable. Advances in machine learning allow automatic extraction of linguistic and paralinguistic features, where lexical diversity, semantic cues, prosody, and speech tone provide complementary information for depression detection [18–21]. Combining these heterogeneous modalities has thus become an active research direction.

Despite these advances, recent state-of-the-art multimodal methods, such as graph-based neural networks and transformer-based fusion frameworks, still face important challenges. For example, graph neural network (GNN)-based approaches capture modality-specific and shared structures but often struggle with inter-modal synchronization when modalities are not temporally aligned [22]. Transformer-based systems such as hierarchical or tensor-based fusion models improve long-term dependency modeling, yet they are limited in simultaneously integrating spectral, spatial, and temporal EEG information with linguistic and paralinguistic signals [23–28]. Moreover, the complexity and cost of EEG acquisition remain barriers to scalability, as most existing approaches require large numbers of electrodes and cumbersome hardware setups.

To address these gaps, this study proposes a multimodal transformer model that integrates EEG signals—transformed into two-dimensional representations preserving spectral, spatial, and temporal

[*] Corresponding author at: Bernoulli Institute, University of Groningen, Groningen, The Netherlands.
E-mail address: n.esmi.rudbardeh@rug.nl (N. Esmi).

https://doi.org/10.1016/j.bspc.2025.109039
Received 14 March 2025; Received in revised form 5 October 2025; Accepted 31 October 2025
Available online 5 November 2025

information—with interview data capturing both linguistic and paralinguistic cues. By employing synchronized multi-head cross-attention and self-attention mechanisms, our model effectively aligns heterogeneous modalities and extracts richer inter- and intra-modal dependencies, overcoming the limitations of previous CNN-, LSTM-, and GNN-based approaches. Furthermore, to improve practicality in real-world applications, we introduce the flexible temporal sequence matching (FTSM) technique for EEG channel selection. This strategy substantially reduces the number of required electrodes while maintaining competitive accuracy, thereby improving portability and lowering deployment costs.

The main contributions of this paper are summarized below:

- We propose a multimodal transformer architecture that jointly models spectral, spatial, and temporal EEG patterns alongside linguistic and paralinguistic interview features. In comparison with existing state-of-the-art approaches, our framework achieves superior depression detection performance by explicitly synchronizing intra- and inter-modal correlations.
- We introduce the FTSM technique for EEG channel prioritization, which allows reducing the number of EEG channels from 128 to 4 with only a marginal drop in accuracy (91% to 84%). This contribution directly addresses the cost and scalability limitations of wearable EEG systems, making the approach more applicable to real-world and remote clinical settings.

The remainder of this article is organized as follows: Section 2 reviews the related work and sets the context for our research. Section 3 describes the proposed methodology, including the architecture of the multimodal transformer and the data transformation techniques. Section 4 presents the experimental setup, datasets, evaluation metrics, and results. Section 5 concludes the article with a summary of the key contributions. Table 1 lists the main notations used throughout the paper.

2. Related work

In this section, we review related work on depression detection, grouped according to the data modality and previously applied methodologies. We cover EEG-based approaches, audio/text-based methods, multimodal strategies, transformer-based frameworks, and EEG channel selection techniques. The aim is to highlight both advancements and persistent gaps that motivate our proposed model.

2.1. Approaches based on modality

EEG has been widely used in depression detection because it directly measures brain activity. Studies have analyzed EEG features across spectral, spatial, and temporal dimensions, using handcrafted features and classical machine learning classifiers such as SVMs or decision trees [6–13]. Although effective to some extent, these methods rely heavily on manual feature extraction and often fail to generalize across datasets.

Behavioral markers of depression, including linguistic [29] and paralinguistic cues, have also been studied extensively. Shin et al. [18] and Bauer et al. [19] analyzed linguistic and speech features, while Sardari et al. [20] focused on paralinguistic indicators. More recent works highlight the role of lexical diversity and prosody in capturing depressive states [21]. Transformer-based large language models have also been fine-tuned for social media posts, showing strong sensitivity to subtle cues of stress and depression [30,31].

Given the multifaceted nature of depression, multimodal strategies that integrate EEG with other behavioral modalities have gained prominence. Zhu et al. [32] and Chen et al. [33] demonstrated that fusing EEG with eye movement or interviews improves accuracy. Zhang et al. [22] proposed a GNN-based modal-shared modal-specific architecture to capture heterogeneity and homogeneity across modalities, though it remains sensitive to temporal misalignment. Recent multimodal studies extend fusion to EEG, audio, and video data, showing the advantage of cross-modal information sharing [34–40].

Table 1
Notations and definitions

Notation	Definition
X	First sequence of length n
Y	Second sequence of length m
$ftsm\_dist$	Distance between the sequences X and Y
D	Cost matrix for X and Y
$d(x_i, y_j)$	Distance between x_i and y_j
$(value, idx)$	Minimum value and its index among neighboring cells
$s(t)$	Input signal
$s(p+q)$	Signal $s(t)$ sampled at time $(p+q)$
q	Current time frame or window position
p	Position of time samples within each window
L	Length of the finite window
$\omega(p)$	Window function applied to the signal
$e^{-j\omega p}$	Frequency-dependent phase shift
$-j\omega p$	Inverse Fourier transform convention
j	Imaginary unit
ω	Frequency component
Q_T	Query token
K_T	Key token
V_T	Value token
d_k	Size of the vector space for the key and query vectors
Z^l	Input to the self-attention block at layer l
W	Weight matrix
T_p	Transpose
$Norm$	Layer normalization
MLP	MultiLayer perceptron
A	Audio sequence
E	EEG sequence
r	Offset range
ϵ	Smoothing factor
t	Time sample
A_t	Feature vectors for the audio at time t
E_t	Feature vectors for the EEG at time t
$\cdot$	Dot product of the vectors
$\|\|A_t\|\|$	Euclidean norms of the audio feature vectors
$\|\|E_t\|\|$	Euclidean norms of the EEG feature vectors
T	Total number of frames in the sequence vectors
o	Possible offset
$e^{similarity(o)}$	Exponential cosine similarity for a particular Offset o
$p(o)$	Predicted probability for each offset o
$label_o$	Represents the ground-truth label for the offset
ϵ	Smoothing factor

2.2. Previously applied methodologies

Transformers have emerged as a powerful tool for multimodal depression detection, owing to their ability to model long-term dependencies and integrate heterogeneous modalities. For instance, TensorFormer [41] uses tensor algebra to retain modality structure while enabling cross-modal interactions. Teng et al. [42] developed an intra- and inter-emotion fusion transformer that balances homogeneous and heterogeneous emotional cues. Fan et al. [43] proposed DepMSTAT, which leverages spatio-temporal attention to jointly analyze EEG and behavioral cues, successfully modeling speech prosody shifts [44]. Li et al. [45] incorporated contrastive learning to improve transformer robustness, while Zhu et al. [46] introduced MTNet, which fuses EEG and eye-tracking for early depression detection. Sun et al. [28] applied graph-based attention mechanisms to audio-based depression detection. Despite these advances, limitations remain. GNN-based methods are restricted in handling modality synchronization, while transformer variants like TensorFormer and DepMSTAT do not fully capture spectral, spatial, and temporal EEG features alongside linguistic and paralinguistic cues. Moreover, none explicitly address EEG channel reduction for portability. In contrast, our proposed model integrates spectral–spatial–temporal EEG features with linguistic and paralinguistic interview data using synchronized multi-head cross-attention and

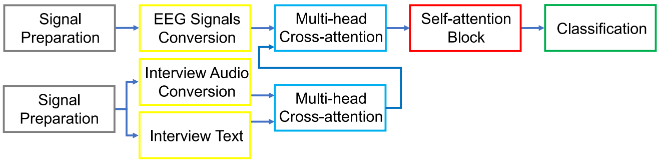


Fig. 1. A general overview of the proposed multimodal transformer model includes the Signal Preparation, Signal Conversion and Projection, Multi-head Cross-attention, Self-attention Block, and Classification.

self-attention, while introducing FTSM for model-agnostic EEG channel selection. This dual novelty addresses both fusion challenges and practical device constraints.

EEG devices typically include 128 or more channels, which hinders real-world usability. To overcome this, channel reduction methods have been explored. Zhang et al. [47] proposed weighting EEG channels to identify and remove low-value ones. Similarly, Wang et al. [38] applied gradient-weighted class activation mapping to rank channel importance. Shen et al. [48] combined kernel-target alignment with binary particle swarm optimization for optimal channel selection. A feature-aggregation-based cross-device transfer method was introduced by Li et al. [49], enabling improved generalization with fewer channels. Zhu et al. [50] applied mutual information to identify EEG channels most correlated with pupil area signals. These works demonstrate the promise of channel reduction, yet most are tied to specific models or feature extraction pipelines.

3. Methodology

An overview of the proposed multimodal transformer model is shown in Fig. 1, and a more detailed representation is given in Fig. 2. The model integrates EEG and interview data within a unified architecture, where EEG signals are transformed to preserve their spectral, spatial, and temporal properties, while the audio data contributes both linguistic and paralinguistic cues. By leveraging multi-head cross-attention, the framework captures both intra- and inter-modal correlations, allowing the integration of neural and behavioral cues to enhance depression detection.

