A Multimodal EEG-Eye Movement Model for Automatic Depression Detection

Hao-Long Yin[1], Jian-Ming Zhang[1], Ren-Jie Dai[1], Wei-Long Zheng[1], Qinyu Lv[23], Zhenghui Yi[23], Bao-Liang Lu[1*]

[1]School of Computer Science, Shanghai Jiao Tong University, China
[2]School of Medicine, Shanghai Jiao Tong University, China
[3]Shanghai Mental Health Center, China
{yinhaolong, jmzhang98, renjiedai, weilong, bllu}@sjtu.edu.cn, {lvqingyu_louis, yizhenghui1971}@163.com

Abstract

Depression is a prevalent mental health disorder characterized by persistent sadness and a diminished interest in daily activities, early detection of depression facilitates timely intervention, mitigating its adverse effects. Electroencephalography (EEG) signals and eye movements are emerging as promising biomarkers for depression detection due to their non-invasive nature and cost-effectiveness. Nevertheless, existing studies suffer from methodological constraints, including low specificity, insufficient sample sizes, limited generalizability, and difficulties in large-scale replication, which collectively undermine their clinical utility. To address these challenges, we collected a large-scale depression dataset comprising EEG and eye movements from 1,060 individuals diagnosed with depression and 1,308 healthy controls. To efficiently leverage multimodal data for automatic depression detection, we propose the **EEG-Eye Movements Model ($\mathbf{E^2Mo}$)**. E^2Mo employs modality-specific encoders to extract discriminative multi-view features from each modality and incorporates a mixture-of-modality-experts architecture with multi pretraining tasks to achieve efficient and robust modality alignment and fusion. Our approach achieves a **70.06% balanced accuracy** by leveraging multi-modal data, demonstrating the effectiveness of integrating EEG signals and eye movements for automatic depression detection.

Code — https://github.com/SJTU-BCMI/E2Mo

Introduction

Depression is a psychiatric syndrome featuring mood changes, emotional lability, disrupted rhythms, cognitive impairment, and abnormal motor behavior (Fava and Kendler 2000; Otte et al. 2016). As a prevalent mental disorder, it has significantly impaired the social functioning and overall well-being of affected individuals. According to World Health Organization (WHO), approximately 280 million people globally live with depression, representing about 3.8% of the world's population, while over 700,000 individuals die by suicide annually (Organization 2023).

Early and accurate diagnosis of depression is crucial for ensuring timely clinical intervention, which significantly benefits affected individuals. In clinical practice, depression diagnosis is challenging due to its heterogeneous nature, which not only limits the identification of specific genetic risk factors and results in inconsistent reliability of current diagnostic criteri but also introduces subjectivity and potential human errors that may lead to misdiagnosis and hinder the prediction of treatment response and disease course (Schatzberg 2019). Consequently, a variety of neural signals have been explored to objectively assess depression. Functional magnetic resonance imaging (fMRI) (Zeng et al. 2014; Magnin et al. 2009; Kerestes et al. 2014), functional near-infrared spectroscopy (fNIRS) (Ma et al. 2020; Chao et al. 2021; Husain et al. 2020) and electroencephalography (EEG) (de Aguiar Neto and Rosa 2019; Wang et al. 2024) have demonstrated promising results in depression detection.

Compared to alternative neural signal-based techniques, EEG offers distinct advantages, including ease of administration, high patient tolerance, and relatively low cost. Recent technological advancements have reinforced the reliability of EEG as a non-invasive method for investigating depression detection(Acharya et al. 2018; Saeedi et al. 2021; Hashempour et al. 2022). Meanwhile, eye movement analysis has demonstrated significant potential in augmenting depression detection from a different perspective(Alghowinem et al. 2013; Li et al. 2016, 2020). Studies have shown that eye movement data can provide complementary insights when combined with EEG in multimodal depression detection frameworks (Zhu et al. 2019, 2020). While existing studies on EEG or eye movement data based depression detection have produced encouraging results, limitations such as low specificity, small sample sizes, limited generalizability, and difficulties in extensive replication hinder their clinical applicability.

To address these challenges, we constructed a large-scale dataset comprising simultaneous EEG and eye movement recordings from 1060 individuals diagnosed with depression, 1,308 healthy controls, totaling 1,474 hours of signal recordings. We propose the EEG-Eye Movement Model (E^2Mo), a novel multimodal model designed for depression detection using EEG and eye movement data. The MoME framework enables efficient processing of different input modalities by leveraging modality-specific experts.

The primary contributions of this research are as follows:

*Corresponding author

- We present a novel large-scale multimodal depression dataset capturing EEG and eye movements. To the best of our knowledge, it represents the largest EEG-based depression dataset available, featuring the highest number of participants and the most extensive total recording duration.
- We propose the EEG-Eye Movement Model (E^2Mo), which utilizes modality-specific encoders to extract discriminative multi-view features from each modality. It incorporates a mixture-of-modality-experts architecture combined with multiple pretraining tasks to achieve efficient and robust alignment and fusion of modalities, achieved effective multimodal depression detection.
- We systematically evaluate E^2Mo against other classifiers under both unimodal and multimodal conditions through rigorous experimentation.

Related Work

EEG Based Depression Detection

Given the growing interest in EEG-based depression diagnosis, numerous detection methods have emerged. Consequently, high-quality and comprehensive EEG datasets are crucial for benchmarking algorithm performance.

Most existing datasets involve resting-state EEG recordings under eyes-open or eyes-closed conditions (Acharya et al. 2018; Uyulan et al. 2021; Mumtaz et al. 2017; Mumtaz and Qayyum 2019; Mao et al. 2018; Zhang et al. 2020). Some studies incorporate task-based paradigms to elicit more distinctive EEG patterns. For example, Cai *et al.* (Cai et al. 2016) used emotion-inducing soundtracks, Mumtaz *et al.* (Mumtaz et al. 2017) employed a 3-stimulus visual oddball task, and Xie *et al.* (Xie et al. 2020) designed an ϵ-stroop task with image stimuli. Table 1 compares our dataset with existing ones, our dataset represents a substantial advancement over existing EEG datasets for depression research. It features the largest sample size to date and incorporates a broader range of experimental tasks to capture more diverse and ecologically valid emotional responses.

Alongside data efforts, algorithmic advances have significantly improved depression detection from EEG (Seal et al. 2021; Ying et al. 2024). Sharma *et al.* (Sharma, Parashar, and Joshi 2021) proposed DepHNN, a CNN-LSTM hybrid achieving high accuracy with low complexity. Wan *et al.* (Wan et al. 2020) introduced HybridEEGNet, a dual-path CNN that extracts synchronous and regional EEG features from resting-state data.

