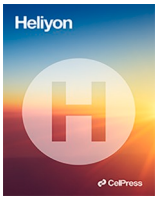
Heliyon 10 (2024) e25959

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Enhancing multimodal depression diagnosis through representation learning and knowledge transfer

Shanliang Yang [a], Lichao Cui [a], Lei Wang [a], Tao Wang [a], Jiebing You [b,*]

[a] School of Computer Science and Technology, Shandong University of Technology, Zibo, 255000, China
[b] Department of Neurology, Zibo Central Hospital, Zibo, 255036, China

ARTICLE INFO

Keywords:
Depression diagnosis
Multimodal data
Representation learning
Knowledge transfer techniques

ABSTRACT

Depression is a complex mental health disorder that presents significant challenges in diagnosis and treatment. This study proposes an innovative approach, leveraging artificial intelligence advancements, to enhance multimodal depression diagnosis. The diagnosis of depression often relies on subjective assessments and clinical interviews, leading to potential biases and inaccuracies. Additionally, integrating diverse data modalities, such as textual, imaging, and audio information, poses technical challenges due to data heterogeneity and high dimensionality. To address these challenges, this paper proposes the RLKT-MDD (Representation Learning and Knowledge Transfer for Multimodal Depression Diagnosis) model framework. Representation learning enables the model to autonomously discover meaningful patterns and features from diverse data sources, surpassing traditional feature engineering methods. Knowledge transfer facilitates the effective transfer of knowledge from related domains, improving the model's performance in depression diagnosis. Furthermore, we analyzed the interpretability of the representation learning process, enhancing the transparency and trustworthiness of the diagnostic process. We extensively experimented with the DAIC-WOZ dataset, a diverse collection of multimodal data from clinical settings, to evaluate our proposed approach. The results demonstrate promising outcomes, indicating significant improvements over conventional diagnostic methods. Our study provides valuable insights into cutting-edge techniques for depression diagnosis, enabling more effective and personalized mental health interventions.

1. Introduction

Depression is a complex mental health disorder that significantly impacts a large portion of the global population. According to the World Health Organization (WHO), approximately 280 million people worldwide are diagnosed with depression [1]. However, diagnosing and treating depression remains a crucial challenge in the field of mental health. The current subjective assessment approaches, such as the Personal Health Questionnaire (PHQ) [2], the Hamilton Depression Inventory (HAMD) [3], often lead to over-diagnosis or misdiagnosis, highlighting the critical need for accurate diagnosis to facilitate effective interventions and treatments for patients. Given the high prevalence and complexity of depression, researchers have turned to explore artificial intelligence-based depression detection methods [4–6]. In recent years, there has been a surge in AI-based depression detection research, which aims to develop diagnostic methods that are more objective, accurate, and reliable. This advancement ultimately seeks to improve treatment

* Corresponding author.
E-mail address: youjiebing2023@163.com (J. You).

https://doi.org/10.1016/j.heliyon.2024.e25959
Received 15 January 2024; Received in revised form 30 January 2024; Accepted 5 February 2024
Available online 10 February 2024

options for individuals with depression. In this context, our study proposes an innovative approach that leverages representation learning and knowledge transfer techniques to enhance multimodal depression diagnosis.

Symptoms of depression in individuals often manifest as prolonged feelings of sadness or loss of pleasure, which can be recognized through facial expressions, vocal tone, and language expression [15]. Consequently, current research is focused on designing artificial intelligence models to identify emotional information from monitoring data sources, aiming to determine whether a patient is suffering from depression. Yang et al. [7] explored audio feature representations for depression diagnosis by employing learnable time-domain filters to acquire biologically meaningful acoustic features. Additionally, they utilized multi-scale spectral attention learning to retain the useful frequency sub-bands. W. C. de Melo et al. [8] effectively captured facial information associated with depressive behaviors from videos by utilizing parallel convolutional layers with varying temporal depths and receptive field sizes. In addition to utilizing audio and visual information, scholars have also commenced leveraging the complementarity of multimodal data to enhance depression diagnosis. Fang et al. [6] conduct a comprehensive analysis of depression-related audio-visual and textual data, introducing a multimodal fusion model augmented with a multi-level attention mechanism (MFM-Att) aimed at enhancing depression detection. Despite certain advancements in AI-based depression diagnosis technology, it continues to face challenges, particularly in extracting discriminative features specific to depression. The principle of multimodal depression diagnosis is illustrated in Fig. 1, where a fusion of visual, audio, and text features is employed to determine the presence of depression. The most crucial aspects highlighted within the schematic are the challenges of multimodal feature representation and the role of background knowledge in supporting depression diagnosis. The diagnostic process for depression necessitates prior knowledge, and the limitation of domain expertise hinders diagnostic performance.

Representation learning enables the automatic acquisition of meaningful representations from data, alleviating the burden of feature engineering and uncovering latent patterns within the data, thereby facilitating improved task processing. For instance, Q. Zheng et al. [10] employ multi-view representation learning to explore valuable information acquired from feature representations and multi-view subspaces. Similarly, M. Dorkenwald et al. [9] utilize representation learning techniques to extract semantic information and motion patterns from videos. Knowledge transfer, on the other hand, permits the acquisition of background knowledge from external knowledge repositories or pre-trained models, assisting the model in comprehending and effectively utilizing contextual information to enhance task performance. For example, J. Yu and G. Liu [11] first extract rule-based knowledge through symbolic models and then integrate it into deep neural network models to obtain a sparse network enriched with knowledge. In light of this, this paper attempts to leverage representation learning and knowledge transfer techniques to enhance the effectiveness of the multimodal depression diagnosis model.

To address the aforementioned technical challenges associated with depression diagnosis, this paper proposes the RLKT-MDD model framework. This framework integrates representation learning with knowledge transfer to enhance the accuracy of depression diagnosis. By employing representation learning, it extracts critical feature representations from multimodal data, ensuring the integrity of information and preserving essential details through data reconstruction. Concurrently, the framework leverages knowledge transfer, utilizing the insights and patterns from pre-trained models. This approach facilitates the application of pre-existing knowledge to the task of diagnosing depression, thereby enabling the framework to make more precise and informed assessments, significantly improving the reliability and effectiveness of depression diagnosis. Specifically, the data preprocessing starts by extracting basic features from the multimodal data, including visual, audio, and textual data, utilizing the Facial Action Coding System [12], COAVREP [13], and BERT-style language model [14]. These extracted features are then fed into a self-encoding representation learning model to acquire unimodal latent representations through feature vector reconstruction. Subsequently, a modality fusion unit is employed to learn the complementary relationships between different modalities, resulting in a multimodal fused feature representation. Finally, knowledge transfer techniques are employed to integrate knowledge from pre-trained models and depression-related datasets into the model, thereby further enhancing the diagnostic performance of the model.

The main contributions of this paper are summarized as follows:

(1) This paper proposes a multimodal depression diagnosis model framework, RLKT-MDD, which integrates representation learning and knowledge transfer. The model addresses the problem of learning representative features for depression-related characteristics through representation learning, resulting in more discriminative multimodal depression features.

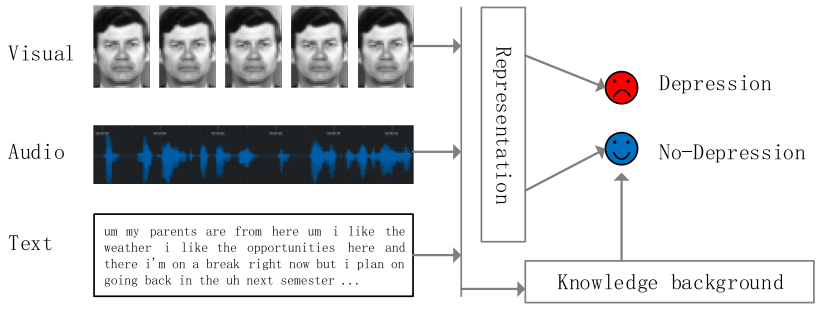


Fig. 1. The schematic of multimodal depression diagnosis incorporates multimodal representations and background knowledge.