The workflow of the model is organized into five main stages, illustrated in Fig. 1. First, in the *signal preparation stage*, raw data is preprocessed and EEG channel prioritization is performed. Second, during the *signal conversion and projection stage*, both EEG and audio signals are transformed into appropriate formats for subsequent processing. Third, the *multi-head cross-attention stage* fuses heterogeneous modalities, enabling the model to learn interactions between EEG and interview features. Fourth, the *self-attention block* further refines the representations by modeling dependencies within each modality. Finally, the *classification stage* assigns labels to distinguish between normal and depressed subjects.

These five steps build on one another in a sequential manner, ensuring a smooth flow from raw data acquisition to final prediction. The following subsections describe each stage in detail.

3.1. Signal preparation

3.1.1. Optimal EEG channels selection

To assess the system's real-world portability while maintaining acceptable levels of relative accuracy, the model is tested using various input scenarios, each corresponding to a different number of channels. To select the sample channels, for each subject, the similarity of signals from each channel with all other channels was measured using FTSM [51], and the same channels were identified. Then the channels were prioritized based on the repetition of similarity in different subjects. This makes it possible to evaluate the priority of channels independently of the model used for depression classification. As shown in Fig. 3, based on priority, four, eight, 16, 32, 64 and 128 channels were selected for analysis, respectively. For example, the four channels with numbers 67, 68, 93, and 94 have the highest priority which are marked in red. Next, the four orange channels, 48, 66, 82, and 84, are added to this set. Subsequently, eight yellow channels, 16 green channels, 32 blue channels, and 64 purple channels are included in the set of channels under examination.

Fig. 4 illustrates a sample scenario where flexible temporal sequence matching is compared with the fixed Euclidean distance method. In "Step-1" the green line indicates spatially similar points between the two signals, while the red points represent their differences. In "Step-2" additional red-filled points are inserted according to the differences so that the two signals can be aligned. The FTSM offers a more flexible and robust approach for comparing time series data, particularly when handling time shifts, distortions, and varying lengths, making it more suitable for real-world applications.

Details of the FTSM are provided in Algorithm 1. The FTSM is used to measure similarity between two temporal sequences that may vary in time or speed. It aligns the sequences in a non-linear fashion to minimize the distance between them. In the algorithm, let $X = (x_1, x_2, \dots, x_n)$ be the first sequence of length n, and $Y = (y_1, y_2, \dots, y_m)$ be the second sequence of length m. Create a cost matrix D of size $(n+1) \times (m+1)$ initialized to infinity (∞), except for the first cell $D(1,1)$, which is set to 0. This extra row and column allow for easier handling of boundary conditions. The cost matrix D is filled using a nested loop. As shown in Eq. (1), in each cell, $D(i,j)$ represents the minimum cumulative cost to align the first i elements of X with the first j elements of Y. The cost to align x_i with y_j is calculated using a distance function, typically the absolute difference or squared difference.

$$D(i,j) = d(x_i, y_j) + \min\{D(i-1,j), D(i,j-1), D(i-1,j-1)\} \tag{1}$$

where $d(x_i, y_j)$ is the distance between x_i and y_j. The nested loops iterate through each cell of the matrix D starting from $D(2,2)$. For each cell $D(i,j)$, calculate the cost and update the matrix using the formula mentioned above. After filling the matrix, the FTSM distance is found at $D(n+1, m+1)$. This value represents the minimum cumulative cost required to align the entire sequences X and Y. Starting from $D(n+1, m+1)$, trace back to $D(1,1)$ by choosing the path that led to the minimum cost in the dynamic programming step. This path represents the optimal alignment between the two sequences. Path$X$ and Path$Y$ indices of the elements from sequence X and Y that are aligned. The pair $(value, idx)$ finds the minimum value and its index among the three neighboring cells: above, left, and diagonal. The FTSM distance is set to the value at the bottom-right corner of the cost matrix D.

3.1.2. Interview audio data preparation

To prepare audio files for conversion to 2D representation and token extraction, a comprehensive data integrity check was performed. This process ensured compatibility of audio file formats, identified and excluded corrupted files, and standardized file names for consistency. Following this, volume normalization was applied across all audio files to maintain uniform loudness levels, enhancing the quality and consistency of subsequent processing steps. Finally, all audio files were resampled to 44.1 kHz sampling rate, ensuring uniformity in data structure and compatibility with the selected processing algorithms. These preparation steps set a robust foundation for accurate 2D generation and reliable token extraction.

3.2. Signal conversion and projection

3.2.1. EEG signal to 2D representation

Studies have shown that each aspect of spatial, spectral, and temporal dimensions in EEG data provides valuable information about individuals' emotions [52]. Simultaneous analysis of all three aspects

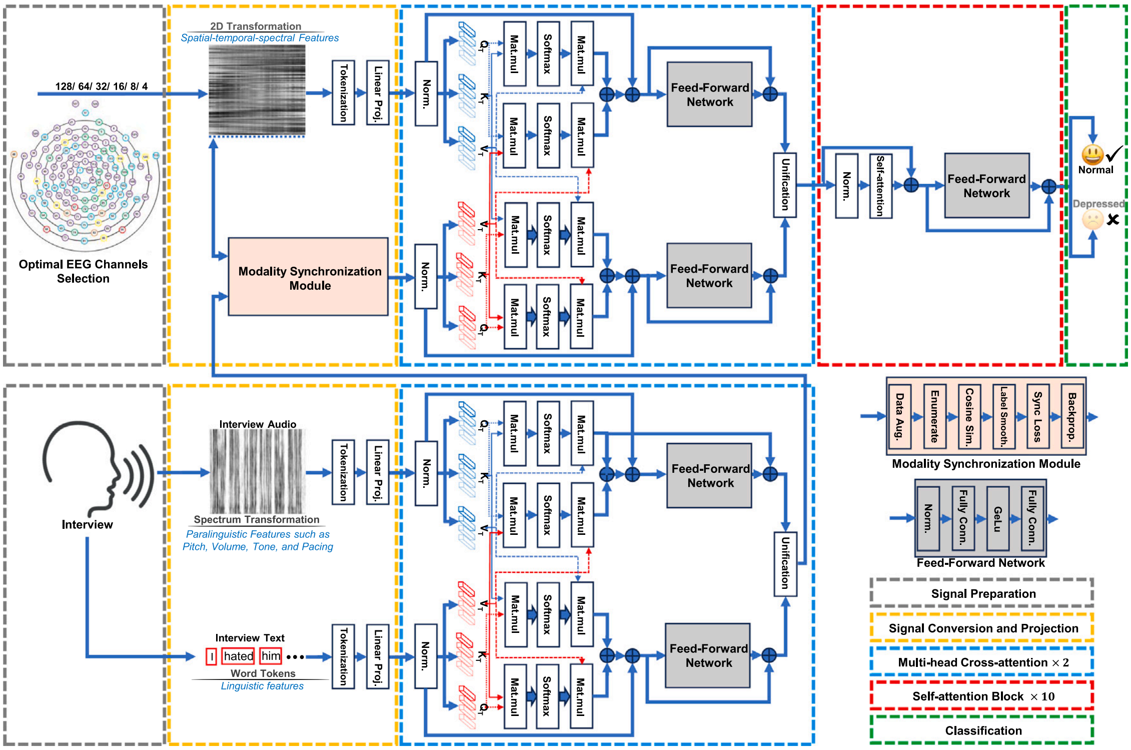


Fig. 2. The overall workflow of the proposed model for depression detection includes five main steps: Signal Preparation, Signal Conversion and Projection, Multi-head Cross-attention, Self-attention Block, and Classification.

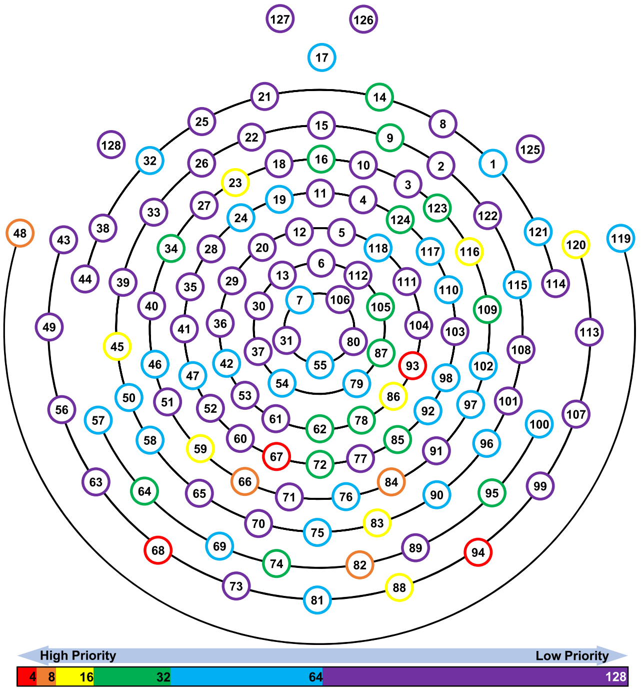


Fig. 3. Priority-based channel selection using flexible temporal sequence matching for four, eight, 16, 32, 64, and 128 channels.