Eye Movement Based Depression Detection

Eye movement abnormalities are well-documented in depression (Carvalho et al. 2015; Stolicyn, Steele, and Seriès 2022). Zhang *et al.* (Zhang et al. 2022) employed eye-tracking tasks (e.g., fixation stability, free-viewing, anti-saccade) with machine learning to extract discriminative features. Takahashi *et al.* (Takahashi et al. 2021) used scan path length and saccade velocity for classification via discriminant analysis.

Most current methods rely on hand-crafted features (e.g., saccade amplitude, fixation duration), which may overlook subtle or nonlinear biomarkers. In contrast, end-to-end deep learning can directly learn from raw signals, capturing richer depression-related patterns such as pupil dynamics and scan path anomalies.

Multimodal Depression Detection

Single-modality signals offer limited insight, while multimodal integration provides a more holistic representation for depression recognition. Scherer *et al.* (Scherer, Stratou, and Morency 2013) employed early fusion and factor analysis on audiovisual features (e.g., voice quality, facial expressivity) in virtual interviews. Zhu *et al.* (Zhu et al. 2025) introduced a Transformer-based mid-fusion model with attention bottlenecks to combine EEG and pupil area signals for mild depression detection.

Although multimodal methods consistently outperform single-modal approaches, few studies have explored the joint use of EEG and eye movement, despite their promise as complementary biomarkers.

Experiment Setup

Stimuli

The experiments were designed to record EEG and eye movement data simultaneously while participants performed a series of audiovisual tasks. Prior research on depression models indicates that individuals with depression exhibit distinct psychological reactions compared to healthy controls upon exposure to external stimuli (Mumtaz et al. 2017; Cai et al. 2016; Cavanagh et al. 2011; Xie et al. 2020). In this study, participants were exposed to a range of stimuli—including video clips, oil paintings, facial expressions, and paired-comparison emotional faces—to assess their varied responses. Each session began and ended with a resting-state period to serve as a baseline reference.

Video clips: Participants viewed video clips that elicited specific emotions. All video stimuli were standardized materials sourced from the emotion-eliciting database of the SEED dataset(Zheng and Lu 2015).

Facial expressions: Participants were presented with photos of positive or negative facial expressions. These stimuli underwent rigorous evaluation and selection by a panel of certified emotion recognition experts.

Oil paintings: Participants were presented with oil paintings representing diverse artistic styles, which were selected from the validated stimulus set in (Ma et al. 2024).

Paired-comparison emotional face (Paired-comparison): Participants viewed two facial photographs exhibiting contrasting emotional valence. These stimulus pairs were derived from the aforementioned facial expression database.

Resting state: During the baseline condition, participants maintained visual fixation on a white cross centered on a black background while remaining motionless.

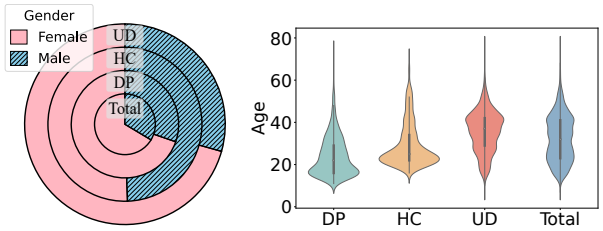
Subjects

The dataset consisted of 1,308 healthy controls (HC) and 1060 patients with depression (DP). Additionally, 4,068 cases remained untranscribed from paper-based diagnostic

Dataset	Task	DP No.	HC No.	Patient age	Channels No.
MODMA (Cai et al. 2020)	Resting, pictures	24	29	16-56	128
	Resting	26	29	16-56	3
REST-MDD (Wu et al. 2021)	Resting	200	200	51.2±17.8 (HC), 53.4±16.4 (DP)	26
HUSM (Mumtaz et al. 2017)	Resting	34	30	38.3±15.6 (HC), 40.3±12.9 (DP)	19
TDBRAIN (Van Dijk et al. 2022)	Resting, audio, pictures	426	47	38.7±19.2	19
Ours	Resting, video clips, oil paintings, facial expressions, paired-comparsion	**1060**	**1308**	23.9±9.3 (HC), 29.1±10.3 (DP)	18

Table 1: Comparison of EEG-based depression datasets in terms of task types, number of subjects, age, and EEG channels. DP: individuals with depression; HC: healthy controls.

records to electronic formats; these undiagnosed (UD) instances were exclusively utilized for pre-training purposes in this study. All DP participants were recruited from specialized psychiatric hospitals and received clinical diagnoses made by experienced psychiatrists in accordance with the DSM-5 diagnostic criteria (Edition et al. 2013). Strict inclusion and exclusion criteria were implemented, with comprehensive data collection including medical histories, primary complaints, and evaluation outcomes from 5 clinician-rated and 14 self-assessment scales. HC participants were recruited from the general population and underwent comprehensive psychiatric evaluations by the same chief physicians to confirm their mental health status. Demographic characteristics, including age and gender distributions, are presented in Figure 1.



(a) Gender distribution. (b) Age violin plots.

Figure 1: Demographic characteristics of study populations: Gender and age distributions. (a) Gender composition: Ring chart displaying female (pink) and male (blue) proportions across depressive patients (DP), healthy controls (HC), undiagnosed individuals (UD), and the combined cohort (Total). (b) Age profiles: Violin plots illustrating nonparametric age distributions, with median trends and density estimates. Sample sizes are annotated (DP: n=1,060; HC: n=1,308; UD: n=4068; Total: n=6436), demonstrating cohort-specific age dispersion patterns.

Protocol

To ensure high-quality data, experiments were conducted in a controlled laboratory setting to reduce noise and other environmental interferences. EEG and eye movement signals are simultaneously recorded using an 18-channel electrode cap with sensors placed according to the 10–20 electrode placement standard and a Tobii Pro Fusion eye tracker, respectively. EEG signals were captured at a sampling rate of 300 Hz, while eye-tracking signals were sampled at 250 Hz.

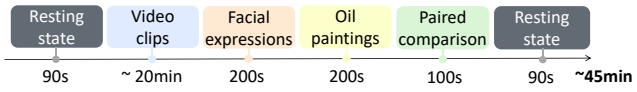


Figure 2: Overview of experimental procedure.

The experimental paradigm (illustrated in Figure 2) consisted of an approximately 45-minute standardized protocol. The session began with a 90-second baseline resting state recording. Participants then viewed six emotion-inducing video clips: three designed to evoke positive affect and three to elicit negative responses. Subsequently, they were exposed to 40 Western oil paintings, followed by 20 facial expression photographs (10 positive and 10 negative). The session continued with 10 paired-comparison emotional face trials before concluding with a final 90-second resting state period. Participants could self-determine the rest duration during intervals between task module switches.

This study was approved by the Scientific & Technical Ethics Committee of Shanghai Mental Health Center. Written informed consent was obtained from all participants prior to the study commencement.