(2) To tackle the lack of prior knowledge in depression diagnosis models, the approach of model knowledge transfer and depression-related data knowledge transfer is employed to augment depression prior knowledge and enhance the diagnostic performance of the model.
(3) Extensive experiments are conducted on the DAIC-WOZ dataset, yielding favorable experimental results, thus validating the effectiveness of the designed representation learning and knowledge transfer modules in this study.

2. Related work

2.1. Depression diagnosis

Depression diagnosis is the process of identifying and assessing the presence and severity of depressive symptoms in an individual's mental health. In traditional practice, health practitioners rely on clinical interviews and self-report scales and inventories (Self-RIs) to assess individuals for clinical depression. The most commonly employed scales in clinical interview settings are the Diagnostic and Statistical Manual of Mental Disorders (DSM) [16] and the Hamilton Depression Rating Scale (HDRS) [17]. The most commonly utilized Self-Reported Instruments (Self-RIs) include the Patient Health Questionnaire (PHQ) [2] and the Hospital Anxiety and Depression Scale (HADS) [18]. However, clinical interviews rely on the diagnostic experience of clinicians, making them susceptible to subjective biases; self-assessment scales, on the other hand, overlook the individualized characteristics of patients and the clinical significance of symptoms. With the rapid advancement of artificial intelligence technology, experts and scholars have begun researching methods for the automated identification of depression, seeking objective diagnostic approaches that are not susceptible to subjective biases.

Depression symptoms can be manifested through various types of data, including textual expressions, facial expressions, voice characteristics, physiological signals, social media activity, and altered cognitive and behavioral patterns. Many researchers have obtained effective results in depression diagnosis from various data sources. S. K. Priya et al. [19] employ natural language processing techniques and machine learning models to extract features from textual data for the identification of depression. Y. Guo et al. [20] conducted research on diagnosing depression from video data. They designed a temporal dilated convolutional network to extract depression-related features from the videos. W. Yang [7] investigated speech-based depression detection and employed an interpretable perspective to learn effective features for depression detection. E. Rejaibi et al. [21] extracted high-level MFCC features from low-level speech features and utilized the LSTM network for clinical depression detection. J. Zhu et al. [22] extracted depression-related features from EEG data and eye-tracking data for mild depression recognition. S. A. Qureshi et al. [23] proposed that emotion intensity recognition and depression detection share similar cognitive patterns. They employed multi-task learning to enhance the effectiveness of depression detection. S. Alghowinem et al. [24] detected depression using data from eye gaze, head posture, and language.

While uni-modal depression diagnosis has achieved certain efficacy, multimodal data possesses evident complementarity and relative advantages. Research on multimodal depression diagnosis has become a hot topic in the field, attracting a considerable number of researchers. Y. Wang et al. [29] extracted users' textual, image, and behavioral data from social media for predicting depression detection. N. Seneviratne and C. Espy-Wilson [28] improved the accuracy of multimodal depression diagnosis by fusing speech features and transcribed text features. M. Muzammel et al. [25] investigated the impact of neural network architectures and multimodal fusion methods on multimodal depression diagnosis tasks. They conducted comparative experiments using the DAIC-WOZ benchmark dataset. W. Xie et al. [26] designed a CNN-LSTM neural network model to extract facial expression features and integrated self-assessment scores to obtain depression diagnosis results. T. Chen et al. [27] leveraged a graph neural network model to capture both the heterogeneity and homogeneity among multimodal data, extracting inter-class and intra-class features. Although the accuracy of depression diagnosis has been enhanced through multimodal fusion, challenges such as the difficulty in extracting and expressing depression-related features, as well as the lack of prior knowledge in depression diagnostic models, remain significant hurdles in multimodal depression diagnosis. This paper primarily focuses on concrete research in the areas of representation learning and knowledge transfer to address these issues.

2.2. Representation learning

Representation learning aims to automatically learn feature representations of data in order to better capture information and patterns within the data. Reference [30] and reference [31] approach deep representation learning from a multi-view perspective, aiming to reduce redundancy while enhancing the discriminative power of features. Currently, many researchers are also leveraging representation learning in the task of depression diagnosis to enhance the capability of representing depression-related features. S. Song et al. [31] introduced two innovative spectral representations to effectively capture and represent the multi-scale temporal dynamics of expressive behavior at the video level, followed by utilizing these spectral representations for depression analysis. X. Zhou et al. [32] employed label distribution learning and deep metric learning to enhance the model's representational capacity, enabling the acquisition of more discriminative spatiotemporal features. Z. Han et al. [33] designed a self-supervised learning approach to learn depression feature representation from speech features and phonemes. M. Niu et al. [34] investigated spatiotemporal representation learning by segmenting the speech amplitude spectrum/video into fixed-length segments and employing a neural network model that integrates spatial and temporal information through an attention mechanism. Autoencoder is an unsupervised learning neural network model and a significant representation learning method. S. Sardari et al. [5] introduced an end-to-end Convolutional Neural Network-based Autoencoder to learn highly relevant and discriminative features from speech data.

2.3. Knowledge transfer

Knowledge transfer refers to the process of acquiring knowledge from one task, domain, or model and applying it to another task, domain, or model. Transferred knowledge can be explicit external knowledge or internal knowledge stored within the parameter space of the pretrained model. Z. Guo et al. [35] constructed a depression lexicon based on domain knowledge of depression and an emotion dictionary, aiming to better extract lexicon features relevant to depression. Y. Lin et al. [36] emphasized the significance of background knowledge by integrating the knowledge of psychological screening tools and diagnostic criteria into the depression diagnostic model. K. Yang et al. [37] initially extracted mental state knowledge from commonsense knowledge and then introduced a knowledge-aware mentalization module designed to attend to the most relevant knowledge aspects. As deep learning technology continues to mature, pretrained models have been increasingly utilized as knowledge-rich agents for various downstream tasks, including depression diagnosis. Here, we present several illustrative examples where pretrained models are employed for knowledge transfer to enhance the effectiveness of depression diagnosis. W. Wu et al. [38,40] transferred speech knowledge and emotion knowledge to depression diagnosis by fine-tuning the foundation models of automatic speech recognition and emotion recognition. Similarly, Y. Dong and X. Yang [39] utilized pretrained models to extract speaker identification features and speech emotion features, employing differences in speech and emotion for depression diagnosis. These researches has shown that employing pretrained models can transfer latent knowledge to new tasks, enhancing the recognition performance of these new tasks.

3. Methodology

In this section, we delve into the comprehensive methodology and model employed in our study to enhance multimodal depression diagnosis. We provide a detailed exploration of the key elements comprising our framework, encompassing architectural design, data preprocessing, multimodal feature fusion, representation learning strategy, knowledge transfer approach, and the depression diagnosis network. Each of these components has been strategically designed to contribute to the overall effectiveness of depression diagnosis. Our aim is to provide a thorough understanding of how our methodology addresses the complexities of depression diagnosis through the seamless integration of representation learning and knowledge transfer. Subsequent sections will elaborate on these aspects, offering insights into the intricate workings of our proposed framework.

3.1. The overall architecture of RLKT-MDD

We present the architectural design of our proposed framework, RLKT-MDD (Representation Learning and Knowledge Transfer for Multimodal Depression Diagnosis). We provide an overview of the high-level structure, outlining the components and their interactions within the framework. We detail how different modules are interconnected to facilitate the seamless integration of representation learning and knowledge transfer for accurate depression diagnosis.

