TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 pp1–19
DOI: 10.26599/TST.2025.9010147
Volume x, Number x, xxxx xxxx

Research Article

Cross-Modal Knowledge Distillation for Depression Recognition: An Explainability Method with EEG and Pupil Area Signals

Yizhou Li†, Yu Cheng†, Xiaowei Li, Jing Zhu*, and Bin Hu*

Abstract: Multimodal physiological signals provide a more reliable data source for depression detection. For instance, combining electroencephalography (EEG) and pupil area (PA) signals can enhance depression recognition. However, EEG acquisition is challenging, limiting the practical use of EEG-based multimodal approaches, while PA signals are more accessible. Additionally, while existing explainability methods for time series models can quantify the contribution of each feature, they often fail to provide a comprehensive understanding of how these contributions drive performance improvements, limiting insights into the underlying mechanisms. To address these limitations and enhance the generalizability of PA-based depression detection models, this paper proposes a cross-modal knowledge distillation method, using an EEG and PA-based multimodal teacher model and a PA-based unimodal student model. Through knowledge distillation, complex multimodal features are transferred to the PA-based model, enhancing its performance. We also introduce Entropy-GradCAM (E-GCAM), an explainability method combining information entropy and gradient-weighted class activation mapping (Grad-CAM), to clarify mechanisms behind the student model's performance gains. Quantitative results show that knowledge-distilled time series models encode more useful information, consistent with observed student model improvements. Experimental results demonstrate that the proposed method achieves optimal performance on two datasets, effectively reducing reliance on multimodal data and increasing the practicality of depression recognition models.

Key words: electroencephalography (EEG); pupil area (PA); knowledge distillation; multimodal fusion; depression detection; explainability method

1 Introduction

Depression is a common yet serious mood disorder that has a profound impact on millions of people worldwide. Its clinical manifestations are diverse, including low mood, loss of interest[1], changes in appetite, and difficulty concentrating[1], often accompanied by sleep disturbances[2], anxiety[3], and chronic pain[4, 5]. The prolonged presence of these symptoms not only severely impacts patients' daily lives but can also lead to a decline in social functioning and even an increased risk of suicide[6]. Therefore, timely and accurate recognition of depression is crucial

• Yizhou Li, Yu Cheng, Xiaowei Li, and Jing Zhu are with Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Lanzhou 730099, China. E-mail: liyzh2023@lzu.edu.cn; chengyu2023@lzu.edu.cn; lixwei@lzu.edu.cn; zhujing@lzu.edu.cn.
• Bin Hu is also with Laboratory of Special Functional Materials and Structural Design (Ministry of Education), Lanzhou University, Lanzhou 730099, China, and also with Engineering Research Center of Open Source Software and Real-Time System (Ministry of Education), Lanzhou University, Lanzhou 730099, China. E-mail: bh@lzu.edu.cn.
† Yizhou Li and Yu Cheng contribute equally to this paper.
* To whom correspondence should be addressed.
Manuscript received: 2025-01-13; revised: 2025-06-16; accepted: 2025-09-16

for effective intervention.

Currently, the assessment of depression primarily relies on clinical interviews and self-report questionnaires. However, these methods have limitations, such as high subjectivity and susceptibility to patients' personal perceptions and social desirability bias. Studies showed that even experienced clinicians find it challenging to accurately identify depression[7, 8]. Therefore, finding a more objective and effective assessment method holds significant clinical importance. To address the limitations of traditional methods, an increasing number of studies began focusing on using physiological signals for depression detection, as these signals provide objective data that are unaffected by subjective factors and allow continuous monitoring of an individual's physiological changes. Among these, electroencephalography (EEG) signals have attracted considerable attention due to their non-invasive nature and high temporal resolution[9–11]. However, EEG data based on a single modality suffer from incomplete information in depression recognition, making it difficult to fully capture the multidimensional characteristics of depression, thereby limiting diagnostic accuracy. Therefore, combining EEG signals with other physiological signals, such as pupil area (PA) signals, to construct a multimodal model holds the potential to fully capture the characteristics of depression and improve recognition accuracy and robustness.

Multimodal physiological signals can make depression detection more accurate and objective, but data acquisition poses certain challenges[12]. In particular, EEG signals are easily affected by factors such as environmental noise and electromyographic interference, and the acquisition process requires multiple electrodes, which demands high-quality equipment. Additionally, subjects need to remain still for extended periods during the experiment, which may impact the feasibility of data collection. In contrast, eye-tracking data acquisition is relatively simple, with stable signal quality and easy-to-operate equipment, making it highly practical. Studies[13, 14] found that the PA or pupil response of patients with depression differs significantly from that of the normal control group, indicating the potential value of PA signals in depression detection. However, unimodal models solely based on PA data often lag behind multimodal models in terms of recognition accuracy and robustness. Therefore, it is necessary to figure out how to achieve satisfactory depression recognition results using only the PA modality during testing after leveraging multimodal information.

Although existing methods have advanced the explainability of deep models, each has its limitations. Gradient explanations[15] are sensitive to noise. Class activation mapping (CAM)[16] depends on specific architectures. Attention maps[17], constrained by particular network layers, can only indicate the model's focus at different time points. Shapley additive explanations (SHAP)[18] is computationally expensive. local interpretable model-agnostic explanations (LIME)[19] may provide different explanation results in different runs. Existing methods can quantify each time point's contribution to the model output, but they often struggle to provide a clear understanding of how these contributions influence performance improvements, hindering deeper insights into the effectiveness of knowledge transfer. Moreover, the lack of universal standards in current explainability methods complicates their application to diverse time series tasks, making it difficult to accurately assess the contribution of individual features.

To address these challenges, this paper proposes a multimodal teacher model based on EEG and PA data, alongside a unimodal student model that only uses PA data. Using knowledge distillation techniques, we achieve the transfer of rich feature knowledge from the multimodal teacher model to the unimodal student model. Additionally, we introduce Entropy-GradCAM (E-GCAM), a novel time series explainable method that integrates information entropy theory with the gradient-weighted class activation mapping (Grad-CAM) technique. This method quantifies and visualizes feature contributions in time series models, providing insight into the mechanisms underlying the performance improvements achieved through knowledge distillation. The contributions of this paper are as follows:

- **Proposed cross-modal knowledge distillation framework:** We propose an EEG and PA-based multimodal model that explores and integrates the intrinsic features and interactions between the EEG and PA modalities through a cross-modal interaction module and an adaptive modality fusion module, ensuring the transfer of critical knowledge during the knowledge distillation process.
- **Development of the E-GCAM explainable method:** We propose a method that combines

information entropy with Grad-CAM to quantify feature contributions in time series models. Results indicate that knowledge distillation enables the student model to encode more effective information.

- **Experiments on two depression recognition datasets** demonstrate that our method significantly enhances the PA-based unimodal model's performance and that E-GCAM effectively quantifies feature contributions, clarifying the mechanism behind performance improvements in time series models under knowledge distillation.

2 Related Work

2.1 Multimodal depression recognition

Traditional affective computing methods primarily rely on unimodal data, such as text[20], audio[21], or physiological signals (e.g., EEG, electrodermal activity)[9, 22]. A unimodal approach typically offers advantages such as simplicity of implementation and low computational cost. However, its performance is often constrained by the quality of the data, and a single data source is insufficient to comprehensively capture the complex characteristics associated with depressive states. In recent years, multimodal approaches have gained widespread attention in depression recognition. Multimodal approaches, by integrating information from multiple data sources, not only provide more comprehensive feature representations but also leverage complementary information across different modalities. This improves the accuracy and robustness of depression recognition, demonstrating potential advantages over unimodal methods.

In the early stages, studies on depression recognition typically relied on handcrafted features combined with traditional machine learning classifiers. In recent years, end-to-end deep learning approaches have gained increasing popularity, particularly models based on attention mechanisms and the Transformer architecture, which demonstrate superior capabilities in feature extraction and representation learning. Liu et al.[23] used a residual network and a convolutional neural network (CNN) to extract features from facial expressions and pupil diameter, further extracting high-dimensional features through attention networks, and finally concatenating them for depression recognition. Chen et al.[24] extracted low-dimensional features and high-dimensional semantic representations from acoustic, visual, and textual modalities. They used an improved MulT network to perform both intra- and inter-modal fusion and employed the fused results for depression classification. Fan et al.[25] used a CNN to extract video and audio features and employed a short-term end-to-end remote photoplethysmography (rPPG) estimation framework to obtain rPPG signals. They then captured intra- and inter-modal dynamic relationships using a multi-layer Transformer module and achieved depression detection through a graph fusion network (GFN) and a multilayer perceptron (MLP).

Due to the high temporal resolution and other advantages of EEG, many researchers focus on multimodal depression recognition studies based on EEG. Ning et al.[26] extracted linear and nonlinear features from EEG and speech signals and performed feature-level fusion to construct a multimodal feature space for training the classifier. Zhang et al.[27] extracted multiple effective features from EEG and speech signals and trained each modality using six representative classifiers. They then used a co-decision tensor to represent the decision diversity and correlation among the classifiers and combined these decisions into a final classification result using a multi-agent strategy. Cai et al.[28] proposed two feature-level fusion methods to integrate features extracted from two modalities: Eyes-closed resting state EEG and emotion-regulation EEG. Various machine learning classifiers were then used to evaluate the performance of the proposed methods. Zhu et al.[29] used PA signals to select EEG electrodes based on mutual information, then extracted features from different EEG bands and PA signals. A denoising autoencoder was employed for bimodal feature fusion.