Methods

In this section, we present the overall framework of the EEG-Eye Movement model (E^{2}Mo). We represent multichannel EEG signals as $X^G \in \mathbb{R}^{C^Y \times T}$, where C denotes the number of electrodes and T represents the number of time points. The eye movement (EYE) data are denoted as $X^Y \in \mathbb{R}^{C^Y \times T}$, where C^Y corresponds to the number of eye-related features. Specifically, the set of channels $\mathcal{C}^Y_{X^Y} = \{c^Y_1, \ldots, c^Y_4\}$ includes the left and right pupil diameters (c^Y_1, c^Y_2), as well as the horizontal and vertical gaze coordinates (c^Y_3, c^Y_4).

Model architecture

We introduce the E^{2}Mo, a multimodal architecture extract feature of EEG and EYE, align and fuse EEG and EYE for efficient multimodal automatic depression detection, as illustrated in Figure 3 (b). We first segment the EEG and

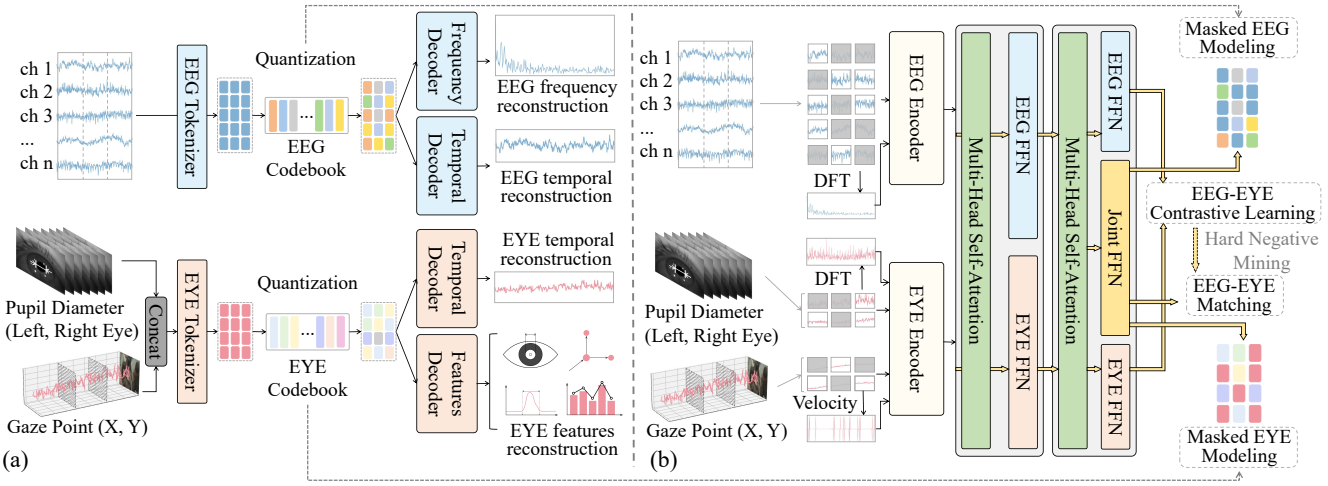


Figure 3: (a) Architecture of the EEG and eye movements (EYE) tokenizers. The EEG tokenizer discretizes signals by reconstructing both the frequency magnitude and raw EEG signals. The EYE tokenizer is trained to reconstruct the raw EYE signals along with handcrafted features. (b) Overview of the E^{2}Mo architecture. EEG and EYE signals are encoded separately, concatenated, and input into the MoME module for multimodal pretraining using three objectives: masked EEG-EYE modeling, contrastive learning, and modality matching, jointly aligning and fusing the two modalities.

EYE data into non-overlapping temporal patches using window sizes ω_G and ω_Y, respectively, i.e., $X^G = \left\{x^G_{i,k} \in \mathbb{R}^{\omega_G}\right\}_{\substack{i=1..C_G \\ k=1..\lfloor t_G/\omega_G \rfloor}}$, $X^Y = \left\{x^Y_{i,k} \in \mathbb{R}^{\omega_Y}\right\}_{\substack{i=1..C_Y \\ k=1..\lfloor t_Y/\omega_Y \rfloor}}$, where t_G and t_Y denote the total duration of EEG and EYE recordings, respectively. The condition $\omega_G s_Y = \omega_Y s_G$ ensures temporal alignment between the modalities, where s_G and s_Y are the sampling rates of the EEG and EYE data.

EEG Encoder EEG signals have been shown to contain numerous potential biomarkers for depression in both temporal and frequency domains (de Aguiar Neto and Rosa 2019). To comprehensively extract these temporal and spectral features, we propose a dual-pathway EEG encoder comprising separate temporal and spectral branches. Each branch consists of a series of convolutional blocks, each followed by group normalization and GELU activation (Hendrycks and Gimpel 2016). The temporal branch processes raw EEG time series, while the spectral branch operates on magnitude spectra obtained via discrete Fourier transform. The resulting feature maps from both branches are fused through element-wise addition.

EYE Encoder Eye movement data comprises sequences of pupil diameters and corresponding gaze points. The temporal and frequency characteristics of pupil diameter are closely linked to emotional processing and emotional states (Brendler et al. 2024; Nguyen et al. 2022), which are often disrupted in individuals with depression when exposed to internal or external stimuli. Therefore, we employ the same architecture used for the EEG encoder in the pupil diameter encoder to effectively capture both temporal and frequency-domain features.

Previous research has shown that individuals with depression tend to exhibit shorter exploration amplitudes, reduced saccade velocities, and ruminative scan paths (Takemoto et al. 2023). To capture these distinctive gaze patterns, we propose a gaze point encoder comprising a temporal encoder that models the sequential dynamics of gaze behavior and a velocity encoder that captures saccadic speed. The outputs of these two modules are combined via element-wise addition to generate the final gaze representation. Both encoders are constructed using a series of convolutional blocks.

Temporal, Channel and Modal-Type Embeddings After the EEG and EYE data are initially processed by their respective encoders, additional learnable [CLS] tokens are appended to the EEG and EYE patches. These patches are then added with temporal, channel and modality-type embeddings to form the final input representation.

Mixture-of-Modality-Experts Transformer Motivated by VLMo (Bao et al. 2022), we adopt a Mixture-of-Modality-Experts (MoME) Transformer as the backbone for multi-modal input processing. Unlike standard Transformers, MoME replaces the feed-forward network with modality-specific experts, dynamically selected by input type and layer index. For unimodal inputs (EEG-only or EYE-only), the corresponding expert (EEG expert or EYE expert) is used. For multimodal inputs (EEG-EYE pairs), lower layers employ separate EEG and EYE experts, while upper layers use a joint expert to capture cross-modal interactions. This expert-switching design enables efficient contextualized representations for both unimodal and multimodal inputs.

EEG and EYE Tokenizer

Prior to pre-training E^{2}Mo, we need to tokenize the EEG and eye movement (EYE) into discrete tokens. The structure of EEG and EYE tokenizer is shown in Figure 3 (a).