The overall architecture of the framework is illustrated in Fig. 2, comprising five levels: data feature preprocessing, modality representation learning, inter-modal fusion unit, depression-related knowledge transfer, and depression diagnosis. The multimodal

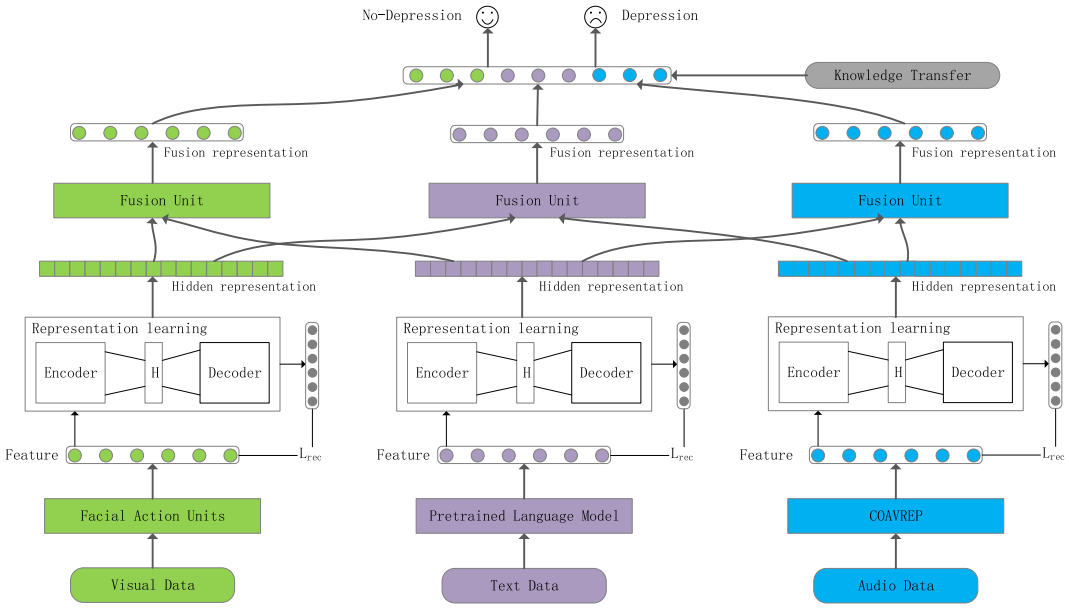


Fig. 2. Architecture of the proposed framework for enhancing multimodal depression diagnosis, encompassing data preprocessing, modality representation learning, inter-modal fusion unit, depression-related knowledge transfer, and depression diagnosis.

data here encompasses visual, textual, and speech modalities. In the data feature preprocessing stage, facial action units system is employed to extract facial features from patients, using facial movement variations as one of the indicators for depression symptom assessment. Pretrained language models are utilized to transform textual string data into computable feature vectors, while COAVREP is employed to extract audio features from speech data. Within the modality feature representation learning component, an Encoder-Decoder structured autoencoder is employed to preserve information integrity while extracting depression-related features. The inter-modal fusion unit utilizes a multi-head attention mechanism to establish interactions among modalities, align modality semantics, harness the complementarity between modalities, and enhance the discriminative ability of depression-related features. For depression-related knowledge transfer, a multi-modal emotion recognition pretrained model and knowledge from multi-modal emotion corpora are leveraged to enrich the model's knowledge background. In the depression diagnosis, the multi-modal features are input into a classifier to determine whether the current patient is suffering from depression. Subsequent sections will provide detailed explanations of the key techniques and components of the proposed framework.

3.2. Data preprocessing

3.2.1. Visual modality

In the visual modality, our approach involves utilizing the Facial Action Coding System [12] for extracting facial features from patients' images. This systematic framework allows us to effectively capture even the most nuanced facial movement variations, which are known to reflect emotional states, including those associated with depression. The Facial Action Coding System employs a standardized set of Facial Action Units (FAUs), which correspond to specific facial muscle movements. These distinct movements, often referred to as FAUs are closely tied to various facial expressions and emotions. Notably, in the realm of depression diagnosis, FAUs assume a pivotal role in assessing emotional states and identifying potential depressive symptoms. By meticulously analyzing these FAUs, we gain insights into the intricacies of emotional expression, thereby enhancing our capacity to discern and evaluate signs of depression in individuals.

3.2.2. Text modality

Converting textual string data into quantifiable features stands as a pivotal endeavor for seamless integration within our framework's architecture. To realize this goal, we harness the capabilities of pretrained language models. BERT (Bidirectional Encoder Representations from Transformers) [14], a notable exemplar, is tailored to serve as a pre-trained language model, with a dedicated focus on acquiring intricate text representations. The utilization of pre-trained language models is of paramount importance, as it effectively captures the nuanced intricacies and emotional nuances inherent in textual content. The model adeptly transforms textual data into feature vectors, which seamlessly align with the requirements of subsequent stages in our framework. By imbuing text with its semantic essence, we amplify the framework's acumen in apprehending linguistic cues intricately tied to depression symptoms.

3.2.3. Audio modality

In the context of the audio modality, our framework utilizes the COAVREP [13] technique for preprocessing audio data. COAVREP, well-regarded for its efficiency in providing swift access to new speech processing algorithms, distills meaningful audio features. This process transforms raw audio signals into concise yet informative representations. This method ensures that crucial acoustic cues are encapsulated within feature vectors, seamlessly aligning with subsequent stages of our framework. This approach reaffirms our dedication to proficiently handling audio data, enhancing our framework's ability to capture vital auditory cues. It also contributes to our comprehensive multimodal strategy, aimed at advancing depression diagnosis through a holistic approach.

3.3. Multimodal feature fusion unit

In order to effectively harness the synergistic potential of diverse modalities and capture their complementary information, we introduce a multimodal feature fusion unit. This unit is specifically designed to seamlessly integrate visual, audio, and textual features using a multi-head attention mechanism, resulting in a cohesive and enriched multimodal fusion representation. The core concept of our multimodal feature fusion unit revolves around the utilization of a multi-head attention mechanism. This mechanism enables the model to dynamically allocate attention weights to different modalities, emphasizing relevant features while diminishing the impact of less informative ones.

Given the visual feature X^v, audio feature X^a, and textual feature X^t, we project each modality's feature into query Q, key K, and value V spaces using learnable projection matrices W_Q^m, W_K^m, W_V^m for modality $m \in \{v,a,t\}$. As shown in the following equations (1)–(3).

$$Q^m = X^m W_Q^m \tag{1}$$

$$K^m = X^m W_K^m \tag{2}$$

$$V^m = X^m W_V^m \tag{3}$$

For each pair of modalities, m and n, we calculate the modality interaction attention scores A^{mn} in equation (4) by performing dot-product attention between the projected queries and keys. Where d_k is the dimensionality scaling factor.

$$A^{mn} = \text{softmax}\left(\frac{Q^m (K^n)^T}{\sqrt{d_k}}\right), m, n \in \{v, a, t\} \tag{4}$$

We compute the weighted sum of values V^n using the attention scores A^{mn} to generate modality-specific context fusion vectors C^{mn}. The formula is denoted as equation (5).

$$C^{mn} = A^{mn} V^n \tag{5}$$

The contextual fusion vectors are concatenated and then linearly transformed using a weight matrix to generate the multimodal fusion representation X^{fused}, where W_o is the weight matrix. The formula is denoted as equation (6).

$$X^{fused} = [C^{va}, C^{vt}, C^{av}, C^{at}, C^{tv}, C^{ta}] \cdot W_o \tag{6}$$

Through this unit, we effectively harness the cross-modal interactions and capture the distinctive cues from each modality. This enriched multimodal fusion representation serves as a comprehensive input for downstream depression diagnosis tasks, leveraging the synergies between different modalities for improved accuracy and robustness.

3.4. Representation learning strategy

In the realm of depression diagnosis, leveraging a combination of pre-processing techniques and feature fusion strategies can provide valuable insights. However, to further enhance the diagnostic accuracy, we introduce a pivotal step before feature fusion, which involves the application of representation learning. Representation learning refers to the process of automatically learning a more informative and compact representation of raw data. In our context, representation learning is employed to enhance the quality and expressiveness of the extracted features before their fusion. Representation learning offers multifaceted benefits by extracting higher-level features that capture underlying data characteristics, potentially revealing subtle depression-related markers, and simultaneously reducing data dimensionality for improved computational efficiency and mitigated overfitting risk.