Current research on multimodal depression recognition primarily focuses on the effective integration and utilization of information from multiple modalities to improve recognition accuracy and reliability. However, while multimodal approaches benefit from incorporating diverse information sources, they also introduce new challenges. Specifically, many methods rely on the fusion of EEG signals with other modalities for depression detection, yet obtaining high-quality EEG data remains a nontrivial task. EEG signals are easily affected by various external interferences and noise[12], including environmental noise, electromagnetic interference, and the subject's own physiological activities, such as eye movements,

muscle movements, and heartbeat. These interferences can significantly impact the quality of EEG signals, thereby increasing the difficulty of data processing. In addition, EEG equipment demands are high[30], typically requiring high-quality electrodes and wires, along with precise placement and good contact to ensure signal accuracy and stability. Such high-demand equipment is not only expensive but also requires professional personnel for operation and maintenance. More importantly, during EEG experiments, subjects typically need to remain still for extended periods to avoid artifacts and noise caused by body movements, which places high demands on their comfort and tolerance. These factors combined make the acquisition and analysis of EEG signals highly challenging.

In contrast, thanks to the mature development of eye-tracking technology, the acquisition of eye-tracking data is not only simple to operate but also provides stable and reliable signals, which can be gathered with small pieces of glasses[31]. Alghowinem et al.[32] extracted eye-tracking features from clinical interview videos and employed Gaussian mixture models (GMMs) and support vector machines (SVMs) for depression detection. Shen et al.[33] trained a psychological state classifier based on SVMs using psychological features extracted from eye-tracking data to classify depression patients and normal controls.

2.2 Cross-modal knowledge distillation

Knowledge distillation aims to transfer the knowledge of a large and complex teacher model to a smaller student model, enabling the student model, with fewer parameters, to achieve accuracy closer to that of the teacher model. This is particularly crucial for mobile devices and edge computing. Hinton et al.[34] first proposed response-based distillation, followed by Romero et al.[35], who introduced feature-based distillation. Knowledge distillation has also been extended to broader learning paradigms beyond standard supervised settings, such as federated and semi-supervised learning[36].

With the rise of multimodal learning, researchers have gradually recognized that knowledge distillation can be performed not only within the same modality but also across different modalities, a process known as cross-modal knowledge distillation. As an early attempt, Gupta et al.[37] utilized knowledge from red, green, and blue (RGB) images to enhance the performance of depth images. Albanie et al.[38] trained a teacher network using facial expression data and employed cross-modal distillation to transfer emotion labels from the visual domain (facial expressions) to the audio domain (speech). This enabled the student model to learn to recognize emotions in speech without having been exposed to labeled audio data. Zhang et al.[39] trained a teacher model using video frame data and transferred the dark knowledge from the visual modality to a student model trained with EEG mean band power features, improving continuous emotion prediction in the EEG modality. Liu et al.[40] proposed a cross-modal knowledge distillation framework called EmotionKD, which distilled the fused multimodal features of EEG and galvanic skin response (GSR) into a unimodal GSR model. Additionally, an adaptive feedback mechanism was introduced to dynamically adjust the distillation process, enhancing the performance of the GSR model in emotion recognition tasks. Despite its broad application potential, cross-modal knowledge distillation still faces several challenges, such as modality heterogeneity and the performance gap between teacher and student models.

In addition, to the best of our knowledge, there is limited work that addresses the problem of multi-to-one cross-modal transfer in the field of depression recognition. We propose a cross-modal knowledge distillation method that transfers the fused information from EEG and PA to a PA-based unimodal model, reducing reliance on EEG and improving the generalization capability of the PA-based unimodal model.

2.3 Explainable methods in time series

Explainability methods for deep neural networks (DNNs) have evolved significantly over time. Early techniques such as gradient explanations[15] and CAM[16] focused on identifying important input regions that influence model predictions. Gradient explanations compute the gradient of input features with respect to the model's output to assess the contribution of each feature, while CAM leverages the weights between the global average pooling layer and the full connected (FC) layer to assign importance scores to different positions in the input. As deep learning architectures incorporating attention mechanisms became more prevalent, Attention maps[17] emerged as a popular interpretation method. These maps visualize the attention weights within the model to highlight which features or time steps are

being emphasized during prediction. More recently, model-agnostic approaches such as SHAP[18] and LIME[19] have gained attention for their general applicability. SHAP assigns feature importance scores using Shapley values, considering all possible combinations of input features, while LIME explains individual predictions by approximating the model locally with an interpretable surrogate model. However, these methods have limitations[41]. Gradient explanations are severely affected by noisy gradients and may not satisfy the continuity of explanations[42]. CAM is dependent on specific network architectures[43]. The magnitude of attention weights does not always reflect the true importance of specific inputs to the model's output[44]. SHAP is computationally expensive, especially for models with many features[18]. LIME may produce inconsistent explanations across different runs due to its reliance on local approximations[19]. Ma et al.[45], building on previous research related to the information bottleneck theory[46, 47], viewed the forward propagation process of DNNs as a layer-by-layer filtering of input information. They proposed a method to explain the model by quantifying the information discarded by the model during this process. This method has been validated in image classification tasks but its application to time series data remains relatively limited.

To address these issues, we integrate information entropy theory with the Grad-CAM technique and propose a novel explainability method. This method not only quantifies the information loss in time series data but also uses Grad-CAM to evaluate the contribution of individual features to model predictions, providing a more detailed and general explainable tool for understanding and optimizing the knowledge distillation process.

3 Methodology

The proposed method aims to (1) train an EEG and PA-based multimodal model, (2) use knowledge distillation to transfer the knowledge captured by the teacher model to PA-based unimodal student model, and (3) apply an information entropy theory-based approach combined with Grad-CAM to explain the performance improvement of the student model. In this section, we first introduce the multimodal teacher model and the unimodal student model, followed by an explanation of the interaction between the teacher and student models, along with the explainability methods used.

3.1 Teacher: The EEG-PA model

In the process of cross-modal knowledge distillation, the key to improving the performance of the unimodal model lies in effectively learning and absorbing the rich knowledge transferred from the multimodal teacher model. Therefore, efficiently capturing the diverse and rich feature information in the multimodal model is a prerequisite for the success of the distillation process. The upper part of Fig. 1 presents the overall structure of the multimodal teacher model. To explore the heterogeneity and correlations between modalities, we design a cross-modal interaction module to capture

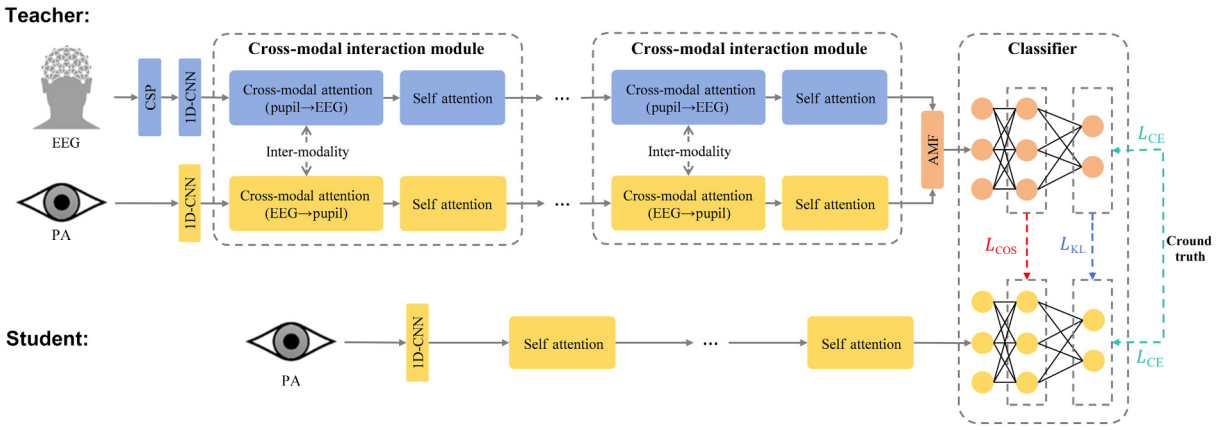

Fig. 1 Framework of our proposed method. The teacher consists of EEG and PA branches, which interact through the cross-modal interaction module and the adaptive modality fusion module. To ensure maximum similarity, the student's PA branch is identical to the teacher's, except for the cross-modal attention. The classification module has the same structure in both models. CSP: Common spatial pattern, AMF: Adaptive modality fusion, 1D: One-dimensional.

the rich knowledge both within and between modalities, and an adaptive modality fusion module to balance the impact of the two modality branches on classification.

3.1.1 Feature extraction

In the multimodal teacher model, CSP is first applied to EEG data to extract task-related spatial features. Then, as with PA, CNN is used to extract temporal features. Let the raw EEG data be $X_E \in \mathbb{R}^{C_1 \times T}$ and the raw PA data be $X_P \in \mathbb{R}^{C_2 \times T}$, where C_1 and C_2 represent the number of channels, and T represents the number of sampling points. The feature extraction process is defined as Eqs (1) and (2),

$$Z_E = \mathrm{CNN}(\mathrm{CSP}(X_E)) \tag{1}$$

$$Z_P = \mathrm{CNN}(X_P) \tag{2}$$

where Z_E denotes the EEG feature representation extracted by CSP followed by CNN, and Z_P denotes the PA feature representation extracted by CNN, both are used for subsequent cross-modal interaction.

3.1.2 Cross-modal interaction module

Each cross-modal interaction (CMI) module includes a cross-modal attention and a self-attention mechanism. The EEG and PA features obtained from feature extraction are divided into n embedding vectors along the temporal dimension, serving as inputs to this module. The EEG features are represented as $Z_E = \left[Z_{E_1}, Z_{E_2}, \ldots, Z_{E_n}\right]$, and the PA features are represented as $Z_P = \left[Z_{P_1}, Z_{P_2}, \ldots, Z_{P_n}\right]$.