EEG Tokenizer The EEG tokenizer is composed of several vital components: VQ encoder, codebook, frequency decoder. The codebook $V_G \in \mathbb{R}^{K_G \times D_G}$ contains K_G discrete D_G-dimension embeddings. Let h_i^G denote the patch representations derived from the VQ encoder. We find the nearest codes of each h_i^G from codebook embeddings $\{v_i^G | i = 1, ..., K_G\}$:

$$z_i^G = \underset{j}{argmin} \|\ell_2(h_i^G) - \ell_2(v_i^G)\|_2,$$

where $j \in \{1, \ldots, K_G\}$ and ℓ_2-normalization is employed, ensuring that the above distance measure is equivalent to cosine similarity. Consequently, an EEG sample is tokenized as $z^G = [z_1^G, \ldots, z_N^G]$.

The EEG decoder predicts the frequency magnitude f^G and reconstructs original signals, the total loss defined as follows:

$$\mathcal{L}_{EEG} = \sum_{x \in \mathcal{D}} \sum_i \Big(\left\| o_i^f - f_i^G \right\|_2^2 + \left\| o_i^t - x_i^G \right\|_2^2 + \left\| \text{sg}(\ell_2(h_i^G)) - \ell_2(v_{z_i^G}^G) \right\|_2^2 + \left\| \ell_2(h_i^G) - \text{sg}(\ell_2(v_{z_i^G}^G)) \right\|_2^2 \Big), \tag{1}$$

where $o_i^{f^G}$ and o_i^t denote the reconstructed frequency magnitude and original signals, and f_i^G is the ground-truth frequency magnitude. where $\mathcal{D}$ represents the whole dataset and sg indicates the stop-gradient operation.

EYE Tokenizer The architecture of EYE tokenizer is similar to EEG tokenizer. The difference is that we propose to predict hand-crafted EYE features and original EYE signals to capture the key information. The hand-crafted EYE features can be formulated as $f^Y \in \mathbb{R}^N$, which can be extracted from a EYE sample $x_Y \in \mathbb{R}^{C^Y \times t_Y}$. The hand-crafted EYE features are described in detail in (Zheng et al. 2018). These features have shown outstanding performance in EYE-based classification tasks (Zheng et al. 2018; Fei et al. 2023).

Let o^f and o^t stand for the output of hand-crafted EYE features decoder and original EYE signals decoder, respectively. The optimizing target for the EYE codebook learning is:

$$\mathcal{L}_{EYE} = \sum_{x \in \mathcal{D}} \sum_i (\left\| o_i^f - f_i^Y \right\|_2^2 + \left\| o_i^t - x_i^Y \right\|_2^2 + \left\| sg(\ell_2(h_i^Y)) - \ell_2(v_{z_i^Y}^Y) \right\|_2^2 + \left\| \ell_2(h_i^Y) - sg(\ell_2(v_{z_i^Y}^Y)) \right\|_2^2), \tag{2}$$

h_i^Y denotes denote the patch representations derived from the EYE VQ encoder, $v_{z_Y}^Y$ denotes the nearest codes of z_i^Y from EYE codebook embeddings. Consequently, an EYE sample is tokenized as $z^Y = [z_1^Y, \ldots, z_N^Y]$.

Multimodal Pre-Training of E^{2}Mo

Masked EEG-EYE Modeling. To learn modality-agnostic representations and encourage EEG-EYE interaction, we randomly mask an r fraction of EEG (x^G) and EYE (x^Y) patches. Encoders convert each modality into embeddings e_i^G ($i \leq N_G$) and e_i^Y ($i \leq N_Y$), replacing masked positions with learnable tokens e_G^M, e_Y^M. The corrupted sequences $\mathbf{e}_G^{\mathcal{M}}, \mathbf{e}_Y^{\mathcal{M}}$ are processed by the MoME Transformer, whose outputs h_i are decoded through a linear head:

$$p(v' \mid e^{\mathcal{M}}) = \text{softmax}(\text{Linear}(h)).$$

The loss over masked tokens is

$$\mathcal{L}_{MEEM} = -\sum_{m_i=1} \left[\log p(v_i^G \mid \mathbf{e}_G^{\mathcal{M}}) + \log p(v_i^Y \mid \mathbf{e}_Y^{\mathcal{M}}) \right].$$

EEG–EYE Contrast. Given a batch of N paired samples, the `[EEG_CLS]` and `[EYE_CLS]` outputs are first projected and normalized to obtain $\hat{h}_i^G$ and $\hat{h}_i^Y$, respectively. The similarity between these representations is computed as

$$s_{i,j}^{G2Y} = \hat{h}_i^{G\top} \hat{h}_j^Y, \qquad s_{i,j}^{Y2G} = \hat{h}_i^{Y\top} \hat{h}_j^G,$$

from which the corresponding probability distributions are derived:

$$p_i^{G2Y} = \frac{\exp(s_{i,i}^{G2Y}/\sigma)}{\sum_j \exp(s_{i,j}^{G2Y}/\sigma)}, \; p_i^{Y2G} = \frac{\exp(s_{i,i}^{Y2G}/\sigma)}{\sum_j \exp(s_{i,j}^{Y2G}/\sigma)}.$$

The bidirectional contrastive loss is then defined using cross-entropy over the $N^2 - N$ negative pairs as

$$\mathcal{L}_{EEC} = -\frac{1}{N} \sum_{i=1}^{N} \left[\log p_i^{G2Y} + \log p_i^{Y2G} \right]. \tag{3}$$

EEG-EYE Matching. The same `[CLS]` vectors feed a binary classifier that predicts whether an EEG and an EYE originate from the same pair, using hard negatives selected as in ALBEF (Li et al. 2021).

Given the concatenated `[EEG_CLS]` and `[EYE_CLS]` embeddings, a binary classifier predicts a matching probability $q_{i,j}$, The matching loss is the binary cross-entropy (BCE) over positive and hard negative pairs:

$$\mathcal{L}_{Ma} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j \in \{i\} \cup \mathcal{N}(i)} \text{BCE}(y_{i,j}, q_{i,j}), \tag{4}$$

where $y_{i,j} = 1$ if (i, j) is a true pair and $y_{i,j} = 0$ otherwise, and $\mathcal{N}(i)$ denotes the set of hard negatives for i.

Notably, in the EEG-EYE Contrast and EEG-EYE Matching tasks, the input patches were not masked.

Stagewise Pre-Training We adopt a stagewise pre-training strategy similar to VLMo (Bao et al. 2022). The model is first pre-trained on EEG-only data with masked modeling task, which only mask and predict EEG patches, the EYE, joint expert are not used. We omit pre-training on EYE-only data, as we found it contributes minimally to the effectiveness of subsequent multimodal pre-training. The resulting model is then used to initialize joint EEG-EYE multimodal pre-training, thereby improving the alignment of multimodal signals.