We employ autoencoders for modality feature representation learning, and the model architecture of the autoencoder is illustrated in Fig. 3. The model incorporates BiLSTM layers within both the encoder and decoder components. The objective is to effectively encode temporal sequences of data, while utilizing pooling and up-sampling layers for dimensionality reduction and enhancement. This approach ensures that the essential information is retained during encoding and decoding processes. The number of layers in the encoder and decoder is comprehensively analyzed in the experimental section. Here, for the purpose of illustration, we present the usage of one layer of BiLSTM and one layer of pooling in the encoder, along with one layer of up-sampling and one layer of BiLSTM in the decoder.

The encoder is designed to capture relevant features from the input data. In our case, the input consists of temporal sequences, making BiLSTM layers a suitable choice due to their ability to capture bidirectional dependencies within the sequences. The encoded representation is obtained from the final hidden states of the BiLSTM layers. The pooling layers are introduced to down-sample the encoded representation, reducing its dimensionality while preserving the most salient information. The input sequence is represented as a sequence of vectors, denoted as $X = [x_1, x_2, \cdots, x_T,]$, where T is the sequence length. The encoder consists of BiLSTM layers followed by pooling layers, which are denoted in equations (7)–(9). The intermediate encoded result, denoted as h_t^{l+1}, serves as latent hidden features for subsequent depression diagnosis.

$$\overrightarrow{h}_t^l = \text{LSTM}\left(\overrightarrow{h}_{t-1}^l, x_t\right) \tag{7}$$

$$\overleftarrow{h}_t^l = \text{LSTM}\left(\overleftarrow{h}_{t+1}^l, x_t\right) \tag{8}$$

$$h_t^{l+1} = Pooling\left(\overrightarrow{h}_t^l, \overleftarrow{h}_t^l\right) \tag{9}$$

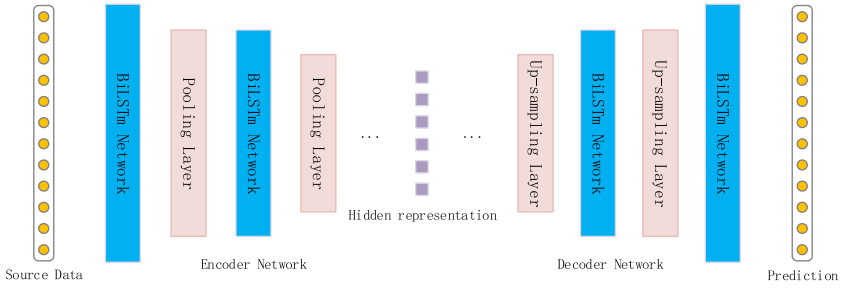


Fig. 3. Representation Learning Model Structure, comprising encoder and decoder components. The encoder incorporates BiLSTM and pooling layers, while the decoder includes BiLSTM and up-sampling layers.

The decoder takes the down-sampled encoded representation and aims to reconstruct the original input data. To achieve this, we utilize up-sampling layers to increase the dimensionality of the representation, which is implemented using bilinear interpolation. The up-sampled representation is then processed through another set of BiLSTM layers. These layers are responsible for capturing the temporal dependencies necessary for accurate reconstruction. Finally, the decoder's output is compared to the input data to quantify the quality of the reconstruction. The up-sampled representation from the encoder is denoted as $U = [u_1, u_2, \cdots, u_T]$, where T is the up-sampled sequence length. The decoder consists of up-sampling layers followed by BiLSTM layers, which are denoted in equation (10)–(12).

$$u_t^{l+2} = \text{Upsampling}\left(h_t^{l+1}\right) \tag{10}$$

$$\overrightarrow{d}_t^{\,l+3} = \text{LSTM}\left(\overrightarrow{d}_{t-1}^{\,l+3}, u_t^{l+2}\right) \tag{11}$$

$$\overleftarrow{d}_t^{\,l+3} = \text{LSTM}\left(\overleftarrow{d}_{t+1}^{\,l+3}, u_t^{l+2}\right) \tag{12}$$

The decoder's output is denoted as r_t, and in order to achieve the goals of information encoding and reconstruction, we employ the mean squared error (MSE) to compute the loss. This loss function quantifies the difference between the reconstructed output r_t and the original input x_t, encouraging the model to minimize this difference during training. The calculation formula is denoted in equations (13) and (14) as shown below.

$$r_t = \left[\overleftarrow{d}_t^{\,l+3}, \overrightarrow{d}_t^{\,l+3}\right] \tag{13}$$

$$L_{rec} = \frac{1}{T}\sum_{t=1}^{T}(x_t - r_t)^2 \tag{14}$$

3.5. Knowledge transfer approach

The integration of knowledge transfer significantly enhances the effectiveness of depression diagnosis. Leveraging external knowledge from existing sources empowers us to enhance diagnostic accuracy by incorporating complementary information. Specifically, we employ two methods of knowledge transfer: fine-tuning pre-trained language models and transfer learning from a multi-modal emotion recognition dataset, both of which bolster the diagnostic capabilities of our framework.

Fine-tuning pre-trained language models is one approach to knowledge transfer. By adapting a pre-trained language model to our depression diagnosis task, we enable it to capture domain-specific linguistic nuances and emotional cues linked to depression. This process involves initializing a pre-trained language model, such as SentiLARE [43], with learned linguistic representations, and then training it on our depression-specific text data. This leverages the model's ability to comprehend intricate textual patterns, thereby enhancing the framework's text-based features and contributing to a more comprehensive diagnostic process.

Another dimension of knowledge transfer involves transfer learning from a multi-modal emotion recognition dataset. In practice, we commence the transfer learning process by training our framework on datasets like CMU-MOSI [41] and CMU-MOSEI [42], which contain diverse multi-modal emotion recognition data. This initial training enriches our framework's grasp of cross-modal relationships and emotional expressions. Following this, we fine-tune our framework with depression-specific data, integrating the insights gained from the multi-modal emotion recognition datasets into our depression diagnosis model.

Both knowledge transfer methods can be used either independently or in combination. Employing a pre-trained language model independently enhances the capability for textual feature representation, while the exclusive use of a multimodal emotion recognition dataset allows for the full utilization of external emotional knowledge related to depression symptoms. When combined, the approach starts with fine-tuning the pre-trained language model, followed by training on the multimodal emotion recognition dataset, and finally, the model is trained on the depression auxiliary diagnostic dataset. This sequence maximizes the benefits of both knowledge transfer methods.

3.6. Depression diagnosis network

The integration of representation learning and knowledge transfer within our framework results in a fortified depression diagnosis network. Through the utilization of representation learning, we are able to extract more nuanced and informative features from the modalities, effectively delineating the underlying patterns associated with depression. Moreover, employing knowledge transfer enhances the network's comprehension of the data by leveraging existing knowledge, thereby improving its capacity for generalization and diagnostic precision.

We employ the multi-modal fusion outputs for depression diagnosis, which are processed through a fully connected neural network. Following this, a softmax classifier is employed to discern whether a patient exhibits depression. For the loss function, we adopt the cross-entropy loss to quantify the difference between the predicted depression probabilities and the actual labels. The formula for the cross-entropy loss with batch size N, labels y_i, and predicted probabilities p_i is denoted in equation (15) as follows.

$$L_{cls} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i) + (1 - y_i)\log(1 - p_i) \tag{15}$$

During the model training process, the employed objective function includes the reconstruction loss from the representation learning strategy and the accuracy loss of the diagnosis recognition results. The calculation formula is denoted in equation (16) as follows, where λ is the weight coefficient of representation learning in the loss function.

$$L_{total} = L_{cls} + \lambda L_{rec} \tag{16}$$

4. Experiments

In this section, we present an extensive analysis of our proposed framework's performance through a series of experiments conducted on depression diagnosis. We begin by providing a description of the dataset used for both training and evaluation purposes in section 4.1. Subsequently, in section 4.2, we outline the implementation details, covering model architectures, hyperparameters, and training protocols. To assess the effectiveness of our approach, we introduce the evaluation indicators employed in section 4.3. Additionally, we evaluate our method's performance against established techniques in the field in section 4.4. For a deeper understanding of the components of our model, we conduct an ablation study in section 4.5, aiming to dissect the influence of distinct factors. We also visually explore the hidden representations acquired by our model in section 4.6, enhancing our insights. Finally, we assess the robustness of our approach to noise in section 4.7, thereby validating its applicability in real-world scenarios.