(1) Cross-modal attention. To achieve cross-modal information exchange and mutual enhancement, while mitigating the distribution differences between the two modalities at the feature level, we design a cross-modal attention mechanism. This cross-modal attention involves bidirectional information exchange, where EEG information is transferred to PA, and simultaneously, PA information is transferred to EEG. This thorough information interaction provides rich and meaningful feature representations for subsequent knowledge distillation. In this way, when distilling to the unimodal model, the PA modality can acquire more comprehensive and accurate knowledge for depression recognition, significantly enhancing the effectiveness of knowledge distillation. The formula for cross-modal attention is as Eqs. (3) and (4),

$$\mathrm{Attention}_{E\to P} = \mathrm{softmax}\left(\frac{Q_E K_P^{\mathrm{T}}}{\sqrt{d}}\right) V_P \tag{3}$$

$$\mathrm{Attention}_{P\to E} = \mathrm{softmax}\left(\frac{Q_P K_E^{\mathrm{T}}}{\sqrt{d}}\right) V_E \tag{4}$$

where $\mathrm{Attention}_{E\to P}$ and $\mathrm{Attention}_{P\to E}$ both denote cross-modal attention between EEG and PA branches. Q_E and Q_P represent the query matrices for the EEG and PA modalities, respectively. K_E^{T} and K_P^{T} are the key matrices, and V_E and V_P are the value matrices. The scaling factor $\sqrt{d}$ is used to stabilize the numerical values. softmax normalizes the scaled dot-product similarity scores into attention weights.

(2) Self attention. While multimodal data can provide a comprehensive and rich perspective, it also introduces additional noise and redundant data. The role of self-attention is to enable the network to focus on key features within a single modality, effectively filtering out irrelevant information[48]. The formula for self-attention is as Eqs. (5) and (6),

$$\mathrm{Attention}_E = \mathrm{softmax}\left(\frac{Q_E K_E^{\mathrm{T}}}{\sqrt{d}}\right) V_E \tag{5}$$

$$\mathrm{Attention}_P = \mathrm{softmax}\left(\frac{Q_P K_P^{\mathrm{T}}}{\sqrt{d}}\right) V_P \tag{6}$$

where $\mathrm{Attention}_E$ and $\mathrm{Attention}_P$ denote the self-attention in the EEG and PA branches, respectively.

3.1.3 Adaptive modality fusion module

To effectively balance the contributions of the two modality branches to the final classification outcome, we propose an AMF module. Figure 2 shows the overall structure of the AMF module. This module first concatenates the high-dimensional features from different modality branches and then generates contribution weights for each modality through several network layers. Finally, these weights are used to apply weighting to the features from both modalities, ensuring that the impact of each modality on the final result can be flexibly adjusted during classification.

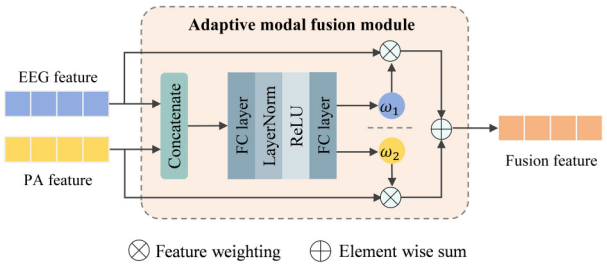

Fig. 2 Structure of the adaptive modality fusion module. ω_1 and ω_2 represent the weights of the two features, respectively. FC: Fully connected, LayerNorm: Layer normalization, ReLU: Rectified linear unit.

After the CMI module, the resulting features for the EEG and PA branches are denoted by Feature_E and Feature_P, respectively. The operation of the AMF module can be defined as Eq. (7),

$$(\text{Weights}_{E,P}, \text{Weights}_P) = \eta(\text{FC}_2(\psi(\text{FC}_1(\text{Concat}(\text{Feature}_E, \text{Feature}_P))))) \quad (7)$$

where Weights_E and Weights_P denote the learned adaptive fusion weights for Feature_E and Feature_P, respectively. η denotes the softmax activation function, FC_i represents the i-th FC layer, ψ represents the ReLU activation function, and Concat represents the tensor concatenation operation. After the softmax layer, two weights are obtained, which are then multiplied with the input features Feature_E and Feature_P, respectively, and subsequently added to obtain the fused feature, defined as Eq. (8),

$$X_{\text{output}}^{\text{AMF}} = \text{Feature}_E \cdot \text{Weights}_E + \text{Feature}_P \cdot \text{Weights}_P \quad (8)$$

where $X_{\text{output}}^{\text{AMF}}$ denotes the final fused feature representation produced by the AMF module, which is subsequently fed into the classifier for prediction.

3.1.4 Classifier

Finally, the fused features obtained from the AMF module are fed into a classification module composed of two FC layers.

In summary, the proposed teacher model can extract heterogeneous and interactive features, which theoretically enhances the classification accuracy and generalization ability of the student model.

3.2 Student: The PA model

In knowledge distillation, the capacity gap is a common issue[49, 50]. Specifically, teacher models typically have a larger capacity, enabling them to learn complex features and patterns with strong representational power. In contrast, student models, with their smaller capacity, are unable to fully match and replicate the teacher model's complex behaviors and outputs. This capacity gap makes it difficult for the student model to effectively learn and absorb the knowledge provided by the teacher model during knowledge distillation, thereby affecting the distillation effectiveness and the final model performance. To minimize the capacity gap, we design a structure in the student model similar to the PA branch in the teacher model. First, a CNN is used to extract temporal features, followed by a self-attention mechanism to capture deeper contextual associations and detailed information. Finally, the same classification module as in the teacher model is used for classification. The lower part of Fig. 1 presents the overall structure of the unimodal student model.

3.3 Knowledge distillation

The right side of Fig. 1 shows the overall structure of the distillation process. We aim for the student model to closely mimic the behavior of the teacher model, meaning that its output probability distribution should approximate that of the teacher model. Therefore, following the original knowledge distillation approach, KL-divergence is used as a loss term. Assuming the output probability distribution of the teacher model is P and that of the student model is Q, the loss term L_{KL} is defined as Eq. (9),

$$L_{\text{KL}} = \sum_i P_i \ln\left(\frac{P_i}{Q_i}\right) \quad (9)$$

where P_i represents the predicted probability of the teacher model for the i-th class, and Q_i represents the predicted probability of the student model for the i-th class.

Additionally, to enable the student model to gain richer representational capacity, the intermediate layer features within the student model should align as closely as possible with the corresponding layer features of the multimodal teacher model. Specifically, we select the output of the first FC layer in the classification module, which typically contains the model's most abstract and high-level feature representations. Moreover, in the process of knowledge distillation from a multimodal model to a unimodal model, differences in feature scales exist between modalities and models. Using mean squared error (MSE) directly as a loss function carries the risk of overfitting. Therefore, cosine similarity is used to capture the directional relationship between feature vectors, reducing the impact of scale differences. This approach enhances the cross-modal transfer of semantic information, improving both the performance and generalization capability of the student model.

Assuming the output feature of the k-th layer in the teacher model is T^k and the corresponding output feature in the student model is S^k, the cosine similarity loss term L_{COS} is defined as Eq. (10),

$$\text{CosineSimilarity}(T^k, S^k) = \frac{T^k \cdot S^k}{||T^k|| ||S^k||} \quad (10)$$

where $\text{CosineSimilarity}(T^k, S^k)$ measures the

directional similarity between the teacher and student features by computing the normalized inner product.

$$L_{\mathrm{COS}} = 1 - \mathrm{CosineSimilarity}(T^k, S^k) \tag{11}$$

To further improve the classification accuracy of the student model, cross-entropy loss is introduced as part of the supervision. Given a one-hot label Y and the output Q of the student model, the cross-entropy loss term L_{CE} is defined as Eq. (12),

$$L_{\mathrm{CE}} = -\sum_i Y_i \ln(Q_i) \tag{12}$$

where Y_i is the i-th component of the one-hot label, and Q_i is the predicted probability of the student model for the i-th class.

The total loss function is the weighted sum of the three loss terms, as shown in Eq. (13),

$$\mathrm{Loss} = \alpha L_{\mathrm{KL}} + \beta L_{\mathrm{COS}} + \gamma L_{\mathrm{CE}} \tag{13}$$

where α, β, and γ are balancing factors used to adjust the relative weights of the different loss terms.

3.4 Entropy-GradCAM-based time series explanation method (E-GCAM)

Currently, effective quantitative methods are lacking to explain why the knowledge distillation process is effective in time series. Inspired by Ma et al.[45], we introduce, for the first time, the concept of knowledge point quantification in time series and propose a novel method for calculating knowledge point quality. The overall computational process of the proposed method is summarized in Algorithm 1.

Algorithm 1 Entropy-GradCAM-based time series explanation method (E-GCAM)

Input: PA feature z_P, τ: The threshold for selecting σ_i, σ_i: The variance of the perturbation added to the i-th token, b_1: The threshold for determining a token as a knowledge point, b_2: The threshold for determining a token as part of the foreground

Output: [num(N_{KP}), λ]

1: Initialize σ_i;
2: Calculate z', δ_f^2;
3: **for** epochs **do**
4: Calculate entropy H_i by Eq. (16);
5: Calculate Loss(σ_i) by Eq. (18);
6: Update σ_i;
7: **end for**
8: Compare each H_i obtained in Line 4 with τ, and select the σ_i corresponding to the entropy closest to τ;
9: Calculate the entropy H_i based on the σ_i obtained in Line 8;
10: Use the entropy H_1 of the first token as the baseline;
11: Determine the knowledge points by Eq. (19);
12: Calculate the contribution c_{i1} and c_{i2} of the i-th token in the two classes using Grad-CAM;
13: $c_{\mathrm{diff}} = |c_{i1} - c_{i2}|$;
14: Determine the foreground by Eq. (20);
15: Calculate the quality λ by Eq. (21);
16: **return** [num(N_{KP}), λ]

