

IICL-DTA : A Contrast Framework for Automatic Depression Detection via Text and Audio

Dongfang Han^{a,b,c}, Guo-Xing Xiang^d, Yi Liang^{a,b,c}, Bo Kong^{a,b,c}, Yuanyuan Liao^{a,b,c}, Askar Hamdulla^{a,b,c}, Turdi Tohti^{a,b,c,*}

^a*School of Computer Science and Technology, Xinjiang University, Urumqi, 830017, China*

^b*Joint International Research Laboratory of Silk Road Multilingual Cognitive Computing, Urumqi, 830017, China*

^c*Xinjiang Key Laboratory of Signal Detection and Processing, Urumqi, 830017, China*

^d*Department of Psychology, Lingnan University, Hong Kong SAR, 999077, China*

Abstract

Automatic Depression Detection (ADD) methods, which can assist physicians in the early diagnosis and analysis of depression, have demonstrated promising performance. However, the current ADD methods face two primary challenges: (1) the ADD methods struggle to effectively capture the subtle differences between depressive and non-depressive samples. These minute variations in depressive cues make it difficult for models to distinguish between depressive and non-depressive samples. (2) Due to the privacy and sensitivity associated with depression, collecting data for ADD is challenging, resulting in small dataset sizes. This low-resource issue has become one of the main challenges for this task. In this paper, we propose an intra- and inter-contrastive learning framework (IICL-DTA) for supervised multimodal automatic depression detection. We employ BERT (Bidirectional Encoder Representations from Transformers) and SpeechFormer++ (a hierarchical efficient framework for paralinguistic speech processing) to extract textual and audio features, respectively. We then employ a projection layer to produce positive and negative feature pairs, which are incorporated into the intra- and inter-contrastive learning loss function. Furthermore, we design a multimodal auxiliary loss based on mixup features of text and audio to mitigate the low-resource data issue. This combined intra- and inter-contrastive learning approach encourages similar fused text–audio features to cluster while pushing dissimilar ones apart. We evaluate our method on three datasets: English DAIC-WOZ, E-DAIC, and Chinese CMDC datasets. Our method achieves F1-Scores (D) of 0.90, 0.80, and 0.97 on these datasets, respectively, outperforming previous methods. These results demonstrate the effectiveness and superiority of our approach.

Keywords: Affective computing; Automatic depression detection (ADD); Multimodal; Transformer; Contrastive learning

1. Introduction

Depression is a common mental disorder typically characterized by persistent low mood and loss of interest in daily activities [1]. It severely impacts patients’ daily functioning. According

*Corresponding author: Turdi Tohti (Email: turdy@xju.edu.cn; ORCID: 0000-0002-1639-8899).

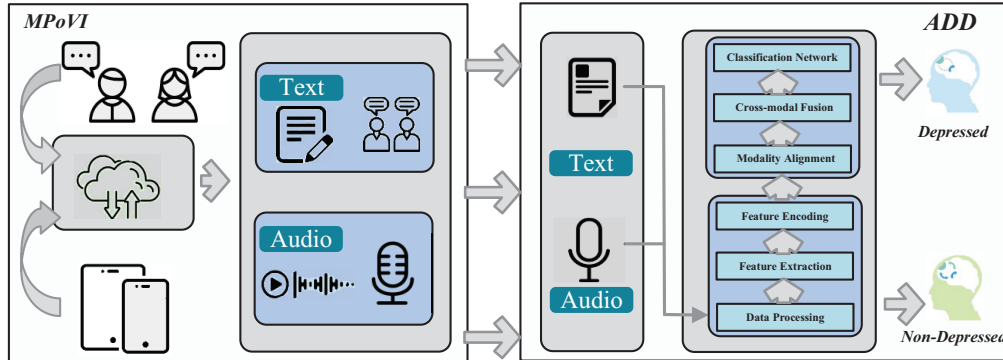


Figure 1: An illustrated example of automatic depression detection (ADD) based on AI algorithms through multimodal information fusion. Text, and audio data provided by mobile applications or virtual interviews (MPoVI) are highly valuable for depression detection.