Data Preprocessing

For the EEG data, a bandpass filter with cutoff frequencies of 0.1 Hz and 70 Hz is applied to remove low-frequency noise and power-line interference. Additionally, a 50 Hz notch filter is employed to specifically suppress power-line noise. To

Method	Bal. Acc.	AUC-PR	AUROC
HybridEEGNet	53.57 ± 1.12	49.27 ± 0.70	55.83 ± 0.76
SPaRCNet	57.91 ± 0.66	55.11 ± 2.05	62.80 ± 1.27
BIOT	58.34 ± 1.85	55.82 ± 1.53	62.61 ± 1.78
MMM	53.72 ± 1.73	48.68 ± 2.35	55.86 ± 3.19
Gram	57.71 ± 0.45	61.74 ± 0.68	62.18 ± 0.81
LaBraM	62.81 ± 0.39	61.51 ± 0.37	68.47 ± 0.25
E^2Mo	**62.96** ± 0.06	**63.37** ± 0.24	**69.19** ± 0.11
E^2Mo *w/o Freq. Rec.*	60.45 ± 0.24	59.20 ± 0.21	66.38 ± 0.18
E^2Mo *w/o Temp. Rec.*	62.22 ± 0.12	63.04 ± 0.25	68.83 ± 0.12
E^2Mo *w/o Spectral Enc.*	60.22 ± 0.44	60.83 ± 0.16	67.42 ± 0.13

Table 2: Performance (Balanced accuracy, AUC-PR and AUROC %) of different methods using EEG.

Subject Group	Train	Validation	Test
DP No.	848	106	106
HC No.	1096	106	106

Table 3: Number of subjects in train, validation, and test set.

improve computational efficiency, the raw EEG signals are downsampled from 300 Hz to 200 Hz.

Regarding the eye movement data, pupil diameter (left and right) and gaze coordinates (x and y) are directly extracted from the raw data without additional preprocessing.

Experimental Results

Implementation Details

Our EEG pre-training dataset consists of 1,266 hours of recordings from the depression dataset introduced in this study, with the remaining 208 hours (of 1,474 total) used for validation and testing. Both eye movement and paired EEG-eye movement data are drawn from the same 1,266-hour dataset.

The dataset uses non-overlapping 4-second windows to extract EEG and eye movement samples. To prevent data leakage, subject-wise partitioning is applied. Table 3 shows the distribution of healthy controls (HC) and depression patients (DP) across the training, validation, and test sets, split in an 8:1:1 ratio based on the smaller group. Excess subjects from the larger group are allocated to the training set to preserve label balance. During fine-tuning, the `[CLS]` token from the final output layer of E^2Mo is fed into a linear classifier for prediction. In the multimodal case, the `[EEG_CLS]` and `[EYE_CLS]` tokens are concatenated as input to the classifier. All model parameters are updated during fine-tuning; no layers are frozen. Results are averaged over four random seeds.

Uni-modal Depression Detection Result

Depression Detection Performance with EEG Data We systematically compare the depression detection performance of six baseline models and our proposed E^2Mo framework. The baselines include HybridEEGNet (Wan et al. 2020), SPaRCNet (Jing et al. 2023), BIOT (Yang, Westover, and Sun 2023), MMM (Yi et al. 2023), Gram (Li, Zheng, and Lu 2025), and LaBraM (Jiang, Zhao, and Lu 2024). Among them, HybridEEGNet is specifically designed for EEG-based depression detection, while the others are general-purpose EEG classification models. All methods are evaluated under identical training conditions to ensure a fair comparison.

E^2Mo is fine-tuned on EEG data after its EEG-only pre-training stage. As reported in Table 2, E^2Mo achieves the highest balanced accuracy of 62.96%, demonstrating its superior capability in leveraging EEG signals for depression detection.

To further analyze the contribution of each EEG-related module, we conduct an ablation study. E^2Mo *w/o Freq. Rec.* and E^2Mo *w/o Temp. Rec.* refer to the removal of frequency or temporal reconstruction during EEG tokenizer training. Both variants show performance degradation, confirming that reconstruction tasks significantly enhance the tokenizer's ability to represent EEG features.

Additionally, E^2Mo *w/o Spectral Enc.* removes the spectral encoder from the EEG encoder. The resulting performance drop highlights the importance of capturing EEG spectral information for depression detection.

Method	Bal. Acc.	AUC-PR	AUROC
CNNTransformer	55.70 ± 0.64	59.35 ± 0.51	65.22 ± 0.83
NST	62.87 ± 0.36	62.45 ± 0.81	67.96 ± 0.24
TCN	62.68 ± 0.13	61.03 ± 0.59	66.61 ± 0.16
TimesNet	62.62 ± 0.50	62.10 ± 0.30	67.78 ± 0.16
E^2Mo	**64.10** ± 0.11	**66.71** ± 0.03	**71.12** ± 0.25
E^2Mo *w/o Temp. Rec.*	63.90 ± 0.25	66.37 ± 0.33	70.67 ± 0.43
E^2Mo *w/o Feature Rec.*	64.05 ± 0.01	65.40 ± 0.02	69.77 ± 0.02
E^2Mo *w/o Spectral Enc.*	63.94 ± 0.11	66.58 ± 0.06	70.84 ± 0.32
E^2Mo *w/o Velocity Enc.*	63.97 ± 0.12	66.46 ± 0.16	70.97 ± 0.26

Table 4: Performance (Balanced accuracy, AUC-PR and AUROC %) of different methods using eye movements.

Depression Detection Performance with Eye Movements Since no existing method directly targets depression detection from raw eye movements, we select several state-of-the-art time-series classification models as baselines, including TCN (Bai, Kolter, and Koltun 2018), TimesNet (Wu et al. 2023), Non-stationary Transformer (stationary) (Liu et al. 2022), and Informer (Zhou et al. 2021). All baselines are trained from scratch under the same protocol.

E^2Mo is fine-tuned based on a model that was initially pre-trained on EYE-only data using a masked modeling task. As shown in Table 4, E^2Mo consistently outperforms all baseline models, validating its ability to extract discriminative features from eye movements for depression detection.

An ablation study is also performed to examine the impact of EYE-related components. E^2Mo *w/o Feature Rec.* and *w/o Temp. Rec.* indicate the removal of hand-crafted EYE feature reconstruction and temporal reconstruction respectively during EYE tokenizer training. Both lead to noticeable performance drops, highlighting their contribution to effective eye movements representation learning.

Method	Bal. Acc.	AUC-PR	AUROC
BDAE	64.71 ± 0.17	65.15 ± 0.34	70.41 ± 0.16
DGCCA-AM	64.42 ± 0.30	64.00 ± 0.63	69.93 ± 0.72
VigilanceNet	64.80 ± 0.37	66.30 ± 0.28	70.03 ± 0.37
MAET	65.42 ± 0.09	66.13 ± 0.34	70.82 ± 0.18
E^2Mo	**70.06** ± 0.04	**73.10** ± 0.03	**78.59** ± 0.01

Table 5: Performance (Balanced accuracy, AUC-PR and AUROC %) of different methods using multimodality.