According to previous studies on the information bottleneck theory[46, 47], the forward propagation process in deep neural networks (DNNs) can be viewed as a layer-by-layer filtering of input information. In this way, given the features of an intermediate layer, it is possible to measure the information discarded from each input unit. In this study, input units are considered as several tokens obtained by sequentially dividing the input sample along the temporal dimension. Specifically, let $z \in \mathbb{R}^n$ denote an object instance, and $f = h(z) \in \mathbb{R}^m$ represent the feature extracted from an intermediate layer of the DNN. A previous study[45] proposed the DNN represents a specific object instance z using a very limited range of features, with the average feature denoted as f. Similarly, for an object instance z, there exists a latent space $X_{\mathrm{obj}} = \{z' \mid \|h(z') - f\|^2 \leqslant \theta\}$, where all perturbed inputs z' around z yield features $h(z')$ that remain within a small neighborhood of the feature f. Here, θ is a small constant. We obtain a perturbed sample z' by adding a small noise Δz to the input z, and measure the maximum noise magnitude that each token can tolerate while ensuring that the change in intermediate features remains sufficiently small. In this way, the entropy $H(Z')$ can be used to quantify how much input information can be discarded,

$$H(Z'), \text{ s.t., } \forall z' \in Z', \|h(z') - h(z)\|^2 \leqslant \theta \tag{14}$$

$$z' = z + \Delta z \sim N\left(z, \sum(\sigma) = \mathrm{diag}\left(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2\right)\right) \tag{15}$$

Assuming the input sample consists of n tokens, independent and identically distributed Gaussian noise is added to these tokens, yielding a set of inputs Z' that correspond to a specific object instance, where σ_i denotes the standard deviation of the Gaussian perturbation for the i-th token. The Gaussian noise allows the entropy $H(Z')$ to be decomposed into token-level entropies $\{H_i\}$,

$$H(Z') = \sum_{i=1}^{n} H_i, \quad H_i = \ln \sigma_i + \frac{1}{2}\ln(2\pi \mathrm{e}) \tag{16}$$

where H_i measures the information discarded at the token level, with a larger H_i indicating that information from the i-th token is discarded more during forward propagation. To obtain the optimal noise, Lagrange multipliers are used to approximate Formula (14) as the loss function shown in Formulas (17) and (18),

$$\min_{\sigma} \text{Loss}(\sigma) \tag{17}$$

$$\text{Loss}(\sigma) = \frac{1}{\delta_f^2} E_{z'=z+\sigma\cdot\delta,\delta\sim N(0,1)}\left[\|h(z')-h(z)\|^2\right] - \mu\sum_{i=1}^{n} H_i \tag{18}$$

where $\sigma = [\sigma_1, \sigma_2, \ldots, \sigma_n]^{\mathrm{T}}$ is the parameter that we aim to learn, represents the noise added to different tokens. $E_{z'=z+\sigma\cdot\delta,\,\delta\sim N(0,1)}$ denotes the expectation over standard Gaussian noise added to z to form $z' = z+\Delta z$, averaged across multiple samples. $\delta_f^2 = \lim_{\phi\to 0^+} E_{z'\sim\mathcal{N}(z,\phi^2)}\left[\|h(z')-h(z)\|^2\right]$ is the inherent variance of intermediate-layer features subject to small input noises with a fixed magnitude ϕ, which is used for normalization. μ is used to balance the two loss terms. $h(z)$ and $h(z')$ denote the intermediate-layer feature representations of the original input z and the perturbed input z', respectively.

For tokens with less discarded information, it is considered that they encode more discriminative information useful for prediction, which we refer to as knowledge points. Specifically, in this experiment, input samples are divided into n tokens along the temporal dimension. As shown in Figs. 3a and 3c, the entropy of discarded information in the first token is relatively high in the majority of samples, indicating that the pupil data in the early stages of stimulus exposure carry sparse knowledge information. Siegle et al.[51] noted that patients with depression often exhibit slower initial responses in emotional stimulus tasks, requiring more time to process this information. This slower response is reflected in both reaction time and pupil response. Oliva and Anikin[52] demonstrated that pupil dilation associated with emotion recognition is not instantaneous but rather accumulates gradually. Therefore, the entropy H_1 of the first token is used as a baseline for determining knowledge points. Specifically, we label token i as a knowledge point when its entropy drop relative to the baseline H_1 exceeds a threshold b_1,

$$H_1 - H_i > b_1 \tag{19}$$

If the entropy of a token is significantly lower than the baseline entropy, it can be considered a knowledge point. Formula (19) is used to determine whether a given token constitutes a knowledge point, where b_1 serves as the threshold for determining knowledge points, allowing for the quantification of knowledge points in the model. Theoretically, the distilled student model encodes more knowledge points than the student model trained from scratch.

Additionally, to evaluate encoding quality in time series, we propose a novel method for calculating knowledge point quality. In image processing, the concepts of foreground and background are commonly utilized, where the foreground typically contains information strongly relevant to classification tasks, while the background often consists of irrelevant information. Analogously, we aim to identify foreground tokens and background tokens in time series data. To achieve this, we employ Grad-CAM to quantify the contribution of each token to the classification outcome. The absolute difference between the two class contributions for each token, c_{i1}

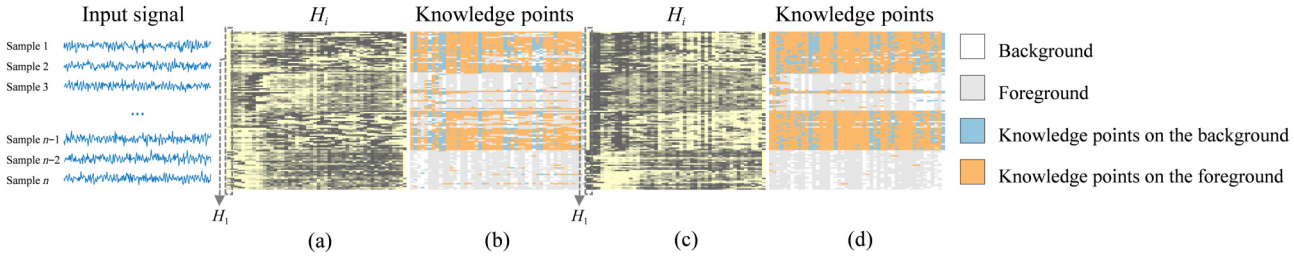

Fig. 3 Knowledge point visualization in EFVP dataset. Dark regions represent low entropy values (H_i) of discarded information. Figures 3a and 3c display the H_i values of students trained from scratch and those subjected to distillation, respectively. Each row represents a different sample, and each column corresponds to a time-segmented token of that sample. In this study, all samples were uniformly divided into 50 tokens. Figures 3b and 3d illustrate the distribution of knowledge points derived from Figs. 3a and 3c, respectively. Knowledge points located in the foreground are considered effective knowledge points.

and c_{i2}, denoted as c_{diff}. This value is used as the basis for determining foreground relevance.

$$c_{\text{diff}} > b_2 \tag{20}$$

If a token's c_{diff} is greater than b_2, the token can be considered as part of the foreground, where b_2 is the threshold for determining foreground, which allows for the quantification of knowledge point quality within the model. Knowledge points in the foreground are regarded as effective knowledge points, and the proportion of effective knowledge points is defined as quality λ.

$$\lambda = \frac{\text{num}(N_F \cap N_{\text{KP}})}{\text{num}(N_{\text{KP}})} \tag{21}$$

where N_{KP} represents the set of knowledge points, and N_F represents the set of foreground tokens, and num denotes the number of elements of a set. Thus $\text{num}(N_F \cap N_{KP})$ counts the number of effective knowledge points, while $\text{num}(N_{KP})$ counts the total number of knowledge points.

4 Result and Discussion

4.1 Dataset

Affective attention dot-probe (AADP) dataset: This is a multimodal dataset comprising EEG and eye-tracking data from 35 patients with depression and 35 normal controls. The data were collected at Gansu Provincial Hospital. The study protocol was approved by the Ethics Committee of Gansu Provincial Hospital (Approval No. 2017-071). Each participant completed 400 trials, where they first viewed a combination of neutral and emotional faces for 500 ms. Following this, they were tasked with identifying the position of a dot on the screen and responding as quickly as possible by pressing a button. EEG data were collected using a 64-channel device with a sampling rate of 1000 Hz, and eye-tracking data were sampled at 1000 Hz.

Emotion free-viewing EEG-pupil (EFVP) dataset: This is another multimodal dataset, comprising EEG and eye-tracking data from 20 patients with mild depression and 20 normal controls. The data were collected from students at Lanzhou University with approval from the Ethics Committee of the Second Hospital of Lanzhou University (Approval No. 2015 A-037). All participants provided written informed consent prior to participation. Each participant completed 40 trials, with each trial involving the viewing of a combination of neutral-neutral or neutral-negative emotional faces for 6 s. EEG data were collected using a 128-channel device with a sampling rate of 250 Hz, and eye-tracking data were sampled at 1000 Hz.

Both datasets are private and are available upon reasonable request by contacting the corresponding author.

4.2 Data preprocessing

In this study, raw EEG data are preprocessed using NetStation software, including the removal of bad electrodes and baseline correction. To effectively suppress physiological artifacts such as electrooculogram (EOG) and electromyogram (EMG) signals, a 0.5 Hz high-pass filter and a 50 Hz low-pass filter are applied to the EEG signals. In addition, the Fast ICA algorithm[3] is employed to isolate and remove EOG artifact interference from the EEG data. For the eye-tracking data, only segments with a data loss rate below 30% are retained. In cases where certain features within a segment have missing values (less than 30%), the missing values are imputed using the mean of the corresponding feature. Previous studies indicated that pupil dilation is a key indicator of cognitive load and is closely associated with neural activity in brain regions implicated in depression[53, 54]. We extract the PA signal from EyeLink data files recorded by the EyeLink 1000 system for subsequent analysis. Missing values are imputed using the mean of the segments immediately before and after the missing interval. Additionally, to align the sampling frequency with that of the EEG signals and reduce processing complexity, the PA signals in the EFVP dataset are resampled to 250 Hz.

4.3 Experimental setup

To obtain a PA-based model suitable for cross-subject deployment, we adopt a subject-independent cross-validation strategy. All trials from the same subject are strictly kept within the same fold to prevent data leakage. Specifically, for the AADP dataset, we perform seven-fold cross-validation by dividing the 70 participants into seven folds: Five folds are used for training, one for validation, and the remaining one for testing. For the EFVP dataset, we conduct ten-fold cross-validation by dividing 40 participants into ten folds: Eight folds are used for training, one for validation, and one for testing. All folds are balanced to ensure an equal number of participants with depression

and healthy controls, thereby maintaining class balance. Classification accuracy and F1 score are used to evaluate classification performance.