4.1. Dataset

The central dataset used in this study to evaluate depression is the DAIC-WOZ depression dataset [44], which played a prominent role in the AVEC2017 challenge. This dataset involves engagements between participants and a virtual animated interviewer named Ellie. It encompasses a total of 189 subjects, designated by identification numbers ranging from 300 to 492. However, certain participants—numbered 342, 394, 398, and 460—were excluded from the dataset due to incomplete information caused by technical constraints.

At the core of the dataset are approximately 50 h of audio-video recordings paired with corresponding audio transcriptions, capturing the interactive interviews held between the virtual interviewer Ellie and the participating subjects. Each recording is associated with both a PHQ-8 score and a PHQ-8 binary label. The PHQ-8 score delineates the severity level of depression for each participant, while the PHQ-8 binary indicates whether the participant is experiencing depression or not. The DAIC-WOZ dataset is split into 80% training, 10% validation, and 10% test sets.

4.2. Implementation details

In this section, we delve into the precise implementation intricacies of our proposed framework for depression diagnosis. We offer an in-depth understanding of the architectures, hyperparameters, and training protocols of the depression diagnosis framework that were utilized in our experimental setup.

The representation learning strategy within our framework is implemented through an encoder-decoder network, which is designed to acquire latent representations that encapsulate crucial features of the data, thereby enhancing feature extraction. The network's architecture consists of a series of layers, each contributing to the transformation of the input data. The specific parameters of the network structure are detailed in Table 1. Notably, the inclusion of a Dense layer with a tanh activation function serves to

Table 1
Detailed parameters of the representation learning network architecture.

Layers	Hyperparameter settings	Output Shape
Input Layer	–	(500, 74)
BiLSTM	Hidden Units: 128, Activation: tanh	(500, 128)
Pooling	Pool size: 2	(250, 128)
Dropout	Rate:0.25	(250, 128)
BiLSTM	Hidden Units: 256, Activation: tanh	(250, 256)
Pooling	Pool size: 2	(125, 256)
Dropout	Rate:0.25	(125, 256)
Flatten	–	32000
Dense	Activation: tanh	100
Dense	Activation: tanh	32000
Reshape	–	(125, 256)
Upsampling	Size: 2	(250, 256)
Dropout	Rate:0.25	(250, 256)
BiLSTM	Hidden Units: 128, Activation: tanh	(250, 128)
Upsampling	Size: 2	(500, 128)
Dropout	Rate:0.25	(500, 128)
BiLSTM	Hidden Units: 74, Activation: tanh	(500, 74)

transform the representation into a higher-level abstraction of 100 dimensions, which subsequently contributes to the depression diagnosis process.

Hyperparameters are crucial for the optimal performance of our framework. The learning rate is set to 0.001, and we utilize the Adam optimizer to update the model's weights during training. The batch size is chosen as 32 to balance computation efficiency and gradient stability. To mitigate overfitting, we apply dropout with a rate of 0.25 after each layer in the neural network. All experiments are performed on a machine equipped with an NVIDIA GPU. We use PyTorch as the deep learning framework for implementing the model. The dataset is preprocessed using standard libraries in Python. In the experiment, 10-fold cross-validation is employed to assess the model's effectiveness, which helps in evaluating the model's generalization ability on new data.

4.3. Evaluation indicators

We provide a comprehensive overview of the evaluation indicators employed to thoroughly evaluate the performance of our depression diagnosis framework. The selection of evaluation metrics is pivotal to conducting a rigorous assessment of the model's capabilities in accurately identifying depression cases. We employ established metrics, including precision, recall, the F1 score, and accuracy, which collectively offer insights into the accuracy, sensitivity, and overall effectiveness of the model in diagnosing depression.

Precision is a measure of the accuracy of the positive predictions made by the model. It calculates the proportion of correctly predicted positive instances among all instances predicted as positive. The formula for precision is denoted in equation (17) as follows. Where TP represents true positives and FP represents false positives.

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

Recall measures the model's ability to correctly identify all positive instances from the actual positive instances. It calculates the proportion of true positives among all actual positive instances. The formula for recall is denoted in equation (18) as follows. Where TP represents true positives and FN represents false negatives.

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

The F1 score is the harmonic mean of precision and recall. It provides a balanced measure that takes into account both false positives and false negatives. The formula for the F1 score is denoted in equation (19) as follows.

$$F1 - Score = \frac{2 \times precision \times recall}{precision + recall} \tag{19}$$

The accuracy metric measures the proportion of correctly predicted instances among the total instances evaluated. It provides an overall assessment of the model's performance. The formula for accuracy is denoted in equation (20) as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

4.4. Comparison with other methods

To validate the effectiveness of the proposed framework, we conducted comparative experiments on the DAIC-WOZ dataset, and the experimental results are presented in Table 2. We conducted experiments under equivalent conditions using the same preprocessed data to ensure a fair comparison of experimental outcomes. The proposed RLKT-MDD model achieves the following metrics: Precision of 80.20%, Recall of 81.00%, F1-Score of 80.60%, and Accuracy of 81.10%. To begin with the conclusion, it can be inferred that the model proposed in this paper outperforms all other models. The comparison involves two categories of models: traditional machine learning and deep learning methods. Within the realm of traditional machine learning, including Logistic Regression, Naive Bayes, Random Forest, and Support Vector Machine, these models exhibit poor performance. This indicates the challenge posed by the intricate nature of acoustic and visual features in depression-related data, particularly considering the approach of flattening and

Tabel 2
Comparison of proposed framework and Baseline methods.

Methods	Model Type	Precision	Recall	F1-Score	Accuracy
Logistic Regression	Logistic Regression	53.20	55.80	54.47	54.12
Naive Bayes	Naive Bayes	56.70	58.40	57.54	57.02
Random Forest	Random Forest	54.80	55.20	55.00	56.33
M. Valstar et al. [45]	SVM	57.90	59.60	58.74	58.41
S. A. Qureshi et al. [46]	Multitask learning	58.00	61.00	59.46	61.11
Z. Huang et al. [47]	FVTC-CNN	72.30	75.70	73.96	74.06
G. Lam et al. [48]	Transformer-CNN	75.10	76.50	75.79	76.25
Z. Han et al. [49]	Spatial-Temporal	79.00	80.50	79.74	80.08
RLKT-MDD (proposed)	Encoder-Decoder	80.20	81.00	80.60	81.10

concatenating such features.

In the sphere of deep learning, S. A. Qureshi et al. [46] employed LSTM neural networks for feature extraction, coupled with multitask training of model parameters. Z. Huang et al. [47] employed convolutional neural networks for vocal channel feature extraction. G. Lam et al. [48]utilized a combination of convolutional neural networks and Transformer structures for data feature extraction. Z. Han et al. [49] introduced a temporal-spatial feature extraction network to learn contextually distant dependencies, effectively enhancing the model's feature extraction capacity and detection efficacy. From the comparative methods, it is evident that effective feature representation is pivotal in enhancing the detection efficacy of the model. The framework proposed in this paper leverages the Encoder-Decoder outcome comprehensively. This allows for the preservation of modality information while representing depression-related features. Simultaneously, the model benefits from knowledge transfer, endowing it with a robust knowledge background. The representation learning and knowledge transfer stand as a critical factor in the model's further efficacy enhancement.