Furthermore, we evaluate the contributions of specific modules. E^2Mo *w/o Pupil Spectral Enc.* and *w/o Velocity Enc.* denote the removal of the pupil spectral encoder and gaze velocity encoder respectively. The degraded performance in both cases underscores the importance of these modules in modeling fine-grained eye dynamics relevant to depression detection.

Multi-modal Depression Detection Result

Table 5 presents the depression detection performance of various multimodal models utilizing both EEG and eye movements data. The baselines include Bimodal Deep AutoEncoder (BDAE) (Liu, Zheng, and Lu 2016), Deep Generalized Canonical Correlation Analysis with Attention Mechanism (DGCCA-AM) (Lan, Liu, and Lu 2020), VigilanceNet (Cheng et al. 2022), and Multimodal Adaptive Emotion Transformer (MAET) (Jiang et al. 2023), all of which have demonstrated strong performance in multimodal EEG and eye movements classification tasks.

E^2Mo is fine-tuned on paired EEG-Eye movements data after its multi-modal pre-training.

Our proposed E^2Mo is fine-tuned from a multimodal pretrained model. As shown in the results, E^2Mo achieves the best overall performance, with a balanced accuracy of 70.06%, AUC-PR of 73.10%, and AUROC of 78.59%. These results clearly demonstrate the effectiveness of our model in capturing and integrating information from both modalities.

Furthermore, multimodal fusion significantly improves detection accuracy compared to unimodal settings. Specifically, E^2Mo outperforms its EEG-only and eye-only counterparts by margins of 7.10% and 5.96% in balanced accuracy, respectively. This highlights the complementary nature of EEG and eye movements and supports the potential of multimodal approaches in advancing depression detection.

Ablation on Multimodal Pre-Training Task We perform an ablation study on the multimodal pre-training task of E^2Mo, as shown in Table 6, to assess the contributions of Masked EEG-EYE Modeling (MEEM), EEG-EYE Contrastive Learning (EEC), and EEG-EYE Matching (EEM). Since EEM depends on hard negative samples generated by EEC, removing EEC necessitates the exclusion of EEM as well. The results demonstrate that removing any of these tasks significantly degrades multimodal depression detection performance, with MEEM having the most pronounced impact. These findings underscore the effectiveness of each task in enhancing the alignment and integration of EEG and eye movement modalities, thereby enhancing their complementarity in multimodal depression detection.

MEEM	EEC	EEM	Bal. Acc.	AUC-PR	AUROC
✓	✓	✓	**70.06** ± 0.04	**73.10** ± 0.03	**78.59** ± 0.01
✓	✓	×	68.71 ± 0.07	72.94 ± 0.03	77.16 ± 0.03
×	✓	✓	66.18 ± 0.08	68.24 ± 0.11	74.06 ± 0.08
×	✓	×	65.09 ± 0.07	68.39 ± 0.07	73.07 ± 0.08
✓	×	×	68.34 ± 0.08	71.79 ± 0.04	76.47 ± 0.00

Table 6: Ablation study on Masked EEG-EYE Modeling (MEEM), EEG-EYE Contrastive Learning (EEC), and EEG-EYE Matching (EEM) tasks.

Task name	Bal. Acc.	AUC-PR	AUROC
Resting state	71.22 ± 0.11	75.74 ± 0.26	80.42 ± 0.17
Video clips	70.29 ± 0.14	73.74 ± 0.08	79.24 ± 0.09
Oil paintings	66.19 ± 0.28	65.92 ± 0.30	72.46 ± 0.16
Facial expressions	**74.88** ± 0.60	**84.43** ± 0.30	**86.08** ± 0.26
Paired-comparison	70.84 ± 0.29	79.42 ± 0.56	81.35 ± 0.32

Table 7: Performance (Balanced accuracy, AUC-PR and AUROC %) of different experiment tasks.

Depression Detection Performance of Experiment Tasks

We evaluate the performance of depression detection across various experimental tasks. The E^2Mo model is trained on the complete training set under optimal conditions, while the test set is divided into five subsets corresponding to distinct experimental tasks. The results are summarized in Table 7. All tasks demonstrated promising effectiveness in detecting depression. The facial expressions task achieved the highest performance, while the Oil paintings task yielded relatively poorer performance.

Conclusion

In this paper, We propose E^2Mo, a novel method that employs modality-specific encoders to extract multi-view discriminative features from modalities and incorporates a mixture-of-modality-experts architecture with multi-task pretraining to enable efficient and robust fusion and alignment of EEG and eye movements modalities. We also developed a large-scale multimodal dataset for automatic depression detection, comprising EEG signals and eye movements. Ablation studies validate the efficacy of the proposed pre-training tasks. Our model achieves a balanced accuracy of 70.06% in multimodal depression detection, underscoring the benefits of integrating EEG signals and eye movements. E^2Mo establishes a robust baseline for future research in depression detection. Furthermore, we evaluate the performance across various experimental tasks, offering valuable insights for designing automatic depression detection paradigms.

Acknowledgments

This work was supported in part by grants from STI 2030-Major Projects+2022ZD0208500, National Natural Science Foundation of China (62376158), Brain Science and Brain-like Intelligence Technology-National Science and Technology Major Project (2025ZD0218900), Medical-Engineering Interdisciplinary Research Foundation of Shanghai Jiao Tong University "Jiao Tong Star" Program (YG2023ZD25, YG2024ZD25 and YG2024QNA03), the Lingang Laboratory (Grant No. LGL-1987), Shanghai Jiao Tong University 2030 Initiative, GuangCi Professorship Program of RuiJin Hospital Shanghai Jiao Tong University School of Medicine, and Shanghai Jiao Tong University SCS-Shanghai Emotionhelper Technology Co., Ltd Joint Laboratory of Affective Brain-Computer Interfaces.

References

Acharya, U.; Oh, S.; Hagiwara, Y.; Tan, J.; Adeli, H.; and Subha, D. 2018. Automated EEG-based screening of depression using deep convolutional neural network. *Comput. Methods Programs Biomed.*, 161: 103.

Alghowinem, S.; Goecke, R.; Wagner, M.; Parker, G.; and Breakspear, M. 2013. Eye movement analysis for depression detection. In *IEEE Int. Conf. Image Process. (ICIP)*, 4220.

Bai, S.; Kolter, J.; and Koltun, V. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Bao, H.; Wang, W.; Dong, L.; Liu, Q.; Mohammed, O.; Aggarwal, K.; Som, S.; Piao, S.; and Wei, F. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Adv. Neural Inf. Process. Syst.*, 35: 32897.