Additionally, EEG and eye-tracking data are collected simultaneously in both datasets. The teacher model is trained using paired EEG and PA data, while the student model is trained using only PA data.

4.4 Results

To verify the effectiveness of our proposed cross-modal knowledge distillation, the distilled unimodal student model is compared with the baseline model based on PA signals and with cross-modal baseline methods applied to the unimodal student model. All methods are evaluated on PA data. The experimental results, shown in Table 1, indicate that the proposed method achieved the highest accuracy and F1 score, demonstrating the effectiveness of the distillation approach. Specifically, our method uses a multimodal teacher model to deeply explore the intrinsic features of EEG and PA modalities and their interactions. Through knowledge distillation, the rich information and valuable insights obtained from the teacher model are efficiently transferred to the student model. This enables the student model to learn not only the high-level abstract features from the teacher model but also the interactions between modalities, resulting in higher accuracy in classification tasks. Methods such as MLA+LSTM[55], CNN+RNN[56], LCTNN[57], and TCNN[58] which rely solely on unimodal data, inevitably face limitations in information availability. Without guidance from multimodal knowledge, their performance reaches a bottleneck. CGAN[59] adopts a cross-domain knowledge transfer strategy but lacks deep modeling of modality-specific complexities, limiting its performance on highly heterogeneous data. scalp-EEG to ear-EEG KD[60] uses identical architectures for teacher and student, failing to showcase the model compression benefits of knowledge distillation. CMTKD[61] focuses on teacher collaboration but lacks deep interaction modeling, limiting the benefits of multi-teacher fusion. EmotionKD[40] uses a dual-stream Transformer and a temporary student for feedback, greatly increasing training cost.

Additionally, to demonstrate the effectiveness of the proposed multimodal teacher model in multimodal depression recognition, it is compared with other baseline models on both datasets. All methods are evaluated on both EEG and PA data, and the experimental results are presented in Table 2. The GSCCA[62] method explores the linear correlation between EEG and eye-tracking data but cannot capture the complex nonlinear interactions between modalities. The DCCA[63] method places excessive emphasis on inter-modal correlations, which may weaken the independent features of each modality. MTS-Net[64] focuses on capturing the heterogeneity of multimodal signals, but it uses simple concatenation for feature fusion, failing to fully capture the complex interactions between modalities. MIBFM[29] relies on high-dimensional feature extraction and feature selection, which limits further improvement in model performance. MFFNN[65] uses CNNs for feature

Table 1 Comparison of the student model with other baseline methods (mean ± standard deviation). * denotes that EmotionKD was not implemented on the AADP dataset due to its high computational resource demands. The best results in terms of the mean accuracy and mean F1 score are highlighted in boldface.

(%)

Method	AADP dataset		EFVP dataset	
	Accuracy	F1 score	Accuracy	F1 score
MLA+LSTM[55]	77.33 ± 9.36	76.27 ± 11.13	73.75 ± 11.65	73.31 ± 11.91
CNN+RNN[56]	77.06 ± 8.43	77.05 ± 10.68	74.00 ± 12.13	72.80 ± 9.46
LCTNN[57]	78.07 ± 4.05	77.85 ± 4.18	74.58 ± 12.21	73.53 ± 13.39
TCNN[58]	78.79 ± 2.22	78.49 ± 2.35	75.00 ± 14.50	74.41 ± 15.34
CGAN[59]	70.44 ± 7.21	71.78 ± 6.36	75.58 ± 19.61	71.53 ± 28.70
Scalp-EEG to ear-EEG KD[60]	76.26 ± 7.98	75.07 ± 10.17	72.83 ± 11.31	72.16 ± 11.52
CMTKD-student[61]	75.58 ± 2.97	74.49 ± 3.05	72.50 ± 13.73	71.57 ± 14.32
EmotionKD-student[40]	*	*	72.67 ± 14.32	73.23 ± 13.48
Scratched-student	75.83 ± 5.91	75.25 ± 5.84	73.25 ± 15.10	71.76 ± 16.77
KD-student (ours)	**81.45** ± 3.29	**80.95** ± 3.52	**78.50** ± 12.46	**77.53** ± 13.26

Table 2 Comparison of the teacher model with other baseline methods (mean ± standard deviation). * denotes that MIBFM was not implemented on this dataset due to the complexity of its procedure and the laborious feature extraction process. The best results in terms of the mean accuracy and mean F1 score are highlighted in boldface.

(%)

Method	AADP dataset		EFVP dataset	
	Accuracy	F1 score	Accuracy	F1 score
GSCCA[62]	71.68 ± 0.06	74.46 ± 0.04	67.58 ± 9.16	67.62 ± 6.08
DCCA[63]	81.75 ± 5.81	82.13 ± 6.49	75.67 ± 16.38	73.36 ± 19.18
MTS-Net[64]	88.56 ± 7.33	88.44 ± 13.17	74.90 ± 18.85	79.09 ± 13.53
MIBFM[29]	*	*	82.00 ± 11.47	85.14 ± 9.25
MFFNN[65]	89.01 ± 4.40	89.89 ± 4.22	77.83 ± 12.79	76.38 ± 14.02
MCAF[66]	88.21 ± 3.38	88.15 ± 3.37	79.58 ± 9.14	78.94 ± 9.57
CMTKD-teacher[61]	73.63 ± 15.05	72.84 ± 14.52	66.67 ± 13.20	65.75 ± 13.73
EmotionKD-teacher[40]	*	*	77.50 ± 6.12	78.08 ± 6.46
KD-teacher (ours)	**90.83** ± 4.09	**90.79** ± 4.12	**86.08** ± 10.74	**85.41** ± 11.44

extraction but struggles to model long-range cross-channel dependencies, limiting its grasp of global physiological patterns. MCAF[66] fuses modalities early via channel-wise concatenation, but EEG's higher channel count can dilute the other modality's contribution. The OurKD-Teacher model, with its CMI and AMF modules, not only models modal heterogeneity, but also captures modal interactivity. Consequently, it achieves state-of-the-art performance among multi-modal models.

4.5 Why did the student improve?

For the student model obtained through knowledge distillation (K) and the student model trained from scratch (S), we measure the knowledge points on the first FC layer in the classification module, as shown in Figs. 3b and 3d. The FC layer integrates and abstracts features from the preceding layers, forming global features that directly impact the final task. By quantifying knowledge points in this layer, it becomes possible to more accurately identify and analyze key input units that significantly impact model performance, thereby supporting model optimization and explainability.

We select three thresholds b_1 and four thresholds b_2, and then compare the knowledge points encoded in the KD student model and the scratched student model.

Table 3 Comparison of the number of knowledge points between the KD student and the scratched student. Bold numbers indicate the better results in terms of the number of knowledge points.

Threshold b_1	Model	num(N_{KP})	
		AADP dataset	EFVP dataset
0.10	KD student	**21.90**	**19.67**
	Scratched student	18.56	13.57
0.15	KD student	**19.39**	**10.68**
	Scratched student	10.45	4.09
0.20	KD student	**17.06**	**6.70**
	Scratched student	4.74	1.45

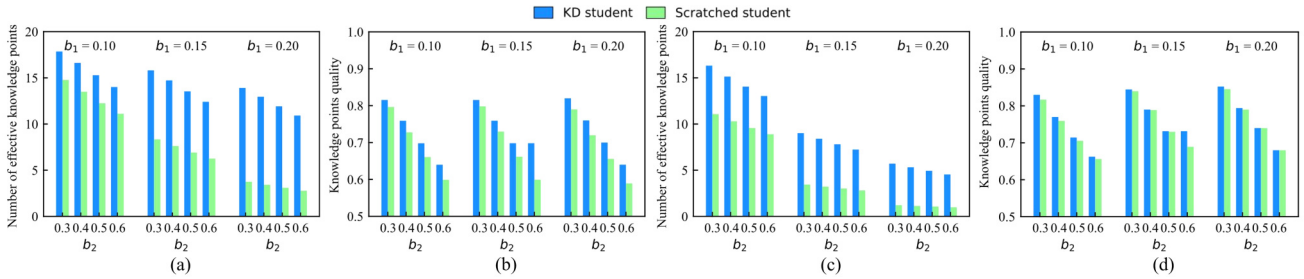

Fig. 4 Comparison of the number of effective knowledge points and knowledge points quality between the KD student model and the scratched student model. Figures 4a and 4b show the number of effective knowledge points and the quality for the two models on AADP dataset, while Figs. 4c and 4d show the number of effective knowledge points and the quality for the two models on EFVP dataset. b_1 denotes the threshold for identifying knowledge points, and b_2 represents the threshold for foreground determination.

Table 3 indicates that the KD student model consistently encodes more knowledge points. As seen in Fig. 4, the KD student model typically encodes a larger number of effective knowledge points and a higher quality λ. This explains the reason for the performance improvement in the KD student model: Knowledge distillation enables the KD student model to encode more knowledge points with higher quality than the scratched student model. This method, combining information entropy and Grad-CAM to quantify useful information in the model, can also be applied to other time series models.

Furthermore, to validate the effective guidance of the multimodal teacher model on the unimodal student model in cross-modal knowledge distillation, we visualize the distribution of discarded information entropy over time for both the teacher and student models. As mentioned earlier, the smaller the entropy, the more discriminative information is encoded. As shown in Fig. 5a, in AADP dataset, where the data duration is 1 s, the EEG branch responds rapidly and significantly to the stimulus (0–150 ms), with its entropy dropping below baseline levels first. This significance is referred to as early knowledge. In contrast, the pupil branch exhibits slower changes. As observed in Fig. 5b, after distillation, the student model alter the original early knowledge distribution, shifting to concentrate in both earlier and later time periods. As shown in Fig. 5c, in EFVP dataset, where the data duration is 6 s, the EEG branch has a limited early knowledge distribution, while the pupil branch gradually exhibits more significant changes in early knowledge over a longer timescale (1.5 to 6 s). As observed in Fig. 5d, after distillation, the student model's early knowledge distribution closely aligns with that of the pupil branch, with the concentration of knowledge further enhanced.