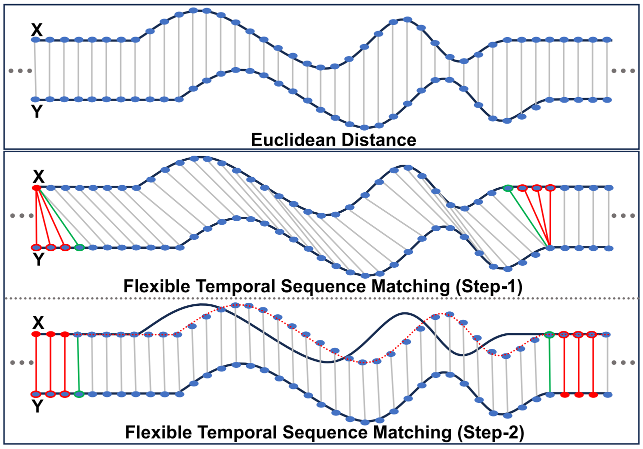

X
Y
Euclidean Distance
Flexible Temporal Sequence Matching (Step-1)
Flexible Temporal Sequence Matching (Step-2)

Fig. 4. Illustration of Euclidean distance and flexible temporal sequence matching for signal alignment. Step 1: Green lines show spatially similar points, and red points denote differences. Step 2: Red-filled points are added to align the two signals.

can help improve the accuracy of the model. In addition, by converting signals to images, different modalities are integrated together in the form of similar patches. For this purpose, EEG data have been transformed into images in such a way that all three aspects are preserved [37,53,54]. As shown in Fig. 5, the vertical axis represents channels (spatial) from top to bottom. Each channel receives five samples per second, which are converted into grayscale pixels (spectral) arranged in a row from left to right (temporal). This process creates an image approximately 70,000 pixels long. Inspired by ViT [55], the image is resized to 224 × 224 pixels and divided into 16 × 16 patches so that in later stages, both local and global features can be extracted from them to distinguish between normal and depressed states. Linear projection is applied, followed by position embedding on the tokens.

3.2.2. Interview audio file to words and 2D representation

In the analysis of emotions from interview data, we recognize the importance of both **linguistic and paralinguistic features.** To

Algorithm 1 Flexible Temporal Sequence Matching

```
Input: Sequences $X = (x_1, x_2, \dots, x_n)$ and $Y = (y_1, y_2, \dots, y_m)$
Output: FTSM distance and alignment path
Initialize cost matrix $D$ of size $(n+1) \times (m+1)$ with $\infty$
$D(1,1) \leftarrow 0$
for $i \leftarrow 2$ to $n+1$ do
    for $j \leftarrow 2$ to $m+1$ do
        $cost \leftarrow |x_{i-1} - y_{j-1}|$
        $D(i,j) \leftarrow cost + \min\{D(i-1,j), D(i,j-1), D(i-1,j-1)\}$
    end for
end for
$ftsm\_dist \leftarrow D(n+1, m+1)$
Initialize empty lists $pathX$ and $pathY$
$i \leftarrow n+1$
$j \leftarrow m+1$
while $i > 1$ and $j > 1$ do
    Insert $i-1$ at the beginning of $pathX$
    Insert $j-1$ at the beginning of $pathY$
    $(value, idx) \leftarrow \min\{D(i-1,j), D(i,j-1), D(i-1,j-1)\}$
    if $idx = 1$ then
        $i \leftarrow i-1$
    else if $idx = 2$ then
        $j \leftarrow j-1$
    else
        $i \leftarrow i-1$
        $j \leftarrow j-1$
    end if
end while
Insert $i-1$ at the beginning of $pathX$
Insert $j-1$ at the beginning of $pathY$
return $ftsm\_dist$, $pathX$, $pathY$
```

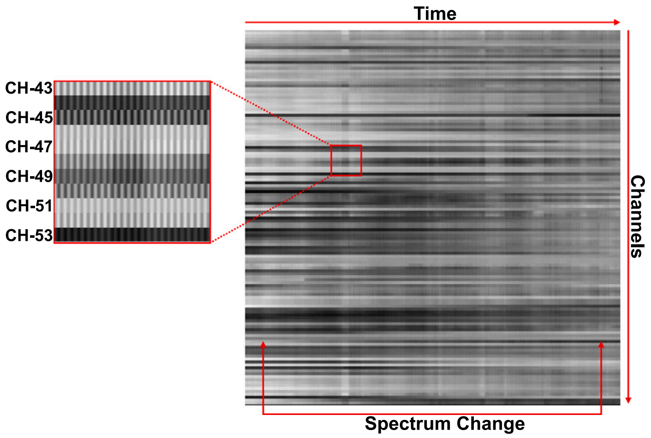


Fig. 5. Transforming EEG signals into images while preserving the spatial information from channels, the spectral details in the gray-scale spectrum, and the temporal aspects over time.

facilitate a comprehensive analysis, the audio input signal undergoes a two-stream preprocessing pipeline. For the **linguistic stream**, the raw audio file is first transcribed into text using an automatic speech recognition system. This transcript is then processed for linguistic analysis. Words are tokenized using a pre-trained tokenizer, which breaks the sentences into sub-word units or words. These tokens are then mapped to numerical embeddings, which are prepared as input for the multi-head cross-attention mechanism.

For the **paralinguistic stream**, we extract features that capture vocal nuances. The audio signal is first converted to a spectrogram using the **Short-Time Fourier Transform (STFT)**. This process involves dividing the signal into small, overlapping frames and computing the Fourier transform for each. The magnitude of the resulting complex values represents the intensity of different frequencies over time, effectively converting the audio signal into a 2D image-like representation. This spectrogram is then used to derive features such as pitch, volume, and spectral centroid, which serve as crucial indicators of emotional state. This 2D representation is then used as input for the cross-attention module, as depicted in Fig. 2.

The STFT of a signal $s(t)$ is defined in Eq. (2). Here, $s(p+q)$ represents the signal $s(t)$ sampled at time $p+q$, where q denotes the position of the current time frame (or window) and p indexes the time samples within each window of length L. The window function $\omega(p)$ is applied to the signal to isolate a segment for analysis, effectively "windowing" it. This ensures that the analysis focuses on a specific portion of the signal within each frame. The term $e^{-j\omega p}$ is a complex exponential representing a sinusoid with a frequency-dependent phase shift. Here, ω is the angular frequency (in radians per sample), and j is the imaginary unit ($j = \sqrt{-1}$). By Euler's formula, $e^{-j\omega p}$ can be expressed as $\cos(\omega p) - j\sin(\omega p)$. This enables the STFT to capture both the amplitude and phase of the frequency component ω for each time frame q.

$$\mathrm{STFT}\{s(t)\}(q,\omega) = \sum_{p=0}^{L-1} s(p+q)\omega(p)e^{-j\omega p} \tag{2}$$

Eq. (3) presents the spectrogram on a decibel scale, which visually represents the magnitude of the STFT. It shows how the spectrum of frequencies in a signal changes over time.

$$\mathrm{Spectrogram}_{dB}(q,\omega) = 20\log_{10}|\mathrm{STFT}\{s(t)\}(q,\omega)| \tag{3}$$

The $\mathrm{STFT}\{s(t)\}(q,\omega)$ is the short-time Fourier transform of the signal $s(t)$. It provides a complex number representing the magnitude and phase of the frequency component ω at the time frame q. $|\mathrm{STFT}\{s(t)\}(q,\omega)|$ is the magnitude of the STFT. The magnitude is calculated by taking the absolute value of the complex number obtained from the STFT, which gives the amplitude of the frequency component ω at the time frame q. Using the logarithm helps in compressing the dynamic range of the amplitudes, making it easier to visualize and interpret. Multiplying the logarithm by 20 converts the magnitude to decibels (dB). This is a common practice in signal processing to express the amplitude of signals on a logarithmic scale, which is more aligned with how humans perceive sound intensity. The spectrogram in dB thus provides a way to visualize how the frequency content of a signal changes over time, with the intensity in the spectrogram plot representing the amplitude of the frequency components in decibels. This helps in identifying patterns, harmonics, and other features in the signal that may not be evident in the time domain alone. To more effectively represent the audio spectrum of human speech, certain assumptions were made. The view range was set between 0 and 5000. Since the voice pitch range for humans is from 85 to 255, the display range for voice pitch was set between 75 and 265. The window length was set to 0.125 s [56,57].

3.3. Multi-head cross-attention

As shown in Fig. 2, in multi-head cross-attention block, a bidirectional multimodal attention mechanism is responsible for fusing modalities at the token level. The model performs both intra-modal and inter-modal attention operations. Intra-modal attention focuses on the connections between elements of the same type of modality, such as EEG images, interview audio files, or interview text. However, inter-modal attention addresses the connections between different modalities. With the help of these two mechanisms, the model can simultaneously extract local patterns and understand the relationships between these patterns across different modalities. As a result, a more comprehensive understanding of the data is created, which can lead to increased model accuracy. Specifically, cross attention is a mechanism that identifies and encodes the relationships between different modalities. It calculates attention scores between the interview audio and interview text modalities and separately examines the total of these connections with the related EEG signal images.