Brendler, A.; Schneider, M.; Elbau, I.; Sun, R.; Nantawisarakul, T.; Pöhlchen, D.; Brückl, T.; et al. 2024. Assessing hypo-arousal during reward anticipation with pupillometry in patients with major depressive disorder: replication and correlations with anhedonia. *Sci. Rep.*, 14(1): 344.

Cai, H.; Gao, Y.; Sun, S.; Li, N.; Tian, F.; Xiao, H.; Li, J.; Yang, Z.; Li, X.; Zhao, Q.; et al. 2020. Modma dataset: A multi-modal open dataset for mental-disorder analysis. *arXiv preprint arXiv:2002.09283*.

Cai, H.; Sha, X.; Han, X.; Wei, S.; and Hu, B. 2016. Pervasive EEG diagnosis of depression using Deep Belief Network with three-electrodes EEG collector. In *IEEE Int. Conf. Bioinformatics Biomed. (BIBM)*, 1239.

Carvalho, N.; Laurent, E.; Noiret, N.; Chopard, G.; Haffen, E.; Bennabi, D.; and Vandel, P. 2015. Eye movement in unipolar and bipolar depression: A systematic review of the literature. *Front. Psychol.*, 6: 1809.

Cavanagh, J.; Bismark, A.; Frank, M.; and Allen, J. 2011. Larger error signals in major depression are associated with better avoidance learning. *Front. Psychol.*, 2: 331.

Chao, J.; Zheng, S.; Wu, H.; Wang, D.; Zhang, X.; Peng, H.; and Hu, B. 2021. fNIRS evidence for distinguishing patients with major depression and healthy controls. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 29: 2211.

Cheng, X.; Wei, W.; Du, C.; Qiu, S.; Tian, S.; Ma, X.; and He, H. 2022. VigilanceNet: Decouple intra-and inter-modality learning for multimodal vigilance estimation in RSVP-based BCI. In *ACM Int. Conf. Multimedia*, 209.

de Aguiar Neto, F.; and Rosa, J. 2019. Depression biomarkers using non-invasive EEG: A review. *Neurosci. Biobehav. Rev.*, 105: 83.

Edition, F.; et al. 2013. Diagnostic and statistical manual of mental disorders. *Am. Psychiatr. Assoc.*, 21(21): 591.

Fava, M.; and Kendler, K. 2000. Major depressive disorder. *Neuron*, 28(2): 335.

Fei, C.; Li, R.; Zhao, L.; Zheng, W.; and Lu, B. 2023. EEG-eye movements cross-modal decision confidence measurement with generative adversarial networks. In *IEEE/EMBS Int. Conf. Neural Eng. (NER)*, 1.

Hashempour, S.; Boostani, R.; Mohammadi, M.; and Sanei, S. 2022. Continuous scoring of depression from EEG signals via a hybrid of convolutional neural networks. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 30: 176.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Husain, S.; Yu, R.; Tang, T.; Tam, W.; Tran, B.; Quek, T.; Hwang, S.; Chang, C.; Ho, C.; and Ho, R. 2020. Validating a functional near-infrared spectroscopy diagnostic paradigm for Major Depressive Disorder. *Sci. Rep.*, 10(1): 9740.

Jiang, W.; Liu, X.; Zheng, W.; and Lu, B. 2023. Multimodal adaptive emotion transformer with flexible modality inputs on a novel dataset with continuous labels. In *ACM Int. Conf. Multimedia*, 5975.

Jiang, W.; Zhao, L.; and Lu, B. 2024. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. In *Int. Conf. Learn. Represent. (ICLR)*.

Jing, J.; Ge, W.; Hong, S.; Fernandes, M.; Lin, Z.; Yang, C.; An, S.; Struck, A.; Herlopian, A.; Karakis, I.; et al. 2023. Development of expert-level classification of seizures and rhythmic and periodic patterns during EEG interpretation. *Neurology*, 100(17): e1750.

Kerestes, R.; Davey, C.; Stephanou, K.; Whittle, S.; and Harrison, B. 2014. Functional brain imaging studies of youth depression: a systematic review. *NeuroImage Clin.*, 4: 209.

Lan, Y.; Liu, W.; and Lu, B. 2020. Multimodal emotion recognition using deep generalized canonical correlation analysis with an attention mechanism. In *IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, 1.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Adv. Neural Inf. Process. Syst.*, 34: 9694.

Li, M.; Cao, L.; Zhai, Q.; Li, P.; Liu, S.; Li, R.; Feng, L.; Wang, G.; Hu, B.; and Lu, S. 2020. Method of depression classification based on behavioral and physiological signals of eye movement. *Complexity*, 2020(1): 4174857.

Li, X.; Cao, T.; Sun, S.; Hu, B.; and Ratcliffe, M. 2016. Classification study on eye movement data: Towards a new approach in depression detection. In *IEEE Congr. Evol. Comput. (CEC)*, 1227.

Li, Z.; Zheng, W.; and Lu, B. 2025. Gram: A Large-Scale General EEG Model for Raw Data Classification and Restoration Tasks. In *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 1.

Liu, W.; Zheng, W.; and Lu, B. 2016. Emotion recognition using multimodal deep learning. In *Neural Inf. Process. (ICONIP)*, 521.

Liu, Y.; Wu, H.; Wang, J.; and Long, M. 2022. Non-stationary Transformers: Exploring the Stationarity in Time Series Forecasting. In *Adv. Neural Inf. Process. Syst.*

Ma, T.; Liu, L.; Zhao, L.; Peng, D.; Lu, Y.; Zheng, W.; and Lu, B. 2024. Detecting Major Depression Disorder with Multiview Eye Movement Features in a Novel Oil Painting Paradigm. In *IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, 1.

Ma, T.; Lyu, H.; Liu, J.; Xia, Y.; Qian, C.; Evans, J.; Xu, W.; Hu, J.; Hu, S.; and He, S. 2020. Distinguishing bipolar depression from major depressive disorder using fnirs and deep neural network. *Prog. Electromagn. Res.*, 169: 73.

Magnin, B.; Mesrob, L.; Kinkingnéhun, S.; Pélégrini-Issac, M.; Colliot, O.; Sarazin, M.; Dubois, B.; Lehéricy, S.; and Benali, H. 2009. Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. *Neuroradiology*, 51: 73.

Mao, W.; Zhu, J.; Li, X.; Zhang, X.; and Sun, S. 2018. Resting state EEG based depression recognition research using deep learning method. In *Brain Inform. (BI)*, 329.

Mumtaz, W.; and Qayyum, A. 2019. A deep learning framework for automatic diagnosis of unipolar depression. *Int. J. Med. Inform.*, 132: 103983.

Mumtaz, W.; Xia, L.; Mohd Yasin, M.; Azhar Ali, S.; and Malik, A. 2017. A wavelet-based technique to predict treatment outcome for major depressive disorder. *PLoS One*, 12(2): e0171409.