Previous studies[67, 68] have that EEG signals are highly sensitive to external stimuli and emotional changes, and can respond rapidly within a short period, reflecting the timely regulation of brain neural activity. Therefore, the student model may learn the rapid response mechanism in the EEG branch, enabling it to adaptively respond to sudden changes in emotional fluctuations or cognitive load. In addition, related studies[51, 52] indicated that pupil typically reflects the regulation of physiological or emotional states, with a longer timescale for responses to cognitive load, emotional changes, and other factors. This also helps explain why the student model primarily relies on EEG signals at the 1-s short timescale and on pupil signals at the 6-s long timescale.

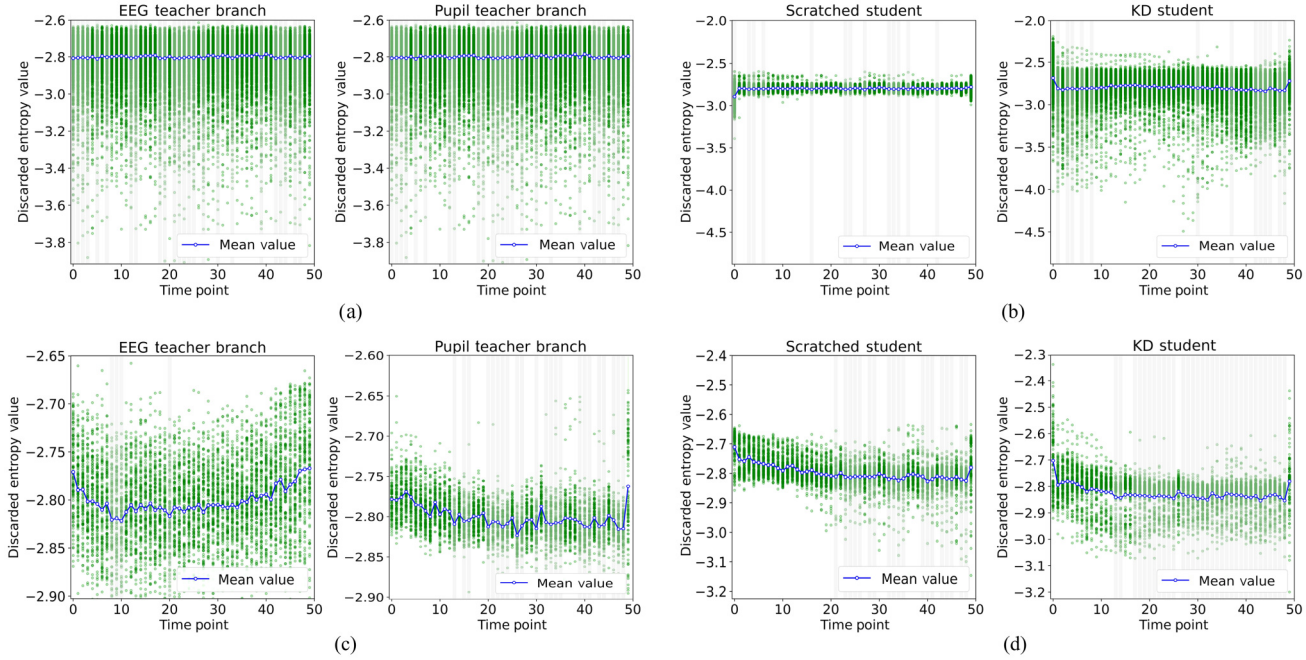

Fig. 5 Distribution of discarded information entropy value H_i at different time stages. The time segments where H_i is significantly lower than the baseline level, as determined by the Wilcoxon signed-rank test, are marked with shaded regions. The baseline level for each model is the mean of its entire dataset. Figures 5a and 5b are from AADP dataset, and Figs. 5c and 5d are from EFVP dataset.

4.6 Ablation experiment

We conduct ablation studies from three dimensions to evaluate the importance of key components in the teacher model, multimodal data sources, and the loss function during distillation. First, we focus on the cross-modal attention mechanism and the AMF module. Second, we compare the performance differences between the multimodal teacher and the unimodal teacher model. Finally, we analyze the impact of different loss functions on the performance of the student model. These experiments clarify the specific contribution of each component to overall performance. The experimental results are shown in Table 4.

By sequentially eliminating the cross-modal attention mechanism and the AMF module to observe changes in overall performance, their importance in the teacher model is validated. During this process, the structure of the student model is kept unchanged, and cross-modal knowledge distillation is performed again using the ablated teacher model. The experimental results show that removing either the cross-modal attention mechanism or the AMF module led to a decline in the teacher model's performance, along with a notable decrease in the performance of the student model obtained through knowledge distillation. This indicates that the cross-modal attention mechanism and the AMF module are essential for enhancing the expressive capability of the teacher model. The cross-modal attention mechanism alleviates the distribution differences between different modalities, while the AMF module ensures the recognition performance of the teacher model. Both contribute to information transfer to the student model.

To further validate the importance of the multimodal model in the knowledge distillation process, the PA modality is removed from the multimodal teacher model, simplifying it to the EEG-based unimodal teacher model. The structure of the new EEG-based unimodal teacher model is kept consistent with that of the PA-based unimodal student model, and cross-modal knowledge distillation is then performed again. The experimental results indicate that when the teacher model is transformed from multimodal to unimodal, its own performance decreases, and the performance of the student model obtained through knowledge distillation also declines. This may be because, compared to the multimodal teacher model, the EEG-based teacher lacks rich knowledge and exhibits modality heterogeneity with the PA-based student.

To assess the importance of different loss functions in the multimodal-to-unimodal knowledge distillation process, the loss functions L_{KL}, L_{COS}, and L_{CE} are

Table 4 Ablation experiment results of the proposed network on AADP Dataset and EFVP Dataset (mean ± standard deviation). "-teacher" represents the ablated teacher model, and "-student" denotes the corresponding student model obtained through distillation from this teacher. In Variant 1, the CMA and AMF components of the teacher model are ablated, where "w/o CMA" indicates the removal of cross-modal attention, "w/o AMF1" replaces AMF with element-wise addition, and "w/o AMF2" replaces it with feature-level concatenation. Variant 2 uses an EEG-based unimodal teacher model for distillation. Variant 3 ablates the loss function in the student model during knowledge distillation. The best results in terms of the mean accuracy and mean F1 score are highlighted in boldface.

(%)

	Method	AADP dataset		EFVP dataset	
		Accuracy	F1 score	Accuracy	F1 score
Variant 1	W/o CMA-teacher	90.16 ± 3.79	90.09 ± 3.82	84.25 ± 9.86	83.70 ± 10.32
	W/o CMA-student	78.83 ± 3.84	78.16 ± 4.27	74.33 ± 14.03	71.95 ± 17.02
	W/o AMF1-teacher	90.21 ± 3.89	90.27 ± 3.81	84.42 ± 9.50	83.66 ± 10.14
	W/o AMF1-student	79.11 ± 0.67	78.32 ± 5.00	74.17 ± 13.05	72.08 ± 15.57
	W/o AMF2-teacher	89.69 ± 5.87	89.51 ± 6.42	85.33 ± 10.42	84.71 ± 10.98
	W/o AMF2-student	78.35 ± 1.99	77.44 ± 2.26	74.08 ± 14.73	72.11 ± 16.94
Variant 2	W/o PA-teacher	86.68 ± 6.27	86.48 ± 6.55	83.08 ± 9.17	82.50 ± 9.62
	W/o PA-student	78.31 ± 3.93	77.62 ± 4.04	75.67 ± 15.53	72.90 ± 19.22
Variant 3	W/o L_{KL}	78.96 ± 3.60	80.61 ± 4.72	74.58 ± 13.66	74.20 ± 13.77
	W/o L_{COS}	78.29 ± 2.04	77.73 ± 1.84	74.25 ± 14.78	78.00 ± 11.94
	W/o L_{CE}	78.54 ± 4.98	78.14 ± 1.91	75.58 ± 14.05	76.40 ± 15.50
Our method	KD student	**81.45** ± 3.29	**80.95** ± 3.52	**78.50** ± 12.46	**77.53** ± 13.26

removed sequentially, and the model's performance was evaluated. The experimental results show that removing any of the loss terms leads to a decrease in model performance, indicating that each loss term plays an indispensable role in the training process. L_{KL} transfers the knowledge of the teacher model through soft labels. L_{COS} helps maintain angular consistency between the teacher and student models. L_{CE} ensures that the student model learns the correct classification information directly from hard labels. Compared to using a single loss term alone, the complete multi-loss function setup better guides the student model to learn richer representations, thereby improving its overall performance.

4.7 Discussion on robustness and practicality

Our proposed framework demonstrates strong robustness and practical value in real-world deployment scenarios. The student model distilled from the multimodal teacher is evaluated under a subject-independent cross-validation setting, where all trials from unseen individuals are used exclusively for testing. This ensures that the performance gains are not limited to subject-specific patterns but instead generalize across individuals, even in the presence of inter-subject variability and potential session-related shifts.

The knowledge distillation framework is not only effective but also practically necessary. While one could theoretically train a large PA-only model on massive unimodal data, collecting such large-scale PA datasets is often costly and time-consuming. In contrast, our method leverages a relatively small set of high-quality multimodal data to train a teacher model that captures rich, depression-relevant patterns. These learned representations are then efficiently transferred to the lightweight, PA-only student model. Empirical results show that our distilled student model significantly outperforms the same-structure model trained solely on PA data, highlighting the effectiveness of cross-modal knowledge transfer.