4.5. Ablation experiment

We embark on ablation experiments in this section, aiming to systematically dissect the components and strategies that contribute to the performance of our proposed framework. Through a sequence of controlled experiments, our objective is to discern the impact of varying modality combinations, assess the effectiveness of our representation learning strategy, evaluate the efficacy of the knowledge transfer approach, and analyze the influence of the parameter λ within the loss function. The insights garnered from these experiments offer a comprehensive grasp of the mechanisms that underlie the success of our framework and illuminate the intricate interplay among its diverse components.

4.5.1. Result of different modality combination

We analyze the impact of different modality combinations on the effectiveness of depression diagnosis. The experimental results are presented in Table 3. The results of the single modality experiments demonstrate varying levels of performance among visual, audio, and text modalities. The visual modality achieves a Precision of 62.35%, Recall of 62.80%, F1-Score of 62.57%, and an Accuracy of 62.50%. The audio modality shows slightly better results with a Precision of 68.41%, Recall of 69.20%, F1-Score of 68.80%, and an Accuracy of 68.75%. Notably, the text modality outperforms both visual and audio modalities, exhibiting a Precision of 74.39%, Recall of 75.60%, F1-Score of 74.99%, and an Accuracy of 74.75%. These results suggest that the textual information carries more discriminative features for depression diagnosis compared to the visual and audio modalities.

In the bimodal combinations, the performance shows interesting trends. When combining the visual and audio modalities, the Precision is 70.25%, Recall is 70.80%, F1-Score is 70.52%, and Accuracy is 70.50%. This combination exhibits an improvement compared to the individual modalities, indicating that the fusion of visual and audio features enhances depression diagnosis. Similarly, the combination of visual and text modalities yields higher results than any of the individual modalities. This suggests that the visual and textual features complement each other, resulting in improved depression diagnosis. On the other hand, the combination of audio and text modalities also produces better results, with a Precision of 76.34%, Recall of 76.90%, F1-Score of 76.62%, and Accuracy of 76.60%. These findings underscore the advantages of fusing different modalities to enhance depression diagnosis, utilizing the unique features from each modality.

The most comprehensive analysis arises from the trimodal fusion, combining visual, audio, and text modalities. This trimodal approach achieves Precision of 80.20%, Recall of 81.00%, F1-Score of 80.60%, and Accuracy of 81.10%. The significant improvement in all metrics compared to the individual modalities and bimodal combinations underscores the effectiveness of cross-modal fusion. The integration of information from visual, audio, and text sources evidently boosts depression diagnosis performance. In summary, the experimental results underline the advantages of multi-modal depression diagnosis.

4.5.2. Effectiveness of the representation learning strategy

We delve into the outcomes of our experiments aimed at dissecting the influence of various encoder and decoder structures within the context of our representation learning strategy. Our objective was to assess how different architectural configurations impact the overall performance of our framework. The experimental results are summarized in Table 4. To ensure consistency and facilitate fair comparison, we employed neural networks with an equal number of layers in both the encoder and decoder. It is important to note that the results table provided specifically highlights the number of layers in the encoder component.

Our findings underline the transformative role of representation learning in the effectiveness of our framework. The encoder, equipped with the ability to autonomously extract meaningful features, acts as the bedrock for constructing a comprehensive

Tabel 3
Performance analysis of different modality combinations for depression diagnosis.

Modality	Precision	Recall	F1-Score	Accuracy
Visual	62.35	62.80	62.57	62.50
Audio	68.41	69.20	68.80	68.75
Text	74.39	75.60	74.99	74.75
Visual + Audio	70.25	70.80	70.52	70.50
Visual + Text	78.21	78.80	78.50	78.50
Audio + Text	76.34	76.90	76.62	76.60
Visual + Audio + Text	80.20	81.00	80.60	81.10

Tabel 4
Impact of representation learning strategy on multilayer BiLSTM and pooling for depression diagnosis.

Encoder Layers	Precision	Recall	F1-Score	Accuracy
BiLSTM*1+Pooling*1	79.50	78.20	78.85	79.70
BiLSTM*2+Pooling*2	80.20	81.00	80.60	81.10
BiLSTM*3+Pooling*3	80.40	79.90	80.15	80.50
BiLSTM*4+Pooling*4	79.80	78.50	79.15	79.30
BiLSTM*5+Pooling*5	79.90	80.10	80.00	79.60

understanding of the underlying data. The configuration featuring two BiLSTM layers coupled with two pooling layers exhibits superior performance. With Precision standing at 80.20%, Recall at 81.00%, F1-Score at 80.60%, and Accuracy peaking at 81.10%, this arrangement signifies the optimized harmony between depth and feature extraction capabilities. However, we observed that an insufficient number of layers impairs the model's capacity to discern intricate patterns, subsequently resulting in decreased performance. Conversely, an excessive number of layers contributes to overfitting, thereby diminishing the model's ability to generalize and consequently yielding suboptimal outcomes.

4.5.3. Effectiveness of the knowledge transfer approach

To validate the significance of the knowledge transfer approach in enhancing the efficiency of our depression diagnosis framework, we conducted experiments, the results of which are presented in Table 5. BERT and SentiLARE, as pre-trained language models, underwent fine-tuning during the training process in the text modality to enrich their textual knowledge background. CMU-MOSI and CMU-MOSEI, two multi-modal emotion recognition datasets, were leveraged to enhance the model's capacity for emotional awareness and understanding through pre-training on these datasets. From the experimental outcomes, it is evident that using SentiLARE and CMU-MOSEI simultaneously achieved the most favorable results, attaining a Precision of 80.20%, Recall of 81.00%, F1-Score of 80.60%, and an impressive Accuracy of 81.10%. These results underscore the effectiveness of the knowledge transfer approach. Particularly notable is the exemplary performance of the CMU-MOSEI dataset, emphasizing the alignment between cross-modal sentiment recognition and depression diagnosis within our framework.

In summary, the knowledge transfer approach emerges as a pivotal component of our framework's success. While the fine-tuning of pre-trained language models enhances the understanding of depression-related textual cues, transferring learning from multi-modal emotion recognition datasets enriches our model with a nuanced understanding of emotional expressions. The combined effect of these strategies facilitates a more comprehensive and accurate depression diagnosis, as evidenced by the results. This multifaceted approach not only strengthens the framework's performance but also opens avenues for future research in multi-modal diagnostic applications.

4.5.4. Effect of parameter λ in the loss function

In order to investigate the impact of the parameter λ within the loss function, which serves as a critical determinant in maintaining equilibrium between representation learning and depression diagnosis outcomes, we conducted experiments, the results of which are presented in Table 6. The parameter λ embedded within the loss function plays a pivotal role in shaping the relationship between the acquired representations and the diagnostic accuracy. Our exploration revealed that the most favorable results were attained when λ equaled 0.10. This specific value strikes an optimal balance, orchestrating a synergy between the acquired representations and the diagnostic accuracy, thereby leading to the framework's peak performance.

Upon closer examination of the results, it becomes apparent that exceedingly small values of λ (such as 0.01 and 0.05) result in performance that is marginally enhanced by the representation learning process. This phenomenon can be attributed to the overly dominant role of the diagnostic task in the loss function, thus downplaying the contribution of representation learning. Conversely, when λ assumes higher values (such as 0.20, 0.30, 0.40, and 0.50), the representation learning component becomes overly prominent within the loss function. This imbalance may hinder the framework's capacity to optimize for accurate depression diagnosis, potentially leading to reduced performance due to skewed priorities. This analysis showcases the necessity of a balanced integration of representation learning and diagnostic accuracy for achieving optimal outcomes in depression diagnosis.

Tabel 5
Effectiveness of knowledge transfer approach in enhancing depression diagnosis performance.

Knowledge Type	Precision	Recall	F1-Score	Accuracy
BERT	78.83	78.50	78.66	79.00
SentiLARE	79.64	80.00	79.82	79.50
CMU-MOSI	79.86	79.88	79.87	79.65
CMU-MOSEI	79.92	79.95	79.93	79.77
BERT + CMU-MOSI	80.01	80.05	80.03	80.03
BERT + CMU-MOSEI	80.08	80.05	80.06	80.13
SentiLARE + CMU-MOSI	80.10	80.11	80.10	80.11
SentiLARE + CMU-MOSEI	80.20	81.00	80.60	81.10

Tabel 6
The results of different parameter λ in the loss function for depression diagnosis.