Nguyen, K.; Liang, W.; Juan, C.; and Wang, C. 2022. Time-frequency analysis of pupil size modulated by global luminance, arousal, and saccade preparation signals using Hilbert-Huang transform. *Int. J. Psychophysiol.*, 176: 89.

Organization, W. H. 2023. Depressive disorder (depression).

Otte, C.; Gold, S.; Penninx, B.; Pariante, C.; Etkin, A.; Fava, M.; Mohr, D.; and Schatzberg, A. 2016. Major depressive disorder. *Nat. Rev. Dis. Primers*, 2(1): 1.

Saeedi, A.; Saeedi, M.; Maghsoudi, A.; and Shalbaf, A. 2021. Major depressive disorder diagnosis based on effective connectivity in EEG signals: a convolutional neural network and long short-term memory approach. *Cogn. Neurodynamics*, 15(2): 239.

Schatzberg, A. 2019. Scientific issues relevant to improving the diagnosis, risk assessment, and treatment of major depression. *Am. J. Psychiatry*, 176(5): 342.

Scherer, S.; Stratou, G.; and Morency, L. 2013. Audiovisual behavior descriptors for depression assessment. In *ACM Int. Conf. Multimodal Interact.*, 135.

Seal, A.; Bajpai, R.; Agnihotri, J.; Yazidi, A.; Herrera-Viedma, E.; and Krejcar, O. 2021. DeprNet: A deep convolution neural network framework for detecting depression using EEG. *IEEE Trans. Instrum. Meas.*, 70: 1.

Sharma, G.; Parashar, A.; and Joshi, A. 2021. DepHNN: A novel hybrid neural network for electroencephalogram (EEG)-based screening of depression. *Biomed. Signal Process. Control*, 66: 102393.

Stolicyn, A.; Steele, J.; and Seriès, P. 2022. Prediction of depression symptoms in individual subjects with face and eye movement tracking. *Psychol. Med.*, 52(9): 1784.

Takahashi, J.; Hirano, Y.; Miura, K.; Morita, K.; Fujimoto, M.; Yamamori, H.; Yasuda, Y.; Kudo, N.; Shishido, E.; Okazaki, K.; et al. 2021. Eye movement abnormalities in major depressive disorder. *Front. Psychiatry*, 12: 673443.

Takemoto, A.; Aispuriete, I.; Niedra, L.; and Dreimane, L. 2023. Depression detection using virtual avatar communication and eye tracking. *J. Eye Mov. Res.*, 16(2): 10.

Uyulan, C.; Ergüzel, T.; Unubol, H.; Cebi, M.; Sayar, G.; Nezhad Asad, M.; and Tarhan, N. 2021. Major depressive disorder classification based on different convolutional neural network models: Deep learning approach. *Clin. EEG Neurosci.*, 52(1): 38.

Van Dijk, H.; Van Wingen, G.; Denys, D.; Olbrich, S.; Van Ruth, R.; and Arns, M. 2022. The two decades brainclinics research archive for insights in neurophysiology (TDBRAIN) database. *Sci. Data*, 9(1): 333.

Wan, Z.; Huang, J.; Zhang, H.; Zhou, H.; Yang, J.; and Zhong, N. 2020. HybridEEGNet: A convolutional neural network for EEG feature learning and depression discrimination. *IEEE Access*, 8: 30332.

Wang, Y.; Peng, Y.; Han, M.; Liu, X.; Niu, H.; Cheng, J.; Chang, S.; and Liu, T. 2024. GCTNet: a graph convolutional transformer network for major depressive disorder detection based on EEG signals. *J. Neural Eng.*, 21(3): 036042.

Wu, C.; Huang, H.; Huang, S.; Chen, I.; Liao, S.; Chen, C.; Lin, C.; Lee, S.; Chen, M.; Tsai, C.; et al. 2021. Resting-state EEG signal for major depressive disorder detection: A systematic validation on a large and diverse dataset. *Biosensors*, 11(12): 499.

Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; and Long, M. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *Int. Conf. Learn. Represent. (ICLR)*.

Xie, Y.; Yang, B.; Lu, X.; Zheng, M.; Fan, C.; Bi, X.; Li, Y.; et al. 2020. Anxiety and depression diagnosis method based on brain networks and convolutional neural networks. In *IEEE Eng. Med. Biol. Soc. (EMBC)*, 1503.

Yang, C.; Westover, M.; and Sun, J. 2023. Biot: Biosignal transformer for cross-data learning in the wild. *Adv. Neural Inf. Process. Syst.*, 36: 78240.

Yi, K.; Wang, Y.; Ren, K.; and Li, D. 2023. Learning topology-agnostic EEG representations with geometry-aware modeling. *Adv. Neural Inf. Process. Syst.*, 36: 53875.

Ying, M.; Shao, X.; Zhu, J.; Zhao, Q.; Li, X.; and Hu, B. 2024. EDT: An EEG-based attention model for feature learning and depression recognition. *Biomed. Signal Process. Control*, 93: 106182.

Zeng, L.; Shen, H.; Liu, L.; and Hu, D. 2014. Unsupervised classification of major depression using functional connectivity MRI. *Hum. Brain Mapp.*, 35(4): 1630.

Zhang, D.; Liu, X.; Xu, L.; Li, Y.; Xu, Y.; Xia, M.; Qian, Z.; Tang, Y.; Liu, Z.; Chen, T.; et al. 2022. Effective differentiation between depressed patients and controls using discriminative eye movement features. *J. Affect. Disord.*, 307: 237.

Zhang, X.; Li, J.; Hou, K.; Hu, B.; Shen, J.; and Pan, J. 2020. EEG-based depression detection using convolutional neural network with demographic attention mechanism. In *IEEE Eng. Med. Biol. Soc. (EMBC)*, 128.

Zheng, W.; Liu, W.; Lu, Y.; Lu, B.; and Cichocki, A. 2018. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Trans. Cybern.*, 49(3): 1110.

Zheng, W.; and Lu, B. 2015. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Trans. Auton. Ment. Dev.*, 7(3): 162.

Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. In *AAAI Conf. Artif. Intell.*, volume 35, 11106.

Zhu, J.; Li, Y.; Yang, C.; Cai, H.; Li, X.; and Hu, B. 2025. Transformer-based fusion model for mild depression recognition with EEG and pupil area signals. *Med. Biol. Eng. Comput.*, 1.

Zhu, J.; Wang, Y.; La, R.; Zhan, J.; Niu, J.; Zeng, S.; and Hu, X. 2019. Multimodal mild depression recognition based on EEG-EM synchronization acquisition network. *IEEE Access*, 7: 28196.

Zhu, J.; Wang, Z.; Gong, T.; Zeng, S.; Li, X.; Hu, B.; Li, J.; Sun, S.; and Zhang, L. 2020. An improved classification model for depression detection using EEG and eye tracking data. *IEEE Trans. Nanobiosci.*, 19(3): 527.