In our model, cross attention works by using attention mechanisms where the query (Q), key (K), and value (V) matrices come from different types of tokenized data, like Q_T. These cross-modal attention scores are then integrated with the intra-modal attention scores to create a unified representation that fuses information from both text and images. The data extracted from individuals' interviews and the EEG signals obtained from them can inherently be asynchronous. This issue can significantly affect the performance of the multi-head cross-attention mechanism. To mitigate the impact of this lack of synchronization, inspired by [58] a synchronization module is propose whose structure is presented in Algorithm 2

In the first step, it simulates real-world asynchronous conditions by randomly shifting the audio or EEG sequence within a defined range (e.g., [−r, r]). In the second step, for each possible audio-EEG (AV) sequence within the range, the model shifts either the audio or EEG and processes the concatenated sequence through the encoder to extract features. In the third step, the cosine similarity between the audio and EEG feature vectors is computed for each possible offset to measure the alignment quality. The formula computes the cosine similarity between the audio A and EEG sequence E over time t. A_t and E_t are the feature vectors for the audio and EEG at time t. $\cdot$ denotes the dot product of the vectors. $\|A_t\|$ and $\|E_t\|$ are the magnitudes (Euclidean norms) of the feature vectors. T is the total number of frames in the sequence. A higher cosine similarity value indicates better alignment between the audio and EEG at that specific offset. In the fourth step, the model smooths the labels for offset predictions to prevent harsh penalties for minor prediction deviations. In the fifth step, the similarities are converted into probabilities using softmax, and the cross-entropy loss is computed between the predicted and ground truth offsets. The softmax function, converts the cosine similarity scores into probabilities for each possible offset o. The $e^{similarity(o)}$ is the exponential of the cosine similarity for a particular offset o. The denominator $\sum_{o'} e^{\text{similarity}(o')}$ in similarity softmax formula is the sum of the exponentials of cosine similarities for all possible offsets. The cross-entropy loss is used to compute the difference between the predicted AE offset probabilities and the ground-truth labels. $p(o)$ is the predicted probability for each offset o, obtained from the softmax function. $label_o$ represents the ground-truth label for the offset, which is 1 for the correct offset and 0 for incorrect offsets (after smoothing). The loss quantifies how well the model's predicted offset probabilities align with the true offsets. Minimizing this loss improves the model's ability to handle AE synchronization. Label smoothing formula smooths the labels to avoid over-penalizing small deviations from the true AE offset. For the true offset, the label is reduced by a small factor ϵ (smoothing factor), and this value ϵ is distributed to adjacent offsets. This prevents the model from being overly confident in predicting the exact offset. Smoothing helps regularize the model, making it more tolerant to small deviations and improving its generalization performance. Finally, in the sixth step, the Sync Loss is used to update the model weights, improving the alignment of asynchronous audio and EEG data.

Algorithm 2 Modality Synchronization Module

```
Input: Audio sequence $A$, EEG sequence $E$, Offset range $r$, smoothing factor $\epsilon$
Output: Adjusted sync between audio and EEG, optimized Sync Loss
Initialization:
- Prepare encoder with initial weights.
- Set offset range $r$ (e.g., [-4,4] frames).
- Initialize loss function (e.g., cross-entropy for classification).
procedure SynchronizationModule($A$, $E$)
    Step 1: Data Augmentation
    - For each training example, artificially introduce AV offset:
            Shift $A$ or $E$ by random offset $\in [-r, r]$
    Step 2: Enumerate Over Possible Offsets
    for each offset $o \in [-r, r]$ do
        - Shift $A$ or $E$ by offset $o$.
        - Compute the fused sequence $F_o$ by concatenating shifted $A$ and $E$.
        - Pass $F_o$ through shared encoder to get encoded features $f_o$.
    end for
    Step 3: Calculate Cosine Similarity
    for each offset $o$ do
        - Calculate average cosine similarity between $A$ and $E$ for offset $o$:
$\text{similarity}(A, E) = \frac{1}{T} \sum_{t=1}^{T} \frac{A_t \cdot E_t}{\|A_t\| \|E_t\|}$
    end for
    Step 4: Apply Label Smoothing
    - For each ground truth offset, smooth the label distribution:
            Subtract smoothing factor $\epsilon$ from true label position and add to adjacent
positions.
    - This prevents harsh penalties for small deviations in offset predictions.
$\text{smoothed\_label}_o = \begin{cases} 1-\epsilon & \text{for the true offset,} \\ \epsilon & \text{for adjacent offsets,} \\ 0 & \text{for all other offsets.} \end{cases}$
    Step 5: Compute Sync Loss
    - Use softmax to convert cosine similarities to probabilities for each offset $o$:
$p(o) = \frac{e^{\text{similarity}(o)}}{\sum_{o'} e^{\text{similarity}(o')}}$
    - Calculate cross-entropy loss between predicted and ground truth offsets:
$L_{\text{sync}} = -\sum_{o} \text{label}_o \cdot \log(p(o))$
    Step 6: Backpropagation and Weight Update
    - Use $L_{\text{sync}}$ to update the encoder weights via backpropagation.
    - Adjust the model's attention and encoder to improve AE alignment.
end procedure
```

3.4. Self-attention block

As shown in Fig. 2, the self-attention block utilizes the self-attention mechanism, which calculates attention scores between all pairs of tokens within a single modality. These scores come from the query (Q), key (K), and value (V) matrices, which are representations of the input data transformed by learned weight matrices. The attention scores are computed using Eq. (4) [59]. In this equation, $\frac{QK^T}{\sqrt{d_k}}$ represents the scaled dot-product of the query and key matrices, normalized by the square root of the key vectors' dimensionality (d_k). Here, d_k is the size of the vector space for the key and query vectors. The softmax function ensures the scores are positive and sum to one. After calculating the attention scores, the outputs are processed through a residual connection, followed by normalization and a multi-layer perceptron (MLP). A residual connection (Z^{l+1}) adds the input to the layer (Z^l) to the output of the attention mechanism. This technique aids in training deep networks by providing a direct path for the gradient during backpropagation, addressing the vanishing gradient problem, and preserving information from the input while learning new features. This process can be summarized by Eq. (5). Variable Z^l denotes the input to the self-attention block at layer l. This variable is a matrix with each row representing a token's representation at that specific layer. Inspired by [53], we stack multiple self-attention blocks to iteratively enhance the input data representations. Each block enables the model to identify more intricate patterns and dependencies within the modality. Attention(Q, K, V) denotes the outcome of the self-attention mechanism for the specified layer. This result is computed utilizing the query (Q), key (K), and value (V) matrices, derived from the input Z^l. As depicted in Eq. (6), matrices W_K, W_Q and W_V serve as weight matrices employed to project the input embeddings into the key (K), query (Q), and value (V) vectors. Q is obtained from the input X^l by applying a weight matrix W_Q to it, representing the token currently under focus. Similarly, K is derived from Z^l by multiplying it with the weight matrix W_K. It depicts the tokens that used for comparison with the query. V is obtained from the input Z^l by the process of multiplying it with a weight matrix W_V. It encapsulates the values compiled from the attention scores. The self-attention mechanism calculates the level of attention each token should allocate to every other token utilizing these matrices. The result is a calculated total of the values (V), with the weights determined by the attention scores as defined in Eq. (4). Norm denotes layer normalization, a method aimed at stabilizing and expediting the training of deep neural networks by standardizing the inputs across the features for every token. It guarantees that the outputs maintain a consistent range of values across the network by ensuring they have an average of zero and a variance of one. After normalization, the data is processed by an MLP feed-forward neural network, which usually consists of two linear layers with a GeLu activation function in between. The model processes EEG data from the entire session along with interview data, using a self-attention block before the classification stage to fully capture the key information from both modalities. The MLP, alongside the self-attention mechanism, helps to identify patterns and relationships in the data before classification.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK_p^T}{\sqrt{d_k}}\right)V \tag{4}$$

$$Z^{l+1} = \text{MLP}(\text{Norm}(\text{Attention}(Q, K, V) + Z^l)) \tag{5}$$

$$Q = Z^l W_Q \quad \& \quad K = Z^l W_K \quad \& \quad V = Z^l W_V \tag{6}$$

3.5. Classification head

Classification is the final step of proposed model. The classification head receives a latent array that has been processed through multiple multimodal and self-attention layers. This array contains the combined features from both the image and text data. The classification head generates raw scores, or logits, for the ‘normal’ and ‘depressed’ classes. These logits are passed through a softmax function to calculate the probability for each class. During training, binary cross-entropy loss is used to compare the predicted scores with the actual labels.

4. Experimental results

4.1. Datasets and experimental setup

This study utilizes the multi-modal open dataset for mental-disorder analysis dataset (MODMA), and distress analysis interview corpus-wizard of oz (DAIC-WOZ). The MODMA is a public resource for major depressive disorder (MDD) research published by Lanzhou University. The MODMA includes 128-channel EEG recordings and audio data from 24 MDD patients and 29 healthy controls. MDD persons were recruited from a hospital setting and diagnosed by psychiatrists. Twenty of the participants (named subjects) were women and 33 were men. The age range of the subjects was 16 to 52 years. The audio data and EEG signals are recorded simultaneously during multimodal emotion elicitation experiments. Each session typically lasts for about 20–30 min, depending on the specific experimental setup and protocol used [60].