Parameter value	Precision	Recall	F1-Score	Accuracy
parameter $\lambda = 0.01$	79.95	80.10	80.02	79.91
parameter $\lambda = 0.05$	79.64	79.80	79.72	79.63
parameter $\lambda = 0.10$	80.20	81.00	80.60	81.10
parameter $\lambda = 0.20$	80.63	80.80	80.71	80.51
parameter $\lambda = 0.30$	79.96	79.70	79.83	79.20
parameter $\lambda = 0.40$	80.05	80.20	80.12	80.11
parameter $\lambda = 0.50$	79.72	79.50	79.61	79.43

4.6. Visualization of the hidden representation

In this section, we present visualizations of the distributions of multimodal representations within the embedding space, aiming to assess the efficacy of the proposed RLKT-MDD method. We conduct a visualization experiment by employing the t-SNE algorithm to project each high-dimensional representation into a two-dimensional feature space. This approach effectively captures the local structure inherent in the high-dimensional features. The outcomes are depicted in Fig. 4, illustrating the visualizations of visual, audio, text, and fusion feature distribution both before and after the representation learning phase. We can observe that, for these modal representations, after representation learning, a more discriminative representation space is achieved, especially in the case of multimodal representations. Visual representations became more informative with discernible features, voice representations revealed consistent and interpretable patterns, and text representations organized into semantically meaningful structures. Moreover, the fusion of these modalities enhanced the overall representation quality, enabling improved cross-modal understanding. By effectively mapping high-dimensional data into a shared feature space, representation learning facilitates meaningful cross-modal associations, promoting a deeper understanding of multimodal data. However, it is worth noting that even after representation learning, certain samples remain challenging to distinguish in the feature space. This poses a substantial challenge in the context of multimodal depression diagnosis. Future efforts should focus on finding more discriminative representations to enhance the diagnostic accuracy and address these remaining challenges in the field.

5. Conclusion

In conclusion, our study introduces an innovative framework that utilizes representation learning and knowledge transfer to enhance multimodal depression diagnosis. We demonstrated the transformative impact of representation learning, wherein the acquisition of more discerning depression-related features through autoencoders facilitates precise diagnosis. The effectiveness of knowledge transfer from pre-trained language models and multi-modal sentiment recognition datasets was evident, leading to improved diagnostic performance. Our findings underscore the intricate interplay between representation learning and diagnostic accuracy, with optimal parameter configurations and strategic modality combinations yielding superior results. This research advances the comprehension of AI-driven depression diagnosis, opening avenues for more effective and personalized interventions in mental health.

While the framework proposed in this paper has achieved performance improvements, it also has its inherent limitations. When employing a pre-trained language model and a multimodal emotion recognition dataset for external knowledge transfer, the model's generalizability may be affected due to the scale and domain constraints of the pre-trained data. To enhance the model's ability to generalize, it is advisable to integrate domain-specific data into the pre-training process, thereby better adapting the model to specific application scenarios. Additionally, instead of relying solely on a single data source, it is beneficial to combine multiple pre-trained data sources from various domains. This approach will enhance the model's capacity to understand and process different types of data more effectively. These will be the directions of effort in future research.

Ethical and informed consent for data used

This study adheres to ethical guidelines and obtained informed consent for the data used. All data sources and participants involved in the research provided their consent for the collection, analysis, and publication of the data. This declaration affirms our commitment to conducting research in an ethical manner and upholding the principles of informed consent.

Data availability and access

The data used in this study are sourced from a publicly available dataset. The dataset can be accessed at the following URL: https://dcapswoz.ict.usc.edu/wwwdaicwoz/. We acknowledge and appreciate the efforts of the dataset creators in making their data available to the research community. This study utilized the dataset in accordance with its stated permissions and guidelines. Researchers interested in accessing and utilizing the data for their own analyses can refer to the provided URL to obtain the necessary files and information.

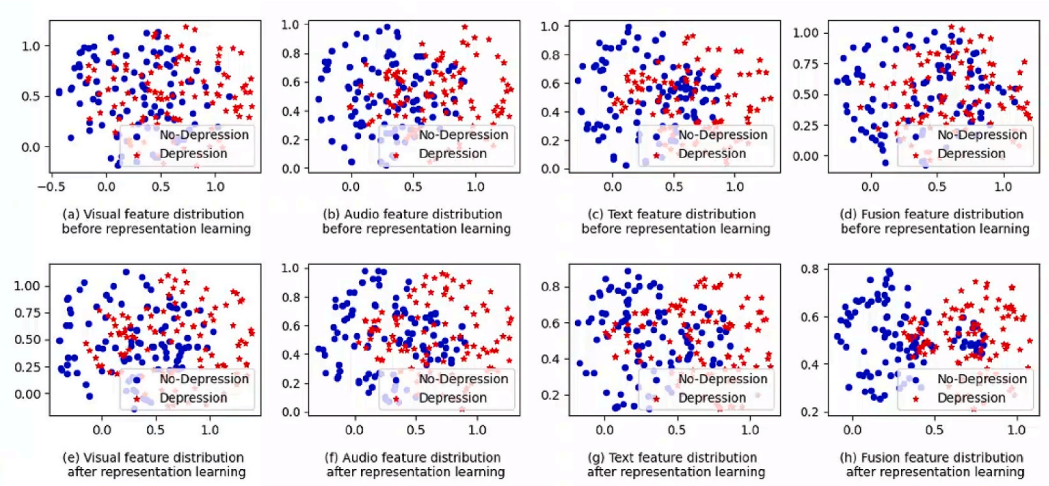


Fig. 4. Visualization contrast of unimodal latent representations and multimodal fusion representations before and after representation learning.

CRediT authorship contribution statement

Shanliang Yang: Writing – original draft, Methodology, Data curation, Conceptualization. **Lichao Cui:** Writing – review & editing, Data curation. **Lei Wang:** Funding acquisition. **Tao Wang:** Funding acquisition. **Jiebing You:** Writing – original draft, Methodology, Data curation, Conceptualization.

Declaration of competing interest

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

Acknowledgements

This work was supported by the Shandong Provincial Natural Science Foundation (No. ZR2021MF017), Key R&D Program of Shandong Province, China (No. 2023RKY01015) and Shandong Provincial Natural Science Foundation (No. ZR2020QF069).