Given the practical limitations of EEG collection—such as the need for expensive equipment, professional setup, and subject compliance—our approach of deploying a high-performance, EEG-free model offers a much more feasible solution for scalable depression screening in community or low-resource settings.

Nonetheless, individual differences and session dependence remain challenging. While our model shows strong generalization overall, performance may still degrade on subjects with highly atypical patterns. Future work may explore subject-adaptive mechanisms, to further enhance robustness without sacrificing deployment efficiency.

5 Conclusion and Future Work

We propose a novel cross-modal knowledge distillation method for depression recognition. In this method, the multimodal teacher model uses a cross-modal interaction module and an adaptive modality fusion module to deeply explore the intrinsic features and interactions between the EEG and PA modalities, capturing rich multimodal information. Subsequently, this information is transferred to the PA-based unimodal student model through knowledge distillation, enhancing the student model's performance. To further clarify the mechanisms driving improvements in the student model, we propose E-GCAM, an explainability method combining information entropy theory with Grad-CAM. Quantitative results indicate that the time series model under knowledge distillation encodes more useful information, consistent with the observed performance improvements in the student model. E-GCAM reveals the mechanisms behind the performance improvements of the student model and is equally applicable for analyzing performance in other time series models. Compared to other baseline models, the proposed method achieved the best results on both datasets. Our work significantly reduces the reliance on EEG signals for depression recognition, thereby expanding the applicability and potential use cases of depression recognition models. Additionally, this explainability method can also be applied to other time series tasks, and when combined with specific scenarios, it facilitates a more comprehensive analysis of model performance. Moreover, the distilled student model does not require retraining for individual subjects and may be used as an auxiliary tool for depression recognition.

Future research will focus on investigating the impact of modality heterogeneity on cross-modal knowledge distillation, an underexplored factor that is crucial for understanding how modality differences influence knowledge transfer efficiency. A better understanding of this aspect will provide key insights for optimizing the distillation process. Additionally, we aim to refine the knowledge point selection mechanism

to more effectively identify and transfer relevant information, thereby improving the performance of knowledge distillation.

Acknowledgment

This work was supported in part by the Brain Science and Brain-like Intelligence Technology-National Science and Technology Major Project (Nos. 2022ZD0208500 and 2021ZD0202000), the National Key Research and Development Program of China (No. 2019YFA0706200), the National Natural Science Foundation of China (Nos. 62372216 and 62102172), the Key Program of Natural Science Foundation of Gansu Province (No. 22JR5RA410), the Gansu Province Science and Technology Program (No. 22JR5RA489), and the Fundamental Research Funds for the Central Universities (No. lzujbky-2022-10).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] A. Pampouchidou, P. G. Simos, K. Marias, F. Meriaudeau, F. Yang, M. Pediaditis, and M. Tsiknakis, Automatic assessment of depression based on visual cues: A systematic review, *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 445–470, 2019.

[2] M. J. Murphy and M. J. Peterson, Sleep disturbances in depression, *Sleep Med. Clin.*, vol. 10, no. 1, pp. 17–23, 2015.

[3] M. Jansson-Fröjmark and K. Lindblom, A bidirectional relationship between anxiety and depression, and insomnia? A prospective study in the general population, *J. Psychosom. Res.*, vol. 64, no. 4, pp. 443–449, 2008.

[4] S. Borgman, I. Ericsson, E. K. Clausson, and P. Garmy, The relationship between reported pain and depressive symptoms among adolescents, *J. Sch. Nurs.*, vol. 36, no. 2, pp. 87–93, 2020.

[5] J. P. Lépine and M. Briley, The epidemiology of pain in depression, *Hum. Psychopharmacol. Clin. Exp.*, vol. 19, no. S1, pp. S3–S7, 2004.

[6] Z. A. E. Sarhan, H. A. El Shinnawy, M. E. Eltawil, Y. Elnawawy, W. Rashad, and M. Saadeldin Mohammed, Global functioning and suicide risk in patients with depression and comorbid borderline personality disorder, *Neurol. Psychiatry Brain Res.*, vol. 31, pp. 37–42, 2019.

[7] A. J. Mitchell, A. Vaze, and S. Rao, Clinical diagnosis of depression in primary care: A meta-analysis, *Lancet*, vol. 374, no. 9690, pp. 609–619, 2009.

[8] J. M. Bostwick and S. Rackley, Recognizing mimics of depression: The '8 Ds', *Curr. Psychiatry*, vol. 11, no. 6, pp. 31–36, 2012.

[9] H. S. Chiang, M. Y. Chen, and L. S. Liao, Cognitive depression detection cyber-medical system based on EEG analysis and deep learning approaches, *IEEE J. Biomed. Health Inf.*, vol. 27, no. 2, pp. 608–616, 2023.

[10] I. M. Spyrou, C. Frantzidis, C. Bratsas, I. Antoniou, and P. D. Bamidis, Geriatric depression symptoms coexisting with cognitive decline: A comparison of classification methodologies, *Biomed. Signal Process. Control*, vol. 25, pp. 118–129, 2016.

[11] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, H. Adeli, and D. P. Subha, Automated EEG-based screening of depression using deep convolutional neural network, *Comput. Methods Programs Biomed.*, vol. 161, pp. 103–113, 2018.

[12] N. Padfield, J. Zabalza, H. Zhao, V. Masero, and J. Ren, EEG-based brain-computer interfaces using motorimagery: Techniques and challenges, *Sensors*, vol. 19, no. 6, p. 1423, 2019.

[13] G. J. Siegle, S. R. Steinhauer, E. S. Friedman, W. S. Thompson, and M. E. Thase, Remission prognosis for cognitive therapy for recurrent depression using the pupil: Utility and neural correlates, *Biol. Psychiatry*, vol. 69, no. 8, pp. 726–733, 2011.

[14] J. Wang, Y. Fan, X. Zhao, and N. Chen, Pupillometry in Chinese female patients with depression: A pilot study, *Int. J. Environ. Res. Public Health*, vol. 11, no. 2, pp. 2236–2243, 2014.

[15] K. Simonyan, A. Vedaldi, and A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv preprint arXiv: 1312.6034, 2013.

[16] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, Learning deep features for discriminative localization, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 2921–2929.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in *Proc. 31st Int. Conf. Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 6000–6010.

[18] S. M. Lundberg and S. I. Lee, A unified approach to interpreting model predictions, in *Proc. 31st Int. Conf. Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 4768–4777.

[19] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 1135–1144.

[20] S. Peng, R. Zeng, H. Liu, L. Cao, G. Wang and J. Xie, Deep broad learning for emotion classification in textual conversations, *Tsinghua Science and Technology*, vol. 29, no. 2, pp. 481–491, 2024.

[21] S. Sardari, B. Nakisa, M. N. Rastgoo, and P. Eklund, Audio based depression detection using Convolutional Autoencoder, *Expert Syst. Appl.*, vol. 189, p. 116076, 2022.

[22] A. Y. Kim, E. H. Jang, S. Kim, K. W. Choi, H. J. Jeon, H. Y. Yu, and S. Byun, Automatic detection of major depressive disorder using electrodermal activity, *Sci. Rep.*, vol. 8, no. 1, p. 17030, 2018.

[23] X. Liu, H. Shen, H. Li, Y. Tao, and M. Yang, Multimodal depression detection based on self-attention network with facial expression and pupil, *IEEE Trans. Computat. Soc. Syst.*, vol. 12, no. 1, pp. 64–76, 2025.

[24] J. Chen, Y. Hu, Q. Lai, W. Wang, J. Chen, H. Liu, G. Srivastava, A. K. Bashir, and X. Hu, IIFDD: Intra and inter-modal fusion for depression detection with multi-modal information from Internet of Medical Things, *Inf. Fusion*, vol. 102, p. 102017, 2024.

[25] H. Fan, X. Zhang, Y. Xu, J. Fang, S. Zhang, X. Zhao, and J. Yu, Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals, *Inf. Fusion*, vol. 104, p. 102161, 2024.

[26] Z. Ning, H. Hu, L. Yi, Z. Qie, A. Tolba, and X. Wang, A depression detection auxiliary decision system based on multi-modal feature-level fusion of EEG and speech, *IEEE Trans. Consum. Electron*, vol. 70, no. 1, pp. 3392–3402, 2024.

[27] X. Zhang, J. Shen, Z. U. Din, J. Liu, G. Wang, and B. Hu, Multimodal depression detection: Fusion of electroencephalography and paralinguistic behaviors using a novel strategy for classifier ensemble, *IEEE J. Biomed. Health Inform.*, vol. 23, no. 6, pp. 2265–2275, 2019.

[28] H. Cai, Z. Qu, Z. Li, Y. Zhang, X. Hu, and B. Hu, Feature-level fusion approaches based on multimodal EEG data for depression recognition, *Inf. Fusion*, vol. 59, pp. 127–138, 2020.

[29] J. Zhu, C. Yang, X. Xie, S. Wei, Y. Li, X. Li, and B. Hu, Mutual information based fusion model (MIBFM): Mild depression recognition using EEG and pupil area signals, *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2102–2115, 2023.

[30] Y. Qin, Y. Zhang, Y. Zhang, S. Liu, and X. Guo, Application and development of EEG acquisition and feedback technology: A review, *Biosensors*, vol. 13, no. 10, p. 930, 2023.

[31] Y. Lu, W. L. Zheng, B. Li, and B. L. Lu, Combining eye movements and EEG to enhance emotion recognition, in *Proc. 24th Int. Conf. Artificial Intelligence*, Buenos Aires, Argentina, 2015, pp. 1170–1176.

[32] S. Alghowinem, R. Goecke, M. Wagner, G. Parker, and M. Breakspear, Eye movement analysis for depression detection, in *Proc. 2013 IEEE Int. Conf. Image Processing*, Melbourne, Australia, 2013, pp. 4220–4224.

[33] R. Shen, Q. Zhan, Y. Wang, and H. Ma, Depression detection by analysing eye movements on emotional images, in *Proc. 2021 IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Toronto, Canada, 2021, pp. 7973–7977.