The DAIC-WOZ dataset is a widely used benchmark for detecting signs of psychological distress, particularly depression. It contains audio, video, and text data from clinical interviews, where a virtual interviewer (controlled by a human “wizard”) engages participants in a structured conversation. The dataset includes 189 sessions, with 142 used for training and 47 for testing. Each session contains transcribed text, audio recordings (16 kHz, 16-bit WAV format), and video (1920 × 1080 resolution), along with facial landmarks, prosodic features, and voice pitch variations. We are used audio files and transcripts in this research [61]. The comparison of channels for selecting channels and generating the image dataset was performed on a system with an Intel Core i7 CPU, 48.0 GB RAM, and an NVIDIA GeForce GTX 1080 graphics card. The implementation and evaluation of the model were conducted on Google Colab. The evaluation metrics included Accuracy, Precision, Recall, and F1-score.

4.2. Evaluation results

To assess the effectiveness of our proposed model, as shown in Table 2 we perform a quantitative comparison against several methods applied to the MODMA dataset, and DAIC-WOZ including HGP-SL [23], AM-GCN [24], SAGE [25], CGIPool [26], SGP-SL [27], MS^2-GNN [22], and G-Atten. According to the experimental results, our model has outperformed the state-of-the-art.

HGP-SL [23] developed a hierarchical graph pooling method that emphasizes structure learning to effectively summarize graph representations. Their approach leverages a structure learning mechanism that dynamically learns the graph structure during the pooling process, leading to enhanced performance in various graph-based tasks. The AM-GCN [24] employs adaptive multi-channel graph convolutional networks to capture the complex relationships in data, achieving notable results in various domains. SAGE [25] proposed a semi-supervised classification method to capture multi-scale structures within data, achieving high accuracy in classification tasks. CGIPool [26] utilizes a graph pooling approach to improve the infomax principle on coarsened graphs, achieving significant results in the task of graph classification. The [27] utilizes self-attention mechanism integrated with EEG-based topological structures and soft labels to enhance depression detection. MS^2-GNN [22] focus on fusing different modalities to improve detection accuracy. Features from each modality are extracted different techniques and then fused using a neural network, which enables the model to learn complex interactions between features from different modalities. Recently, [28] introduced an approach for depression detection based on audio signals using a GNN framework. This method first employs a gated recurrent unit (GRU) to capture time-series dependencies in audio features and then constructs two sequential graph neural networks. The first network models frame-level features within each audio sample, while the second one captures inter-sample relationships.

Our multimodal transformer model demonstrates significant advancements in depression detection, achieving the highest accuracy of 91.22% on the MODMA and 94.17% on the DAIC-WOZ dataset, outperforming previous models such as AM-GCN, MS^2-GNN, and G-Atten. To validate these improvements statistically, we conducted 10 independent runs of our model on both datasets. For MODMA, the mean accuracy was 91.20% with a standard deviation of 0.75% and a 95% confidence interval (CI) of [90.79%, 91.73%], while for DAIC-WOZ, the mean accuracy was 94.18% with a standard deviation of 0.71% and a 95% CI of [93.61%, 94.75%]. Paired t-tests revealed statistically significant improvements over all baseline models ($p < 0.0001$), except for G-Atten on MODMA, where p = 0.0006, indicating a robust performance advantage. By integrating EEG and audio data, our model captures a richer representation through spectro-temporal and linguistic features, leveraging advanced attention mechanisms and image transformation techniques. As will be discussed in Section 4.3, this approach not only ensures higher accuracy, precision, recall, and F1-score, respectively 91.22%, 92.34%, 90.15%, and 91.23% for MODMA, and 94.17%, 96.14%, 94.87%, and 95.50% for DAIC-WOZ, but also exhibits robustness and better generalization, particularly in real-world scenarios where some EEG channels might be missing. This highlights the model’s practical applicability and improved usability in constrained data acquisition environments.

4.3. Ablation study

4.3.1. Channel and modality selection using FSTM

To investigate the impact of reducing the number of channels, as performed by the FTSM algorithm, on measurement accuracy, the accuracy of the model was examined with four, eight, 16, 32, 64, and 128 channels. Additionally, the impact of each modality on the model’s accuracy when channel selection is performed was assessed. As shown in Fig. 6, including any of the modalities along with EEG significantly increases the model’s accuracy. The highest accuracy is achieved when all three EEG, audio, and text (E,A,T) modalities are considered. It is important to note that due to the large number of parameters in vision transformer models compared to traditional models like CNNs, along with the limited dataset size, there is a significant gap between training and validation performance, indicating overfitting. However, adding modalities and utilizing attention mechanisms and modality synchronization greatly reduces the extent of this overfitting. According to Fig. 7 EEG played the most important role in classification. Additionally, the audio file, which provides paralinguistic features, often plays a more significant role than the text, which contains linguistic features.

Table 2
Comparison of recent research studies that employed deep learning and transformer models to detect depression using MODMA and DAIC-WOZ datasets, ranked by accuracy.

Dataset	Model	ACC.%	PRE.%	REC.%	F1-.%	Stats (Mean ±Std, p-value)
MODMA	HGP-SL [23]	58.49	53.57	62.50	62.50	– ($p < 0.0001$ vs. Ours)*
	AM-GCN [24]	64.86	58.82	62.50	60.61	– ($p < 0.0001$ vs. Ours)*
	SAGE [25]	67.92	64.00	66.67	65.30	– ($p < 0.0001$ vs. Ours)*
	CGIPool [26]	73.58	69.23	75.00	72.00	– ($p < 0.0001$ vs. Ours)*
	SGP-SL [27]	84.91	80.77	87.50	84.00	– ($p < 0.0001$ vs. Ours)*
	MS^2-GNN [22]	86.49	82.35	87.50	84.85	– ($p < 0.0001$ vs. Ours)*
	G-Atten. [28]	90.35	88.25	**90.33**	89.15	– (p = 0.0006 vs. Ours)
	Ours	**91.22**	**92.34**	90.15	**91.23**	91.20 ± 0.75, CI 95% [90.79%, 91.73]
DAIC-WOZ	AM-GCN [24]	54.35	29.41	35.71	32.26	– ($p < 0.0001$ vs. Ours)*
	HGP-SL [23]	60.71	57.19	59.12	58.14	– ($p < 0.0001$ vs. Ours)*
	SAGE [25]	68.51	67.02	65.98	66.50	– ($p < 0.0001$ vs. Ours)*
	CGIPool [26]	74.19	70.79	72.28	71.53	– ($p < 0.0001$ vs. Ours)*
	SGP-SL [27]	79.63	78.90	80.88	79.88	– ($p < 0.0001$ vs. Ours)*
	MS^2-GNN [22]	80.43	64.71	78.57	70.97	– ($p < 0.0001$ vs. Ours)*
	G-Atten. [28]	92.21	92.36	92.18	92.23	– ($p < 0.0001$ vs. Ours)
	Ours	**94.17**	**96.14**	**94.87**	**95.50**	94.18 ± 0.71, CI 95% [93.61%, 94.75]

*Note: p-values for baseline models are based on paired t-tests using 10 runs, comparing the mean accuracy of ours (91.20% for MODMA, 94.18% for DAIC-WOZ) with the respective baseline accuracies. For MODMA, G-Atten p-value is 0.0006; for DAIC-WOZ, all p-values are < 0.0001.

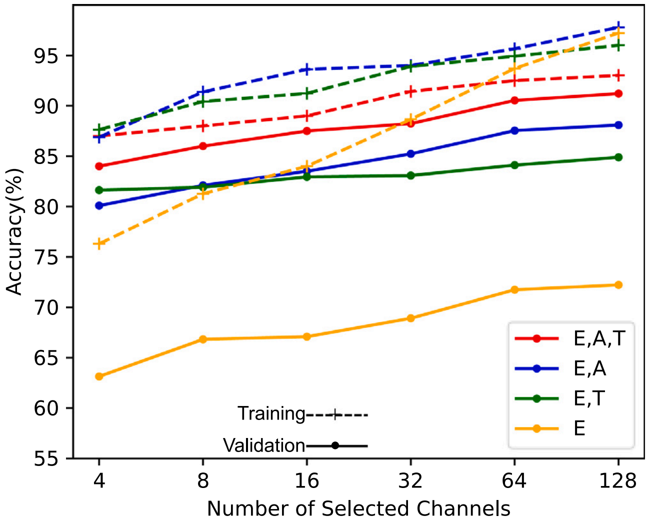


Fig. 6. The training and validation accuracy of the model on the MODMA dataset in different modalities, including EEG, audio, and text (E,A,T); EEG and audio (E,A); EEG and text (E,T); and EEG alone (E), across sets of four, eight, 16, 32, 64, and 128 channels.

4.3.2. *Multi-head cross-attention, modality synchronization, and self-attention*

The multi-head cross-attention block and self-attention block are vital components for enhancing the model's comprehension of hidden information across various modalities, as shown in Table 3. Removing either block leads to a significant decrease in accuracy, while the impact of the Modality Synchronization Module is comparatively smaller. When the multi-head cross-attention block is removed, a normalization and unification stage is introduced after the signal conversion and projection block. The data is then passed directly to the self-attention block. Cross-attention helps the model combine information from different modalities and capture their relationships, so removing it reduces or eliminates these interactions. When the self-attention block is removed, a normalization stage follows unification, leading directly to the classification stage. Self-attention helps the model understand long-term dependencies between input tokens. Removing it can weaken the model's ability to grasp these dependencies, especially in tasks that involve complex or long-term relationships.