to 2017 data from the World Health Organization (WHO), more than 300 million people worldwide are affected by depression [2]. This figure has continued to rise, especially following the COVID-19 pandemic [3]. Early diagnosis of depression can significantly reduce the global disease burden [4, 5, 6]. Clinical scales for early depression screening often use the Patient Health Questionnaire-8 (PHQ-8) scores by employing a cutoff of 10 to classify individuals as depressed (D) or non-depressed (ND) [7]. The PHQ-8 consists of eight items, each asking respondents how often over the past two weeks they have been bothered by symptoms such as loss of pleasure (anhedonia), feeling down, sleep disturbances, lack of energy, appetite changes, low self-esteem, difficulty concentrating, and psychomotor slowing. However, large-scale screening significantly increases clinicians' workload and associated costs. Furthermore, diagnosis often requires consultation with multiple experienced clinicians, which not only consumes considerable clinician time and effort but may also delay patients' access to timely treatment and recovery. Unlike traditional approaches, deep learning-based automated depression detection (ADD) methods rely on conversational data collected via human or virtual interviews. These methods learn from and make predictions based on multiple modalities—such as text and audio—as illustrated in Figure 1. Their lower cost and higher efficiency make ADD methods particularly promising for applications in sentiment analysis, remote healthcare, and mental health assessment.

In the early stages of ADD, researchers focused on traditional machine learning methods and handcrafted unimodal features to capture depression-related signals. For example, Cohn et al. [8] utilized facial actions and vocal prosody, while Stratou et al. [9] and Scherer et al. [10] explored nonverbal behaviors such as posture and gesture. Yu et al. [11] proposed multimodal approaches by modeling verbal and nonverbal dynamics in dialogues. With the advancement of deep learning, recent works have shifted toward multimodal supervised learning frameworks. Haque et al. [12] combined 3D facial expressions with speech to assess depression severity. Al Hanai et al. [13] applied audio-text sequence modeling, and Chen et al. [14] introduced a hierarchical framework for paralinguistic processing. Yang et al. [15] further improved depression detection by enhancing multimodal representation learning through knowledge transfer. In addition, Gimeno-Gómez et al. [16] focused on identifying nonverbal cues from video content to support multi-modal depression analysis. These deep learning-based methods generally outperform traditional approaches in both accuracy and robustness when modalities provide complementary

ID	GT Label	Gender	Audio	Text
#304	[ND, PHQ-8=6]	Female		He's my boyfriend I've been knowing him for about 6 months now.
#319	[D, PHQ-8=13]	Male		I homeschool my ten-year-old son and I'm a homemaker I have five children and married 22 years.

E-DAIC Dataset

ID	GT Label	Gender	Audio	Text
#011	[ND, PHQ-8=2]	Female		嗯，我最近身体状态 挺好的，嗯，也没感 冒，可能我是东北人 觉得这边还不是很 冷。
#056	[D, PHQ-8=20]	Female		身体状况，嗯，还 还可以，就是小毛病不 断，大毛病好像没什 么。

CMDC Dataset

Figure 2: Examples of high semantic similarity and subtle differences in dialogues between non-depressed (ND) and depressed (D) interviewees on the E-DAIC and CMDC datasets.

information. For example, if a speaker discusses negative topics with an incongruously upbeat tone, the model can detect depression by leveraging the contrast between textual and vocal cues. However, depressed (D) and non-depressed (ND) samples may be semantically very similar, exhibiting only extremely subtle differences that are difficult to detect (see Figure 2). Furthermore, DL-based ADD methods heavily rely on large multimodal datasets—including text, audio, and visual streams—annotated with ground truth (GT) labels. On the one hand, privacy concerns limit sample sizes. On the other hand, ground truth annotation by medical experts is labor-intensive and time-consuming. Such high inter-sample similarity, combined with limited data scale from privacy constraints and collection costs, presents a major challenge for ADD tasks.

Considering the aforementioned challenges, we propose a supervised multimodal (text and audio) contrastive learning method for ADD, especially when dealing with small-scale datasets such as DAIC-WOZ. Our approach uses supervised contrastive learning to handle the synonymous and differential feature sets we construct, and adopts a mixup-based data augmentation strategy to expand the training data by combining different samples within each batch. We focus on text and audio modalities because prior research [17, 18] has shown that their combination offers the strongest complementarity and generally yields the best performance. In our method, we utilize the feature values of four key types of features as input for multi-task learning: 1) Synonymous features: The projection space derived from similar original features tends to exhibit similar semantic information. 2) Differential features: The projection space derived from dissimilar original features typically displays different semantic information. 3) Mixed features: Features obtained by mixing different samples within the same batch tend to create linear and smooth decision boundaries. 4) Fused features: Fused features are generated by combining synonymous and differential features from the text and audio modalities, while preserving the original characteristics of each modality, and serve as the model’s primary optimization-target inputs.