References

[1] World Health Organization, "Depression," https://www.who.int/news-room/fact-sheets/detail/depression, Accessed: March 31, 2023.
[2] K. Kroenke, R.L. Spitzer, J.B.W. Williams, The PHQ-9, J. Gen. Intern. Med. 16 (2001) 606–613.
[3] J. Endicott, J. Cohen, J. Nee, et al., Hamilton depression rating scale: extracted from regular and change versions of the schedule for affective disorders and schizophrenia, Arch. Gen. Psychiatr. 38 (1) (1981) 98–103.
[4] E. Rejaibi, A. Komaty, F. Meriaudeau, et al., MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech, Biomed. Signal Process Control 71 (2022) 103107.
[5] S. Sardari, B. Nakisa, M.N. Rastgoo, et al., Audio based depression detection using convolutional autoencoder, Expert Syst. Appl. 189 (2022) 116076.
[6] M. Fang, S. Peng, Y. Liang, C.-C. Hung, S. Liu, A multimodal fusion model with multi-level attention mechanism for depression detection, Biomed. Signal Process Control 82 (2023) 104561.
[7] W. Yang, J. Liu, P. Cao, R. Zhu, Y. Wang, J.K. Liu, F. Wang, X. Zhang, Attention guided learnable time-domain filterbanks for speech depression detection, Neural Network. 165 (2023) 135–149.
[8] W.C. de Melo, E. Granger, A. Hadid, A deep multiscale spatiotemporal network for assessing depression from facial dynamics, IEEE Transactions on Affective Computing 13 (2022) 1581–1592.
[9] M. Dorkenwald, F. Xiao, B. Brattoli, J. Tighe, D. Modolo, SCVRL: shuffled contrastive video representation learning, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW), 2022, pp. 4131–4140.
[10] Q. Zheng, J. Zhu, Z. Li, Z. Tian, C. Li, Comprehensive multi-view representation learning, Inf. Fusion 89 (2023) 198–209.
[11] J. Yu, G. Liu, Knowledge transfer-based sparse deep belief network, IEEE Trans. Cybern. (2022) 1–12.
[12] P. Ekman, E. Rosenberg, What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS), vol. 68, Oxford University Press, 2005, pp. 83–96.
[13] G. Degottex, J. Kane, T. Drugman, T. Raitio, S. Scherer, COVAREP — a collaborative voice analysis repository for speech technologies, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP), 2014, pp. 960–964.
[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional Transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186.
[15] C. Otte, S.M. Gold, B.W. Penninx, C.M. Pariante, A. Etkin, M. Fava, D.C. Mohr, A.F. Schatzberg, Major depressive disorder, Nat. Rev. Dis. Prim. 2 (2016) 16065.
[16] J.R. Vittengl, R.B. Jarrett, E. Ro, L.A. Clark, How can the DSM-5 alternative model of personality disorders advance understanding of depression? J. Affect. Disord. 320 (2023) 254–262.
[17] N. Cheffi, O. Chakroun-Walha, R. Sellami, R. Ouali, D. Mnif, F. Guermazi, F. Issaoui, M. Lajmi, B. Benamar, J. Damak, N. Rekik, J. Masmoudi, Validation of the Hamilton depression rating scale (HDRS) in the Tunisian dialect, Publ. Health 202 (2022) 100–105.

[18] A.S. Zigmond, R.P. Snaith, The hospital anxiety and depression scale, Acta Psychiatr. Scand. 67 (1983) 361–370.
[19] K.P. Shanthosam, P.K. Kasirajan, EliteVec: feature fusion for depression diagnosis using optimized long short-term memory network, Intelligent Automation & Soft Computing 36 (2023) 1745–1766.
[20] Y. Guo, C. Zhu, S. Hao, R. Hong, Automatic depression detection via learning and fusing features from visual cues, IEEE Transactions on Computational Social Systems (2022) 1–8.
[21] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, A. Othmani, MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech, Biomed. Signal Process Control 71 (2022) 103107.
[22] J. Zhu, S. Wei, X. Xie, C. Yang, Y. Li, X. Li, B. Hu, Content-based multiple evidence fusion on EEG and eye movements for mild depression recognition, Comput. Methods Progr. Biomed. 226 (2022) 107100.
[23] S.A. Qureshi, G. Dias, M. Hasanuzzaman, S. Saha, Improving depression level estimation by concurrently learning emotion intensity, IEEE Comput. Intell. Mag. 15 (2020) 47–59.
[24] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, M. Breakspear, Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors, IEEE Transactions on Affective Computing 9 (2018) 478–490.
[25] M. Muzammel, H. Salam, A. Othmani, End-to-end multimodal clinical depression recognition using deep neural networks: a comparative analysis, Comput. Methods Progr. Biomed. 211 (2021) 106433.
[26] W. Xie, C. Wang, Z. Lin, X. Luo, W. Chen, M. Xu, L. Liang, X. Liu, Y. Wang, H. Luo, M. Cheng, Multimodal fusion diagnosis of depression and anxiety based on CNN-LSTM model, Comput. Med. Imag. Graph. 102 (2022) 102128.
[27] T. Chen, R. Hong, Y. Guo, S. Hao, B. Hu, MS 2 -GNN: Exploring GNN-Based Multimodal Fusion Network for Depression Detection, IEEE Transactions on Cybernetics, 2022.
[28] N. Seneviratne, C. Espy-Wilson, Multimodal depression classification using articulatory coordination features and hierarchical attention based text embeddings, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP), 2022, pp. 6252–6256.
[29] Y. Wang, Z. Wang, C. Li, Y. Zhang, H. Wang, Online social network individual depression detection using a multitask heterogenous modality fusion approach, Inf. Sci. 609 (2022) 727–749.
[30] X. Jia, X.-Y. Jing, X. Zhu, S. Chen, B. Du, Z. Cai, Z. He, D. Yue, Semi-supervised multi-view deep discriminant representation learning, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2021) 2496–2509.
[31] S. Song, S. Jaiswal, L. Shen, M. Valstar, Spectral representation of behaviour primitives for depression analysis, IEEE Transactions on Affective Computing 13 (2022) 829–844.
[32] X. Zhou, Z. Wei, M. Xu, S. Qu, G. Guo, Facial depression recognition by deep joint label distribution and metric learning, IEEE Transactions on Affective Computing 13 (2022) 1605–1618.
[33] Z. Han, Y. Shang, Z. Shao, J. Liu, G. Guo, T. Liu, H. Ding, Q. Hu, Spatial-temporal feature network for speech-based depression recognition, IEEE Transactions on Cognitive and Developmental Systems (2023) 1, 1.
[34] M. Niu, J. Tao, B. Liu, J. Huang, Z. Lian, Multimodal spatiotemporal representation for automatic depression level detection, IEEE Transactions on Affective Computing 14 (2023) 294–307.
[35] Z. Guo, N. Ding, M. Zhai, Z. Zhang, Z. Li, Leveraging domain knowledge to improve depression detection on Chinese social media, IEEE Transactions on Computational Social Systems 10 (2023) 1528–1536.
[36] Y.S. Lin, L.K. Tai, A.L.P. Chen, The detection of mental health conditions by incorporating external knowledge, J. Intell. Inf. Syst. 61 (2023) 497–518.
[37] K. Yang, T. Zhang, S. Ananiadou, A mental state Knowledge–aware and Contrastive Network for early stress and depression detection on social media, Inf. Process. Manag. 59 (2022) 102961.
[38] W. Wu, C. Zhang, P.C. Woodland, Self-supervised representations in speech-based depression detection, in: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP), 2023, pp. 1–5.
[39] Y. Dong, X. Yang, A hierarchical depression detection model based on vocal and emotional cues, Neurocomputing 441 (2021) 279–290.
[40] W. Wu, M. Wu, K. Yu, Climate and weather: inspecting depression detection via emotion recognition, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP), 2022, pp. 6262–6266.
[41] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos, 2016. ArXiv. abs/1606.06259.
[42] A. Bagher Zadeh, P.P. Liang, S. Poria, E. Cambria, L.-P. Morency, Multimodal Language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2236–2246.
[43] P. Ke, H. Ji, S. Liu, X. Zhu, M. Huang, SentiLARE: sentiment-aware language representation learning with linguistic knowledge, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6975–6988.
[44] J. Gratch, R. Artstein, G.M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, et al., The distress analysis interview corpus of human and computer interviews, in: International Conference on Language Resources and Evaluation, 2014, pp. 3123–3128.
[45] M. Valstar, M. Pantic, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge, 2016.
[46] S.A. Qureshi, S. Saha, M. Hasanuzzaman, G. Dias, Multitask representation learning for multimodal estimation of depression level, IEEE Intell. Syst. 34 (2019) 45–52.
[47] Z. Huang, J. Epps, D. Joachim, Exploiting vocal tract coordination using dilated CNNS for depression detection in naturalistic environments, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP), 2020, pp. 6549–6553.
[48] G. Lam, H. Dongyan, W. Lin, Context-aware deep learning for multi-modal depression detection, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP), 2019, pp. 3946–3950.
[49] Z. Han, Y. Shang, Z. Shao, J. Liu, G. Guo, T. Liu, H. Ding, Q. Hu, Spatial-temporal feature network for speech-based depression recognition, IEEE Transactions on Cognitive and Developmental Systems (2023) 1, 1.