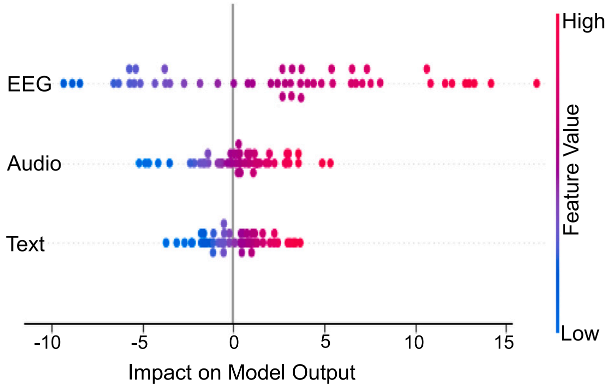


Fig. 7. Scatter plot of SHapley Additive exPlanations (SHAP) values illustrating the impact of EEG, Audio, and Text features on model predictions for the MODMA dataset.

Table 3
Ablation studies of the proposed model's multi-head cross-attention and self-attention blocks using the MODMA dataset, including the effect of the modality synchronization module on multi-head cross-attention.

Cross-att.	Sync. Modu.	Self-att.	ACC.%	PRE. %	REC.%	F1-.%
✗	✗	✗	58.52	56.12	59.62	57.83
✗	✗	✓	62.16	61.84	62.32	62.10
✓	✗	✗	64.63	63.92	65.02	64.46
✓	✗	✓	87.16	88.84	86.84	87.89
✓	✓	✗	75.08	78.12	76.10	77.10
✓	✓	✓	**91.22**	**92.34**	**90.15**	**91.23**

4.4. *Impact of hyperparameters setting*

The performance of a deep learning model is highly influenced by its hyperparameter settings. To develop a highly reliable deep learning model, it is essential to carefully optimize these hyperparameters. The grid search algorithm is a straightforward and efficient method for parameter optimization, commonly employed in hyperparameter tuning for deep learning models [37]. To optimize the hyperparameters of the proposed model, including learning rate, batch size, and maximum epoch, a grid search strategy is utilized. In particular, Fig. 8(a) shows that when the learning rate is set to 0.0001, the proposed method achieves the highest recognition accuracy. However, as the learning

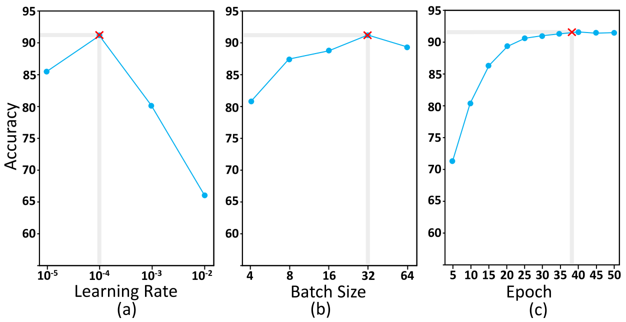


Fig. 8. Accuracy comparison with different hyperparameters. (a) Learning rate. (b) Batch size. (c) Epoch.

rate increases to 0.1, the recognition accuracy drops significantly because the learning rate is too high, causing the model to oscillate around the global optimum. Furthermore, lowering the learning rate to 0.00001 causes a decline in the model's recognition accuracy, as the reduced learning rate results in the model underfitting. Fig. 8(b) shows that with a batch size of 8, the proposed method achieves the lowest recognition accuracy. This is due to the instability in the gradient descent process caused by the smaller batch size. As the batch size grows, the model's performance improves, but once the batch size hits 32, the performance levels off. Fig. 8(c) shows that the model's performance improves up to epoch 38, suggesting that choosing 38 epochs is a sensible decision for this model.

4.5. *Channel attention analysis of EEG subjects*

In this section, we analyze channel attention by visualizing the channel weights learned by the proposed model. Fig. 9 displays the distribution of channel attention on the scalp for three healthy and three MDD subjects from the MODMA dataset. The red areas on the scalp indicate channels with large weights, while the blue areas indicate channels with small weights. The red areas are significantly smaller than the blue areas, suggesting that only a few EEG channels are strongly correlated with depression recognition, while many channels are irrelevant. Due to individual differences, no significant visual distinctions were observed between the healthy and MDD classes. However, relatively more scattered red spots were seen in depressed individuals. Fewer red spots occur in the front and left parts of the brain, while more red spots appear in the back of the head and the right side.

5. Conclusions

This study presents a synchronized multimodal transformer model that integrates EEG signals and interview data to enhance depression detection, extracting spectral, spatial, and temporal features from EEG via 2D mapping and linguistic/paralinguistic cues from audio. Employing self-attention and multi-head cross-attention mechanisms alongside a synchronization module, the model captures inter- and intra-modal correlations, achieving a 4.7% accuracy improvement and 10% precision boost on MODMA and DAIC-WOZ datasets compared to state-of-the-art methods. The FTSM algorithm optimizes EEG channel selection, reducing channels from 128 to 4 while maintaining 84% accuracy, thereby lowering costs and improving device portability. However, limitations include limited generalizability beyond evaluated datasets, susceptibility to real-world noise, and ethical concerns like data privacy and misdiagnosis risks.

Future Research Directions. Based on the current results, several directions can be explored: (i) evaluation on larger and more diverse datasets to improve generalization across populations, (ii) development of noise-robust and domain-adaptive methods to handle real-world recording artifacts, (iii) incorporation of privacy-preserving and fairness-aware learning frameworks such as federated learning to address ethical concerns, (iv) design of adaptive multimodal fusion strategies to remain effective when some modalities are missing, (v) real-time deployment using lightweight wearable EEG systems for clinical and telehealth applications, and (vi) personalized and longitudinal modeling for tracking depressive symptoms over time.

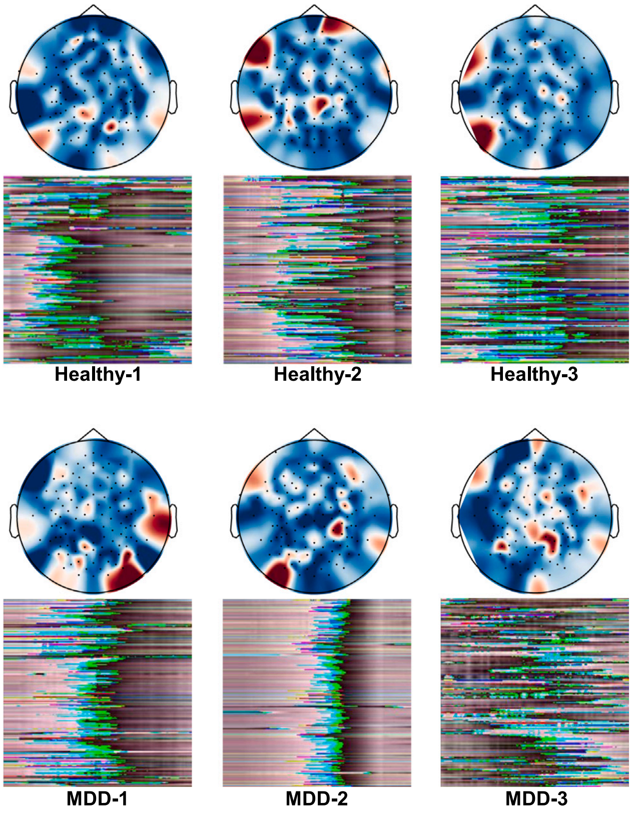


Fig. 9. Display of the channel attention distribution on the scalp surface and on the image related to EEG signals for three healthy subjects and three MDD subjects.

CRediT authorship contribution statement

Nima Esmi: Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Asadollah Shahbahrami:** Writing – review & editing, Validation, Supervision, Conceptualization. **Georgi Gaydadjiev:** Writing – review & editing. **Peter de Jonge:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] World-Health-Organization, Depressive disorder (depression), 2023.

[2] Guanyu Chen., Tianyi Shi, Baoxing Xie, Zhicheng Zhao, Zhu Meng, Yadong Huang, Jin Dong, SwinDAE: Electrocardiogram quality assessment using 1D swin transformer and denoising AutoEncoder, IEEE J. Biomed. Health Inf. 27 (12) (2023) 5779–5790, http://dx.doi.org/10.1109/JBHI.2023.3314698.

[3] Soheil Zabihi, Elahe Rahimian, Amir Asif, Arash Mohammadi, Trahgr: Transformer for hand gesture recognition via electromyography, IEEE Trans. Neural Syst. Rehabil. 31 (1) (2023) 4211–4224, http://dx.doi.org/10.1109/TNSRE.2023.3324252.

[4] Jialai Yin, Minchao Wu, Yan Yang, Ping Li, Fan Li, Wen Liang, Zhao Lv, Research on multimodal emotion recognition based on fusion of electroencephalogram and electrooculography, IEEE Trans. Instrum. Meas. 73 (1) (2024) 1–12, http://dx.doi.org/10.1109/TIM.2024.3370813.

[5] Seyed Reza Shahamiri, Vanshika Lal, Dhvani Shah, Dysarthric speech transformer: A sequence-to-sequence dysarthric speech recognition system, IEEE Trans. Neural Syst. Rehabil. 31 (1) (2023) 3407–3416, http://dx.doi.org/10.1109/TNSRE.2023.3307020.