The contributions of this work can be summarized as follows:

- We propose IICL-DTA, a novel intra-inter contrastive decomposition framework for automated depression detection. Its core contrastive learning module decomposes the objective into two independent losses: an intra-sample contrastive loss that focuses on enhancing decomposed modal synonymous and differential features, and an inter-sample contrastive loss that performs supervised contrast on utterance-level combined synonymous features based on labels. Through joint optimization, IICL-DTA effectively utilizes both decompo-

sition and contrastive information. Furthermore, a customized data sampler with a cache pool module allows us to sample more negative pairs for the learning task.

- We propose decomposing text and audio modalities into synonymous features and differential features. This is motivated by the observation that depression detection requires jointly capturing pathological signals from text and audio and distinguishing between cross-modal consistent and modality-specific information. Through a multi-task prediction loss that relies on the representations of each decomposed modality, the model is enabled to learn from both multimodal and unimodal predictions simultaneously.
- We propose a feature-level Mixup data augmentation method. By separately mixing sample features and their corresponding soft labels, data diversity is increased, facilitating the model in learning smoother decision boundaries, thereby enhancing generalization ability in scenarios with limited data.

The rest of this paper is organized as follows: Section 2 introduces related work. Section 3 presents the theoretical basis and specific details of the proposed IICL-DTA. Section 4 describes the datasets and experimental setup. Section 5 reports the experimental results, followed by discussion and analysis. Finally, Section 6 concludes the paper and outlines future work directions.

2. Related Work

In this section, we review automated depression detection methods, contrastive learning methods, and data augmentation methods to provide a better understanding of our work.

2.1. Automatic Depression Detection

Cohn et al. [8] first proposed the concept of automated depression detection (ADD) by automatically measuring facial actions and vocal prosody in clinically diagnosed depressed patients, i.e., measuring facial expressions via manual FACS coding and Active Appearance Modeling (AAM), and audio prosody via pitch extraction. Subsequently, various methods combining hand-crafted features with machine learning [8, 9, 11, 10, 12, 19, 20, 21] were proposed to enhance ADD performance. These methods, primarily developed in the early stages, relied on feature engineering and statistical models, and were thus labeled "traditional methods." In recent years, there has been a surge in the application of deep learning (DL) techniques in ADD tasks. Some studies [12, 22, 23, 24, 25] have employed Convolutional Neural Networks (CNNs), using selected textual, or audio and visual feature points, as input for ADD. Another category of deep learning-based methods [13, 26, 27, 28] utilizes Long Short-Term Memory (LSTM) networks as the backbone model and employs temporal data features as model input. Furthermore, methods based on a combination of CNN and LSTM [29, 30, 31] have also been introduced to achieve high performance, especially on multimodal data [29]. Recently, [14, 16, 32] proposed methods employing Transformer (including BERT) architectures to obtain ADD representations; these methods can also be extended to multimodal data, achieving high performance. The Transformer architecture is particularly advantageous due to its superior parallelism and robustness. However, none of the aforementioned methods have considered addressing difficult-to-classify samples, such as learning to distinguish the semantic information in samples where the differences between depressed and non-depressed states are subtle, to further improve ADD performance.

2.2. Contrastive Learning

Contrastive learning is a self-supervised learning technique widely adopted in computer vision tasks, which allows deep learning models to map high-dimensional images or videos into low-dimensional feature embeddings without requiring labeled data [14, 16, 32, 33, 34, 35, 36, 37, 38]. Its main objective is to ensure that features from different views of the same sample (termed positive pairs) are pulled closer, while features from different samples (termed negative pairs) are pushed apart. This approach has found extensive use in pre-training models, thereby facilitating training for downstream tasks such as image classification [36], audio classification [39], and video analysis [40, 41]. Akbari et al. [42] used multimodal contrastive learning to train a Video-Audio-Text Transformer (VATT) for aligning video-text and video-audio pairs, thereby achieving state-of-the-art results on various computer vision tasks. Chen et al. [36] proposed a self-supervised framework, SimCLR, which learns visual representations through a contrastive loss between augmented views of the same image sample. Khosla et al. [43] extended self-supervised contrastive learning to supervised scenarios, i.e., performing contrastive learning on samples from different classes; this supervised setting is more stable with respect to hyperparameters. We designed a novel intra-inter contrastive learning framework that utilizes intra-sample and inter-sample contrast for text and audio modalities. It enhances text and audio modal representations by unifying the two intra-inter contrastive loss terms and incorporating a customized pairing strategy. Furthermore, unlike previous work that obtains positive pairs through data augmentation, we propose a data sampler with a cache pool to retrieve similar samples as positive pairs. The cache pool significantly increases the scale of negative pairs.