[34] G. Hinton, O. Vinyals, and J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv: 1503.02531, 2015.

[35] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio. FitNets: Hints for thin deep nets, arXiv preprint arXiv: 1412.6550, 2015.

[36] E. Shang, H. Liu, J. Zhang, R. Zhao and J. Du, Ensemble knowledge distillation for federated semi-supervised image classification, *Tsinghua Science and Technology*, vol. 30, no. 1, pp. 112−123, 2025

[37] S. Gupta, J. Hoffman, and J. Malik, Cross modal distillation for supervision transfer, in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 2827–2836.

[38] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, Emotion recognition in speech using cross-modal transfer in the wild, in *Proc. 26th ACM Int. Conf. Multimedia*, Seoul, Republic of Korea, 2018, pp. 292–301.

[39] S. Zhang, C. Tang, and C. Guan, Visual-to-EEG cross-modal knowledge distillation for continuous emotion recognition, *Pattern Recognit.*, vol. 130, p. 108833, 2022.

[40] Y. Liu, Z. Jia, and H. Wang, EmotionKD: A cross-modal knowledge distillation framework for emotion recognition based on physiological signals, in *Proc. 31st ACM Int. Conf. Multimedia*, Ottawa, Canada, 2023, pp. 6122–6131.

[41] X. Zhou, C. Liu, J. Zhou, Z. Wang, L. Zhai, Z. Jia, C. Guan, and Y. Liu, Interpretable and robust AI in EEG systems: A survey, arXiv preprint arXiv: 2304.10755, 2025.

[42] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, Gradient-based attribution methods, in *Explainable AI*: *Interpreting*, *Explaining and Visualizing Deep Learning*, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K. R. Müller, eds. Cham, Switzerland: Springer, 2019, pp. 169–191.

[43] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, Explainable AI: A review of machine learning interpretability methods, *Entropy*, vol. 23, no. 1, p. 18, 2020.

[44] S. Jain and B. C. Wallace, Attention is not explanation, in *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Minneapolis, MN, USA, 2019, pp. 3543–3556.

[45] H. Ma, H. Zhang, F. Zhou, Y. Zhang, and Q. Zhang, Quantification and analysis of layer-wise and pixel-wise information discarding, in *Proc. 39th Int. Conf. Machine Learning*, Baltimore, MD, USA, 2022, pp. 14664–14698.

[46] R. Shwartz-Ziv and N. Tishby, Opening the black box of deep neural networks via information, arXiv preprint arXiv: 1703.00810, 2017.

[47] N. Wolchover, New theory cracks open the black box of deep learning, https://www.quantamagazine.org/new-theory-cracks-open-the-black-box-of-deep-learning-

20170921/, 2018.

[48] X. Zhang, X. Wei, Z. Zhou, Q. Zhao, S. Zhang, Y. Yang, R. Li, and B. Hu, Dynamic alignment and fusion of multimodal physiological patterns for stress recognition, *IEEE Trans. Affect. Comput.*, vol. 15, no. 2, pp. 685–696, 2024.

[49] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, Improved knowledge distillation via teacher assistant, in *Proc. 34th AAAI Conf. Artificial Intelligence*, New York, NY, USA, 2020, pp. 5191–5198.

[50] E. Guermazi, A. Mdhaffar, M. Jmaiel, and B. Freisleben, MulKD: Multi-layer knowledge distillation via collaborative learning, *Eng. Appl. Artif. Intell.*, vol. 133, p. 108170, 2024.

[51] G. J. Siegle, E. Granholm, R. E. Ingram, and G. E. Matt, Pupillary and reaction time measures of sustained processing of negative information in depression, *Biol. Psychiatry*, vol. 49, no. 7, pp. 624–636, 2001.

[52] M. Oliva and A. Anikin, Pupil dilation reflects the time course of emotion recognition in human vocalizations, *Sci. Rep.*, vol. 8, no. 1, p. 4871, 2018.

[53] B. Hu, D. Majoe, M. Ratcliffe, Y. Qi, Q. Zhao, H. Peng, D. Fan, F. Zheng, M. Jackson, and P. Moore, EEG-based cognitive interfaces for ubiquitous applications: Developments and challenges, *IEEE Intell. Syst.*, vol. 26, no. 5, pp. 46–53, 2011.

[54] H. D. Critchley, J. Tang, D. Glaser, B. Butterworth, and R. J. Dolan, Anterior cingulate activity during error and autonomic response, *NeuroImage*, vol. 27, no. 4, pp. 885–895, 2005.

[55] F. Tao, X. Ge, W. Ma, A. Esposito, and A. Vinciarelli, Multi-local attention for speech-based depression detection, in *Proc. 2023 IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Rhodes Island, Greece, 2023, pp. 1–5.

[56] I. Y. Susanto, T. Y. Pan, C. W. Chen, M. C. Hu, and W. H. Cheng, Emotion recognition from galvanic skin response signal based on deep hybrid neural networks, in *Proc. 2020 Int. Conf. Multimedia Retrieval*, Dublin, Ireland, 2020, pp. 341–345.

[57] P. Hou, X. Li, J. Zhu, and B. Hu, A lightweight convolutional transformer neural network for EEG-based depression recognition, *Biomed. Signal Process. Control*, vol. 100, p. 107112, 2025.

[58] T. C. Sweeney-Fanelli and M. H. Imtiaz, ECG-based automated emotion recognition using temporal convolution neural networks, *IEEE Sensors J.*, vol. 24, no. 18, pp. 29039–29046, 2024.

[59] X. Yan, L. M. Zhao, and B. L. Lu, Simplifying multimodal emotion recognition with single eye movement modality, in *Proc. 29th ACM Int. Conf. Multimedia*, Virtual Event, 2021, pp. 1057–1063.

[60] M. Anandakumar, J. Pradeepkumar, S. L. Kappel, C. U. S. Edussooriya, and A. C. de Silva, A knowledge distillation framework for enhancing ear-EEG based sleep staging with scalp-EEG data, in *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics*, Oahu, HI, USA, 2023, pp. 514–519.

[61] C. Pham, T. Hoang, and T. T. Do, Collaborative multi-teacher knowledge distillation for learning low bit-width deep neural networks, in *Proc. 2023 IEEE/CVF Winter Conf. Applications of Computer Vision*, Waikoloa, HI, USA, 2023, pp. 6424–6432.

[62] X. Zhang, J. Pan, J. Shen, Z. U. Din, J. Li, D. Lu, M. Wu, and B. Hu, Fusing of electroencephalogram and eye movement with group sparse canonical correlation analysis for anxiety detection, *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 958–971, 2022.

[63] J. L. Qiu, W. Liu, and B. L. Lu, Multi-view emotion recognition using deep canonical correlation analysis, in *Proc. 25th Int. Conf. Neural Information Processing*, Siem Reap, Cambodia, 2018, pp. 221–231.

[64] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series, *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 4, pp. 758–769, 2018.

[65] B. Fu, C. Gu, M. Fu, Y. Xia, and Y. Liu, A novel feature fusion network for multimodal emotion recognition from EEG and eye movement signals, *Front. Neurosci.*, vol. 17, p. 1234162, 2023.

[66] J. Yin, M. Wu, Y. Yang, P. Li, F. Li, W. Liang, and Z. Lv, Research on multimodal emotion recognition based on fusion of electroencephalogram and electrooculography, *IEEE Trans. Instrum. Meas.*, vol. 73, p. 6502012, 2024.

[67] M. Bayer, K. Ruthmann, and A. Schacht, The impact of personal relevance on emotion processing: Evidence from event-related potentials and pupillary responses, *Soc. Cogn. Affect. Neurosci.*, vol. 12, no. 9, pp. 1470–1479, 2017.

[68] C. E. Waugh, E. Z. Shing, and B. M. Avery, Temporal dynamics of emotional processing in the brain, *Emot. Rev.*, vol. 7, no. 4, pp. 323–329, 2015.

Yizhou Li received the bachelor degree from Lanzhou University, China, in 2021. He is currently a PhD candidate at Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, China. His research interests mainly focus on affective computing based on physiological data.

Yu Cheng received the BEng degree in computer science and technology from Shandong University, China, in 2022. He is currently a master student at Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, China. His research interests mainly focus on depression recognition based on physiological data.

Bin Hu received the PhD degree from the Institute of Computing Technology, Chinese Academy of Sciences, China, in 1998. He is currently with Lanzhou University, China, where he is affiliated with the Laboratory of Special Functional Materials and Structural Design (Ministry of Education) and the Engineering Research Center of Open Source Software and Real-Time System (Ministry of Education). He was a recipient of many research awards, including the 2014 China Overseas Innovation Talent Award, the 2016 Chinese Ministry of Education Technology Invention Award, the 2018 Chinese National Technology Invention Award, and the 2019 WIPO-CNIPA Award for Chinese Outstanding Patented Invention. He is also the TC co-chair of computational psychophysiology in the IEEE Systems, Man, and Cybernetics Society (SMC), the TC co-chair of cognitive computing in IEEE SMC, and the vice-chair of the TC 9.1. Economic, Business, and Financial Systems on Social Media at the International Federation of Automatic Control. He is also a member-at-large of the ACM China Council and the vice-chair of the China Committee of the International Society for Social Neuroscience. His main research areas include affective computing, pervasive computing, psychophysiological computing, and artificial intelligence.

Xiaowei Li received the MS degree in computer software and theory in 2015 and the PhD degree in software engineering in 2015 from Lanzhou University, China, in 2005 and 2015, respectively. He is currently a professor at Lanzhou University, China. He is also with Shandong Academy of intelligent Computing Technolog, China. His research interests focus on affective computing, ubiquitous computing, biosignal processing, and data mining.

Jing Zhu received the MS degree in economics and finance from Xi'an Jiaotong University, China, in 2010, and the PhD degree in computer application technology from Lanzhou University, China, in 2021. She is currently an associate professor at Lanzhou University, China. Her research interests mainly focus on affective computing, ubiquitous computing, and data mining.