[6] Hong Peng, Chen Xia, Zihan Wang, Jing Zhu, Xin Zhang, Shuting Sun, Jianxiu Li, Xiaoning Huo, Xiaowei Li, Multivariate pattern analysis of EEG-based functional connectivity: A study on the identification of depression, IEEE Access 7 (1) (2019) 92630–92641, http://dx.doi.org/10.1109/ACCESS.2019.2927121.

[7] Xiaowei Li, Xin Zhang, Jing Zhu, Wandeng Mao, Shuting Sun, Zihan Wang, Chen Xia, Bin Hu, Depression recognition using machine learning methods with different feature generation strategies, Artif. Intell. Med. 99 (1) (2019) 1–15, http://dx.doi.org/10.1016/j.artmed.2019.07.004.

[8] Subha D. Puthankattil, Paul K. Joseph, Classification of EEG signals in normal and depression conditions by ANN using RWE and signal entropy, Mech. Med. Biol. 12 (04) (2012) 1–13, http://dx.doi.org/10.1142/S0219519412400192.

[9] Behshad Hosseinifard, Mohammad Hassan Moradi, Reza Rostami, Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal, Comput. Meth. Prog. Bio 109 (3) (2013) 339–345, http://dx.doi.org/10.1016/j.aeue.2018.09.015.

[10] Emrah Aydemir, Turker Tuncer, Sengul Dogan, Raj Gururajan, U. Rajendra Acharya, Automated major depressive disorder detection using melamine pattern with EEG signals, Appl. Intell. 51 (9) (2021) 6449–6466, http://dx.doi.org/10.1007/s10489-021-02426-y.

[11] Maie Bachmann, Laura Päeske, Kaia Kalev, Katrin Aarma, Andres Lehtmets, Pille Ööpik, Jaanus Lass, Hiie Hinrikus, Methods for classifying depression in single channel eeg using linear and nonlinear signal analysis, Comput. Methods Progr. Biomed. 155 (1) (2018) 11–17, http://dx.doi.org/10.1016/j.cmpb.2017.11.023.

[12] Yalin Li, Bin Hu, Xiangwei Zheng, Xiaowei Li, EEG-based mild depressive detection using differential evolution, IEEE Access 7 (1) (2018) 7814–7822, http://dx.doi.org/10.1109/ACCESS.2018.2883480.

[13] Ah Young Kim, Eun Hye Jang, Seunghwan Kim, Kwan Woo Choi, Hong Jin Jeon, Han Young Yu, Sangwon Byun, Automatic detection of major depressive disorder using electrodermal activity, Sci. Rep. 8 (1) (2018) 1–9, http://dx.doi.org/10.1038/s41598-018-35147-3.

[14] Ayan Seal, Rishabh Bajpai, Jagriti Agnihotri, Anis Yazidi, Enrique Herrera-Viedma, Ondrej Krejcar, DeprNet: A deep convolution neural network framework for detecting depression using EEG, IEEE Trans. Instrum. Meas. 70 (1) (2021) 1–13, http://dx.doi.org/10.1109/TIM.2021.3053999.

[15] Pristy Paul Thoduparambil, Anna Dominic, Surekha Mariam Varghese, EEG-based deep learning model for the automatic detection of clinical depression, Phys. Eng. Sci. Med. 43 (4) (2020) 1349–1360, http://dx.doi.org/10.1007/s13246-020-00938-4.

[16] Vivek Sharma, Neelam Rup Prakash, Parveen Kalra, Depression status identification using autoencoder neural network, Biomed. Signal Process. Control. 75 (2022) 103568, http://dx.doi.org/10.1016/j.bspc.2022.103568.

[17] Surbhi Soni, Ayan Seal, Anis Yazidi, Ondrej Krejcar, Graphical representation learning-based approach for automatic classification of electroencephalogram signals in depression, Comput. Biol. Med. 145 (1) (2022) 1–13, http://dx.doi.org/10.1016/j.compbiomed.2022.105420.

[18] Daun Shin, Kyungdo Kim, Seung-Bo Lee, Changwoo Lee, Ye Seul Bae, Won Ik Cho, Min Ji Kim, C. Hyung Keun Park, Eui Kyu Chie, Nam Soo Kim, et al., Detection of depression and suicide risk based on text from clinical interviews using machine learning: possibility of a new objective diagnostic marker, Front. Psychiatry 13 (1) (2022) 1–11, http://dx.doi.org/10.3389/fpsyt.2022.801301.

[19] Jonathan F. Bauer, Maurice Gerczuk, Lena Schindler-Gmelch, Shahin Amiriparian, David Daniel Ebert, Jarek Krajewski, Bjö Schuller, Validation of machine learning-based assessment of major depressive disorder from paralinguistic speech characteristics in routine care, Depress. Anxiety 2024 (1) (2024) 1–12, http://dx.doi.org/10.1155/2024/9667377.

[20] Sara Sardari, Bahareh Nakisa, Mohammed Naim Rastgoo, Peter Eklund, Audio based depression detection using convolutional autoencoder, Expert Syst. Appl. 189 (1) (2022) 1–13, http://dx.doi.org/10.1016/j.eswa.2021.116076.

[21] Na Wang12, Raymond Chiong13, Raja Kamil, Weijia Zhang, Syed Abdul Rahman Al-Haddad, Normala Ibrahim, Depression detection using speech audio and text: A comprehensive review focusing on deep learning methods, Authorea (2024) 1–47, http://dx.doi.org/10.22541/au.172474607.71425235/v1.

[22] Tao Chen, Richang Hong, Yanrong Guo, Shijie Hao, Bin Hu, MS^2-GNN: Exploring GNN-based multimodal fusion network for depression detection, IEEE Trans. Cybern. 53 (12) (2023) 7749–7759, http://dx.doi.org/10.1109/TCYB.2022.3197127.

[23] Zhen Zhang, Jiajun Bu, Martin Ester, Jianfeng Zhang, Zhao Li, Chengwei Yao, Huifen Dai, Zhi Yu, Can Wang, Hierarchical multi-view graph pooling with structure learning, IEEE Trans. Knowl. Data Eng. 35 (1) (2023) 545–559, http://dx.doi.org/10.1109/TKDE.2021.3090664.

[24] Xiao Wang, Meiqi Zhu, Deyu Bo, Peng Cui, Chuan Shi, Jian Pei, Am-gcn: Adaptive multi-channel graph convolutional networks, in: Proc. Int. Conf. Knowl. Discov. Data Min, 2020, pp. 1243–1253, http://dx.doi.org/10.1145/3394486.3403177.

[25] Jia Li, Yu Rong, Hong Cheng, Helen Meng, Wenbing Huang, Junzhou Huang, Semi-supervised graph classification: A hierarchical graph perspective, in: WWW Conf., 2019, pp. 972–982, http://dx.doi.org/10.1145/3308558.3313461.

[26] Yunsheng Pang, Yunxiang Zhao, Dongsheng Li, Graph pooling via coarsened graph infomax, in: Proc. Int. ACM Conf. Res. Dev. Inf. Retriev., 2021, pp. 2177–2181, http://dx.doi.org/10.1145/3404835.3463074.

[27] Tao Chen, Yanrong Guo, Shijie Hao, Richang Hong, Exploring self-attention graph pooling with EEG-based topological structure and soft label for depression detection, IEEE Trans. Affect. Comput. 13 (4) (2022) 2106–2118, http://dx.doi.org/10.1109/TAFFC.2022.3210958.

[28] Chenjian Sun, Min Jiang, Linlin Gao, Yu Xin, Yihong Dong, A novel study for depression detecting using audio signals based on graph neural network, Biomed. Signal Process. Control. 88 (2024) 105675, http://dx.doi.org/10.1016/j.bspc.2023.105675.

[29] Nima Esmi, Asadollah Shahbahrami, Yasaman Nabati, Bita Rezaei, Georgi Gaydadjiev, Peter de Jonge, Stress detection through prompt engineering with a general-purpose LLM, Acta Psychol. 260 (2025) 105462, http://dx.doi.org/10.1016/j.actpsy.2025.105462.

[30] Loukas Ilias, Spiros Mouzakitis, Dimitris Askounis, Calibration of transformer-based models for identifying stress and depression in social media, IEEE Trans. Comput. Soc. Syst. 11 (2) (2023) 1979–1990, http://dx.doi.org/10.1109/TCSS.2023.3283009.

[31] Nima Esmi, Asadollah Shahbahrami, Georgi Gaydadjiev, Peter de Jonge, Suicide ideation detection based on documents dimensionality expansion, Comput. Biol. Med. 192 (2025) 110266, http://dx.doi.org/10.1016/j.compbiomed.2025.110266.

[32] Jing Zhu, Shiqing Wei, Xiannian Xie, Changlin Yang, Yizhou Li, Xiaowei Li, Bin Hu, Content-based multiple evidence fusion on EEG and eye movements for mild depression recognition, Comput. Methods Progr. Biomed. 226 (1) (2022) 1–11, http://dx.doi.org/10.1016/j.cmpb.2022.107100.

[33] Tao Chen, Richang Hong, Yanrong Guo, Shijie Hao, Bin Hu, MS^2-GNN: Exploring GNN-based multimodal fusion network for depression detection, IEEE Trans. Cybern. 53 (12) (2023) 7749–7759, http://dx.doi.org/10.1109/TMC.2022.3140430.