2.3. Data Augmentation

Data augmentation techniques have been widely applied in computer vision and natural language processing tasks, aiming to expand the scale of labeled datasets and improve model resilience to input noise. For visual image tasks, geometric and photometric transformations (e.g., cropping, flipping, rotation, resizing, color distortion, and Gaussian blur) [36, 44] are common methods. For text data tasks, simpler syntactic transformations such as word deletion or re-ordering have been shown to be beneficial for improving robustness and performance [44]. Furthermore, semantic augmentation techniques (such as back-translation, synonym replacement, or generative models) have also been actively explored [45, 46, 47]. However, such methods often involve higher implementation complexity, depend on external knowledge (like thesauri) or additional models, and are thus primarily applicable to tasks with adequate resources. We propose a feature-level Mixup data augmentation method that, unlike previous work focusing on data-level augmentation, requires no external knowledge or additional models. It encourages the model to learn smoother decision boundaries through the design of corresponding soft labels.

3. Methodology

This section first introduces the problem definition of the ADD task, followed by the overall framework of IICL-DTA. Subsequently, we provide a detailed description of the contrastive feature decomposition process for learning multimodal representations, and finally, we elaborate on the mixed features and soft labeling process of the Mixup feature augmentation method.

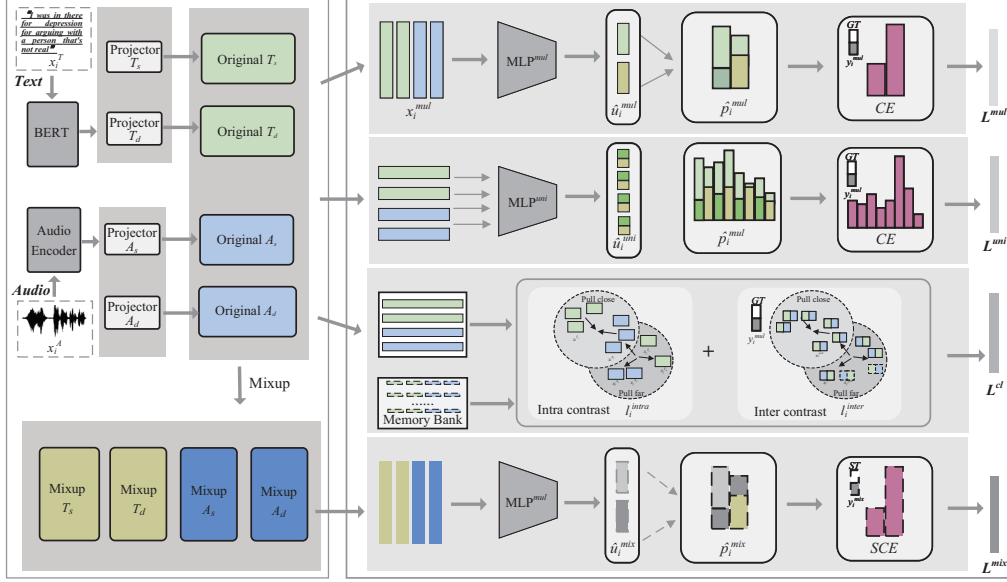


Figure 3: Overview of the proposed IICL-DTA framework. It mainly includes: a feature decomposition and mixing module, a multimodal prediction module (including text and audio data), a unimodal prediction module, a contrastive learning module, and a mixup prediction module.