[34] Ziye Zhang, Aiping Liu, Yikai Gao, Xinrui Cui, Ruobing Qian, Xun Chen, Distilling invariant representations with domain adversarial learning for cross-subject children seizure prediction, IEEE Trans. Cogn. Dev. Syst. 16 (1) (2024) 202–211, http://dx.doi.org/10.1109/TCDS.2023.3257055.

[35] Xiaobing Du, Xiaoming Deng, Hangyu Qin, Yezhi Shu, Fang Liu, Guozhen Zhao, Yu-Kun Lai, Cuixia Ma, Yong-Jin Liu, Hongan Wang, MMPosE: Movie-induced multi-label positive emotion classification through EEG signals, IEEE Trans. Affect. Comput. 14 (4) (2023) 2925–2938, http://dx.doi.org/10.1109/TAFFC.2022.3221554.

[36] Mingyi Sun, Weigang Cui, Shuyue Yu, Hongbin Han, Bin Hu, Yang Li, A dual-branch dynamic graph convolution based adaptive transformer feature fusion network for EEG emotion recognition, IEEE Trans. Affect. Comput. 13 (4) (2022) 2218–2228, http://dx.doi.org/10.1109/TAFFC.2022.3199075.

[37] Jie Luo, Weigang Cui, Song Xu, Lina Wang, Xiao Li, Xiaofeng Liao, Yang Li, A dual-branch spatio-temporal-spectral transformer feature fusion network for EEG-based visual recognition, IEEE Trans. Ind. Inf. 20 (2) (2024) 1721–1731, http://dx.doi.org/10.1109/TII.2023.3280560.

[38] Yongling Xu, Yang Du, Ling Li, Honghao Lai, Jing Zou, Tianying Zhou, Lushan Xiao, Li Liu, Pengcheng Ma, AMDET: Attention based multiple dimensions EEG transformer for emotion recognition, IEEE Trans. Affect. Comput. 15 (3) (2024) 1067–1077, http://dx.doi.org/10.1109/TAFFC.2023.3318321.

[39] Yingdong Wang, Qingfeng Wu, Shuocheng Wang, XiQiao Fang, Qungsheng Ruan, MI-EEG: Generalized model based on mutual information for EEG emotion recognition without adversarial training, Expert Syst. Appl. 244 (1) (2024) 1–11, http://dx.doi.org/10.1016/j.eswa.2023.122777.

[40] Beilin Li, Jiao Wang, Zhifen Guo, Yue Li, Automatic detection of schizophrenia based on spatial–temporal feature mapping and LeViT with EEG signals, Expert Syst. Appl. 224 (1) (2023) 1–13, http://dx.doi.org/10.1016/j.eswa.2023.119969.

[41] Hao Sun, Yen-Wei Chen, Lanfen Lin, TensorFormer: A tensor-based multimodal transformer for multimodal sentiment analysis and depression detection, IEEE Trans. Affect. Comput. 14 (4) (2023) 2776–2786, http://dx.doi.org/10.1109/TAFFC.2022.3233070.

[42] Shiyu Teng, Jiaqing Liu, Yue Huang, Shurong Chai, Tomoko Tateyama, Xinyin Huang, Lanfen Lin, Yen-Wei Chen, An intra-and inter-emotion transformer-based fusion model with homogeneous and diverse constraints using multi-emotional audiovisual features for depression detection, IEICE Trans. Info. Syst. 107 (3) (2024) 342–353, http://dx.doi.org/10.1587/transinf.2023HCP0006.

[43] Huiting Fan, Xingnan Zhang, Yingying Xu, Jiangxiong Fang, Shiqing Zhang, Xiaoming Zhao, Jun Yu, Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals, Inf. Fusion. 104 (1) (2024) 1–11, http://dx.doi.org/10.1016/j.inffus.2023.102161.

[44] Yongfeng Tao, Minqiang Yang, Huiru Li, Yushan Wu, Bin Hu, DepMSTAT: Multimodal spatio-temporal attentional transformer for depression detection, IEEE Trans. Knowl. Data Eng. 36 (7) (2024) 2956–2966, http://dx.doi.org/10.1109/TKDE.2024.3350071.

[45] Meiling Li, Yuting Wei, Yangfu Zhu, Siqi Wei, Bin Wu, Enhancing multimodal depression detection with intra-and inter-sample contrastive learning, Inf. Sci. 684 (1) (2024) 1–15, http://dx.doi.org/10.1016/j.ins.2024.121282.

[46] Feiyu Zhu, Jing Zhang, Ruochen Dang, Bingliang Hu, Quan Wang, MTNet: Multimodal transformer network for mild depression detection through fusion of EEG and eye tracking, Biomed. Signal Process. Control. 100 (2025) 106996, http://dx.doi.org/10.1016/j.bspc.2024.106996.

[47] Yangting Zhang, Kejie Wang, Yu Wei, Xinwen Guo, Jinfeng Wen, Yuxi Luo, Minimal EEG channel selection for depression detection with connectivity features during sleep, Comput. Biol. Med. 147 (2022) 1–9, http://dx.doi.org/10.1016/j.compbiomed.2022.105690.

[48] Jian Shen, Xiaowei Zhang, Xiao Huang, Manxi Wu, Jin Gao, Dawei Lu, Zhijie Ding, Bin Hu, An optimal channel selection for EEG-based depression detection via kernel-target alignment, IEEE J. Biomed. Health Inf. 25 (7) (2020) 2545–2556, http://dx.doi.org/10.1109/JBHI.2020.3045718.

[49] Fang Liu, Pei Yang, Yezhi Shu, Niqi Liu, Jenny Sheng, Junwen Luo, Xiaoan Wang, Yong-Jin Liu, Emotion recognition from few-channel EEG signals by integrating deep feature aggregation and transfer learning, IEEE Trans. Affect. Comput. 15 (3) (2024) 1315–1330, http://dx.doi.org/10.1109/TAFFC.2023.3336531.

[50] Jing Zhu, Changlin Yang, Xiannian Xie, Shiqing Wei, Yizhou Li, Xiaowei Li, Bin Hu, Mutual information based fusion model (MIBFM): mild depression recognition using EEG and pupil area signals, IEEE Trans. Affect. Comput. 14 (3) (2022) 2102–2115, http://dx.doi.org/10.1109/TAFFC.2022.3171782.

[51] Jiaping Zhao, Laurent Itti, shapeDTW: Shape dynamic time warping, Pattern Recognit. 74 (1) (2018) 171–184, http://dx.doi.org/10.1016/j.patcog.2017.09.020.

[52] Weijian Mai, Fengjie Wu, Xiaoting Mai, Learning spatial–spectral–temporal EEG representations with dual-stream neural networks for motor imagery, Biomed. Signal Process. Control. 92 (2024) 106003, http://dx.doi.org/10.1016/j.bspc.2024.106003.

[53] Hong-Yu Zhou, Yizhou Yu, Chengdi Wang, Shu Zhang, Yuanxu Gao, Jia Pan, Jun Shao, Guangming Lu, Kang Zhang, Weimin Li, A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics, Nat. Biomed. Eng 7 (6) (2023) 743–755, http://dx.doi.org/10.1038/s41551-023-01045-x.

[54] Rui Li, Chao Ren, Sipo Zhang, Yikun Yang, Qiqi Zhao, Kechen Hou, Wenjie Yuan, Xiaowei Zhang, Bin Hu, STSNet: a novel spatio-temporal-spectral network for subject-independent EEG-based emotion recognition, Health Inf. Sci. Syst. 11 (1) (2023) 25, http://dx.doi.org/10.1007/s13755-023-00226-x.

[55] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, An image is worth 16x16 words: Transformers for image recognition at scale, 2020, http://dx.doi.org/10.48550/arXiv.2010.11929, arXiv Preprint.

[56] Jana Van Canneyt, Jan Wouters, Tom Francart, Enhanced neural tracking of the fundamental frequency of the voice, IEEE Trans. Biomed. Eng. 68 (12) (2021) 3612–3619, http://dx.doi.org/10.1109/TBME.2021.3080123.

[57] Toe Aung, David Puts, Voice pitch: a window into the communication of social power, Curr. Opin. Psychol. 33 (1) (2020) 154–161, http://dx.doi.org/10.1016/j.copsyc.2019.07.028.

[58] Jiahong Li, Chenda Li, Yifei Wu, Yanmin Qian, Unified cross-modal attention: Robust audio-visual speech recognition and beyond, IEEE/ACM Trans. Audio Speech Lang. Process. 32 (1) (2024) 1941–1953, http://dx.doi.org/10.1109/TASLP.2024.3375641.

[59] Andrea Galassi, Marco Lippi, Paolo Torroni, Attention in natural language processing, IEEE Trans. Neural Netw. Learn. Syst. 32 (10) (2020) 4291–4308, http://dx.doi.org/10.1109/TNNLS.2020.3019893.

[60] Hanshu Cai, Zhenqin Yuan, Yiwen Gao, Shuting Sun, Na Li, Fuze Tian, Han Xiao, Jianxiu Li, Zhengwu Yang, Xiaowei Li, et al., A multi-modal open dataset for mental-disorder analysis, Sci. Data 9 (1) (2022) 178, http://dx.doi.org/10.1038/s41597-022-01211-x.

[61] Fabien Ringeval, Bjö Schuller, State-of-mind, detecting depression with AI, and cross-cultural affect recognition, 2019, pp. 3–12, http://dx.doi.org/10.1145/3347320.3357688.