3.1. Problem Formulation

To facilitate subsequent description, we denote each interview participant as P . For each P , N consecutive utterances are predefined as $(x_1, x_2, \dots, x_N)$. Furthermore, each utterance sample x_i includes text (T) and audio (A) modalities, with their feature representations denoted as $x_i^T \in \mathbb{R}^{d_T}$ and $x_i^A \in \mathbb{R}^{d_A}$, respectively. We represent the sequences of all text and audio utterances in a session as $X^T = [x_1^T; x_2^T; \dots; x_N^T] \in \mathbb{R}^{N \times d_T}$ and $X^A = [x_1^A; x_2^A; \dots; x_N^A] \in \mathbb{R}^{N \times d_A}$, respectively. The ADD task aims to predict the GT label y_i^{mul} for each interview utterance sample x_i from predefined categories, i.e., mapping x_i to its predicted multimodal label $\hat{y}_i^{\text{mul}}$. Here, $y_i^{\text{mul}} = \text{D}$ (1) represents the GT label for an utterance from participant P being depressed, while $y_i^{\text{mul}} = \text{ND}$ (0) represents the GT label for an utterance from participant P being non-depressed.

3.2. Model Architecture

The overall framework of IICL-DTA is shown in Figure 3. IICL-DTA includes feature extraction, modality fusion, modal feature decomposition, intra-modal and inter-modal contrastive learning, and feature mixing. First, given a sample, we use corresponding feature extractors to encode the two modalities. Specifically, we use the [CLS] classification token of the BERT model to encode text (T). We use the Speechformer++ encoder to encode the audio modality (More details of Speechformer++ can be found in [14]). Then, we use different projection modules to decompose the text and audio modalities into a synonymous feature (i.e., T_s and A_s in the figure) and a differential feature (i.e., T_d and A_d in the figure). Each projection layer consists of layer normalization, a linear layer with ReLU activation, a Tanh activation function, and a dropout layer. Finally, the 4 decomposed original features and 4 mixed features are updated

and fused separately, and the IICL-DTA model is trained using the following multi-task learning objective function:

$$L^{\text{total}} = L^{\text{mul}} + \alpha L^{\text{uni}} + \beta L^{\text{cl}} + \gamma L^{\text{mix}} \quad (1)$$

where L^{mul} represents the multimodal prediction loss for the ADD task, L^{uni} represents the uni-modal prediction loss, L^{cl} represents the contrastive loss, and L^{mix} represents the prediction loss for mixup mixed features. α , β , and γ are hyperparameters that balance the contribution of each regularization component to the overall loss L^{total} . A detailed introduction to each loss term is as follows.

3.2.1. Multimodal Prediction Loss

We use a Multi-Layer Perceptron (denoted as MLP^{mul}) with a ReLU activation function and Dropout regularization as the multimodal classifier, which consists of a fully connected neural network with two hidden layers. We concatenate all 4 decomposed modal features to obtain the classifier's input fused feature $\mathbf{x}_i^{\text{mul}}$. Specifically, for each discourse sample x_i , its fused feature is represented as $\mathbf{x}_i^{\text{mul}} = [\mathbf{x}_i^{T_s} \parallel \mathbf{x}_i^{T_d} \parallel \mathbf{x}_i^{A_s} \parallel \mathbf{x}_i^{A_d}]$, where $\parallel$ denotes the vector concatenation operation. We input this fused feature into the MLP to obtain a logits vector $\hat{\mathbf{u}}_i^{\text{mul}} \in \mathbb{R}^C$, which represents the predicted scores for each class, where C is the number of classes. The logits vector $\hat{\mathbf{u}}_i^{\text{mul}} = [\hat{u}_{i,1}^{\text{mul}}, \hat{u}_{i,2}^{\text{mul}}, \dots, \hat{u}_{i,C}^{\text{mul}}]$, where $\hat{u}_{i,j}^{\text{mul}}$ corresponds to the predicted logit value for class j . Subsequently, the logits $\hat{\mathbf{u}}_i^{\text{mul}}$ are converted into a predicted probability distribution $\hat{\mathbf{p}}_i^{\text{mul}} = [\hat{p}_{i,1}^{\text{mul}}, \hat{p}_{i,2}^{\text{mul}}, \dots, \hat{p}_{i,C}^{\text{mul}}]$ using the Softmax function, representing the probability of sample i belonging to each class. The predicted logits vector $\hat{\mathbf{u}}_i^{\text{mul}}$ and the multimodal prediction probability $\hat{p}_{i,c}^{\text{mul}}$ for class c are given by:

$$\hat{\mathbf{u}}_i^{\text{mul}} = \text{MLP}^{\text{mul}}(\mathbf{x}_i^{\text{mul}}) = \text{MLP}^{\text{mul}}([\mathbf{x}_i^{T_s} \parallel \mathbf{x}_i^{T_d} \parallel \mathbf{x}_i^{A_s} \parallel \mathbf{x}_i^{A_d}]) \quad (2)$$

$$\hat{p}_{i,c}^{\text{mul}} = \frac{\exp(\hat{u}_{i,c}^{\text{mul}})}{\sum_{j=1}^C \exp(\hat{u}_{i,j}^{\text{mul}})} \quad (3)$$

where c represents the class, whose index is also c , $c \in \{1, \dots, C\}$, C denotes the number of classes, and $\exp(\cdot)$ represents the natural exponential function. The model's predicted label $\hat{\mathbf{y}}_i^{\text{mul}}$ is obtained by taking the Argmax of the probability distribution $\hat{\mathbf{p}}_i^{\text{mul}}$, i.e., selecting the class with the highest probability to get the final classification result.

Denote the set of samples in a batch as B . For a given discourse sample $x_i \in B$, its ground truth (GT) label is $\mathbf{y}_i^{\text{mul}}$. We represent it using a one-hot encoded vector, where $\mathbf{y}_i^{\text{mul}} = [y_{i,1}^{\text{mul}}, \dots, y_{i,C}^{\text{mul}}]$. In this vector, if sample i belongs to class c , then the component $y_{i,c}^{\text{mul}} = 1$, and all other components are 0. We calculate the multimodal prediction loss using cross-entropy error:

$$L^{\text{mul}} = -\frac{1}{n} \sum_{i=1}^n \sum_{c=1}^C y_{i,c}^{\text{mul}} \log(\hat{p}_{i,c}^{\text{mul}}) \quad (4)$$

where n represents the number of samples in batch B .

3.2.2. Unimodal Prediction Loss

For each discourse sample x_i , we initially decompose it into four distinct unimodal feature representations: $[x_i^{T_s}, x_i^{T_d}, x_i^{A_s}, x_i^{A_d}]$. Let k denote the type of unimodal feature, where $k \in \{T_s, T_d, A_s, A_d\}$. Each of these four unimodal feature vectors, $x_i^{(\text{uni},k)}$, is individually fed into a weight-shared unimodal MLP classifier, referred to as MLP^{uni} . This process yields an independent predicted logits vector, $\hat{u}_i^{(\text{uni},k)} \in \mathbb{R}^C$, for each of the four feature types. Specifically, the unimodal prediction loss is computed as follows:

$$\hat{u}_i^{(\text{uni},k)} = \text{MLP}^{\text{uni}}(x_i^{(\text{uni},k)}) = \text{MLP}^{\text{uni}}([x_i^{T_s}, x_i^{T_d}, x_i^{A_s}, x_i^{A_d}]) \quad (5)$$

$$\hat{p}_{i,c}^{(\text{uni},k)} = \frac{\exp(\hat{u}_{i,c}^{(\text{uni},k)})}{\sum_{j=1}^C \exp(\hat{u}_{i,j}^{(\text{uni},k)})} \quad (6)$$

$$L^{\text{uni}} = \sum_{k \in \{T_s, T_d, A_s, A_d\}} \left(-\frac{1}{n} \sum_{c=1}^C \log(\hat{p}_{i,c}^{(\text{uni},k)}) \right) \quad (7)$$

Specifically, the unimodal prediction loss L^{uni} is computed by summing the cross-entropy losses independently generated by each of the four unimodal features, followed by averaging over the batch. For supervision, each unimodal feature's predicted label $\hat{y}_i^{(\text{uni},k)}$ utilizes the same multimodal ground truth label y_i^{mul} . This means that each decomposed feature is regularized to perform prediction independently.

By imposing separate prediction constraints on the synonymy and discrepancy features, compelling each decomposed feature to independently predict the common multimodal label, the model is encouraged to enhance the discriminative power of each feature. Synonymy features are tasked with learning common cues across modalities, while discrepancy features are intended to distill modality-specific discriminative information. The summation of these independent losses acts as a regularization term, driving all decomposed features to learn effective representations that contribute to the main task.

3.2.3. Contrastive Loss

We further regularize the learned feature decomposition through a joint contrastive loss, which (1) performs synonymy contrast among the decomposed features within the same sample, and (2) conducts discrepancy contrast between samples based on their label-based differentiation objective. The contrastive loss is denoted as:

$$L^{\text{cl}} = \frac{1}{n} \sum_{i=1}^n (\beta_1 l_i^{\text{intra}} + \beta_2 l_i^{\text{inter}}) \quad (8)$$

where l_i^{intra} and l_i^{inter} are the intra-sample and inter-sample contrastive losses for sample i , respectively. β_1 and β_2 are weighting coefficients. The detailed derivation will be presented in Section 3.3.

3.2.4. Mixup Loss

We compute the Mixup loss using soft targets (ST) and soft cross-entropy (SCE). Specifically, for sample i in the batch, we calculate the discrepancy between the predicted probability

distribution $\hat{p}_i^{\text{mix}}$ of the mixed feature and the ST label y_i^{soft} . The Mixup loss L^{mix} can be expressed as:

$$L^{\text{mix}} = -\frac{1}{n} \sum_{i=1}^n \sum_{c=1}^C y_{i,c}^{\text{soft}} \log(\hat{p}_{i,c}^{\text{mix}}) \quad (9)$$

where n is the number of samples in batch B , $y_{i,c}^{\text{soft}}$ is the soft label component for sample i corresponding to class c , and $\log(\hat{p}_{i,c}^{\text{mix}})$ is the LogSoftmax output computed on class c from the logits of the mixed feature for sample i . The detailed derivation will be presented in Section 3.4.

3.3. Intra and Inter Sample Contrastive Feature Decomposition

We unify the intra- and inter-sample contrastive learning with feature decomposition within the InfoNCE framework [36], augmented with a negative sample queue [48] to increase the number of sampled instances. We leverage labels to guide supervised contrastive learning among samples [43]. The intra- and inter-sample contrastive loss is given by Equation 8. We will now detail the intra-sample contrastive loss l_i^{intra} and the inter-sample contrastive loss l_i^{inter} .

3.3.1. Intra-Sample Contrastive Loss

The intra-sample contrastive loss l_i^{intra} is designed to optimize the decomposed features within a discourse sample x_i . These features include the text synonymy feature $\mathbf{x}_i^{T_s}$, the audio synonymy feature $\mathbf{x}_i^{A_s}$, the text discrepancy feature $\mathbf{x}_i^{T_d}$, and the audio discrepancy feature $\mathbf{x}_i^{A_d}$. We construct a set $\mathbf{x}_B^{\text{dec}}$ containing all decomposed features from the n discourse samples in each batch B , where $\mathbf{x}_B^{\text{dec}} = \{\mathbf{x}_j^k \mid j \in \{1, \dots, n\}, k \in \{T_s, T_d, A_s, A_d\}\}$. We treat each of the four decomposed features as an anchor, and the loss for sample i is then averaged over the losses derived from these anchors. Specifically, for an anchor $\mathbf{x}_i^{\text{dec}_a}$, its contrastive loss $l_i^{\text{intra,dec}_a}$ and the intra-sample contrastive loss l_i^{intra} are given by the following equations:

$$l_i^{\text{intra,dec}_a} = -\left(\frac{\sum_{\mathbf{x}_j^{\text{dec}_p} \in \mathbf{P}_i^{\text{intra}}} \exp(\text{sim}(\mathbf{x}_i^{\text{dec}_a}, \mathbf{x}_j^{\text{dec}_p})/\tau) + \epsilon}{\sum_{\mathbf{x}_j^{\text{dec}_m} \in \mathbf{K}_i^{\text{intra}}} \exp(\text{sim}(\mathbf{x}_i^{\text{dec}_a}, \mathbf{x}_j^{\text{dec}_m})/\tau) + \epsilon} \right) \quad (10)$$

$$l_i^{\text{intra}} = \frac{1}{4} \sum_{k \in \{T_s, T_d, A_s, A_d\}} l_i^{\text{intra,dec}_k} \quad (11)$$

where $\text{sim}(\cdot)$ is cosine similarity, τ is the temperature coefficient, and ϵ is a small smoothing term. dec_a denotes the anchor term from the 4 decomposed features, dec_p denotes the positive sample term, and dec_m denotes the contrastive term.

Specifically, the within-sample positive sample set $\mathbf{P}_i^{\text{intra}}$ and contrastive set $\mathbf{K}_i^{\text{intra}}$ are given by:

$$\mathbf{P}_i^{\text{intra}} = \begin{cases} \{\mathbf{x}_i^{A_s}\} & \text{if } \text{dec}_a = T_s \\ \{\mathbf{x}_i^{T_s}\} & \text{if } \text{dec}_a = A_s \\ \emptyset & \text{if } \text{dec}_a = T_d \text{ or } \text{dec}_a = A_d \end{cases} \quad (12)$$

Discrepancy features (T_d, A_d) have no intra-sample positive pairs, as they are designed to capture modality-specific information that should not be aligned across modalities within the same sample.

$$\mathbf{K}_i^{\text{intra}} = \{\mathbf{x}_j^{\text{dec}_m} \in \mathbf{x}_B^{\text{dec}} \mid (j, \text{dec}_m) \neq (i, \text{dec}_a)\} \quad (13)$$

where Equation 12 discusses different cases to define the positive sample set. j represents the sample index, dec_m is the contrastive feature type $\in \{T_s, T_d, A_s, A_d\}$, and the contrastive set $\mathbf{K}_i^{\text{intra}}$ contains all other decomposed features in set $\mathbf{x}_B^{\text{dec}}$ except for the anchor $\mathbf{x}_i^{\text{dec}_a}$ itself.

3.3.2. Inter-Sample Contrastive Loss

The inter-sample contrastive loss l_i^{inter} performs supervised contrastive learning at the discourse level, aiming to enhance the model's ability to discriminate between discourses with different depression labels. First, for each discourse x_i , we construct a combined synonymy feature. This feature is obtained by combining and normalizing its text synonymy feature $\mathbf{x}_i^{T_s}$ and audio synonymy feature $\mathbf{x}_i^{A_s}$, denoted as $\mathbf{x}_i^{\text{syn}} = \text{Norm}(\mathbf{x}_i^{T_s} + \mathbf{x}_i^{A_s})$. We then form a set $\mathbf{x}_B^{\text{syn}}$ containing all combined synonymy features from the n discourses in the current batch, where $\mathbf{x}_B^{\text{syn}} = \{\mathbf{x}_j^{\text{syn}} \mid j \in \{1, \dots, n\}\}$.

Note that we construct a Memory Bank (M) to ensure that negative samples with different labels are always available in the current batch for effective contrast, especially when the batch size is small or the class distribution within the batch is imbalanced. Specifically, this memory bank stores the combined synonymy features $\mathbf{x}^{\text{syn}}$ and their corresponding ground truth labels y^{mul} from previous training iterations. During the current batch's training, we combine the current batch's $\mathbf{x}_B^{\text{syn}}$ with a set of negative sample features $\mathbf{x}_M^{\text{neg}}$ filtered from M , forming an extended feature set $\mathbf{x}^{\text{comb}} = \mathbf{x}_B^{\text{syn}} \cup \mathbf{x}_M^{\text{neg}}$. Here, the features in $\mathbf{x}_M^{\text{neg}}$ have the same form as $\mathbf{x}^{\text{syn}}$, and their corresponding labels are different from the labels of all samples in the current batch.

For a discourse synonymy feature $\mathbf{x}_i^{\text{syn}}$ in the current batch used as an anchor, denoted as $\mathbf{x}_i^{\text{syn}_a}$, its inter-sample contrastive loss is given by the following equation:

$$l_i^{\text{inter}} = -\frac{1}{|\mathbf{P}_i^{\text{inter}}| + \epsilon} \sum_{\mathbf{x}_j^{\text{syn}_p} \in \mathbf{P}_i^{\text{inter}}} \log \left(\frac{\exp(\text{sim}(\mathbf{x}_i^{\text{syn}_a}, \mathbf{x}_j^{\text{syn}_p})/\tau)}{\sum_{\mathbf{x}_{j'}^{\text{syn}_m} \in \mathbf{K}_i^{\text{inter}}} \exp(\text{sim}(\mathbf{x}_i^{\text{syn}_a}, \mathbf{x}_{j'}^{\text{syn}_m})/\tau) + \epsilon} \right) \quad (14)$$

where $\text{sim}(\cdot)$ is cosine similarity, τ is the temperature coefficient, and ϵ is a small smoothing term. syn_a denotes the anchor term originating from the synonymy features, syn_p denotes the positive sample term, and syn_m denotes the contrastive term. $|\mathbf{P}_i^{\text{inter}}|$ denotes the number of positive samples in the set $\mathbf{P}_i^{\text{inter}}$. The inter-sample positive sample set $\mathbf{P}_i^{\text{inter}}$ is defined as $\mathbf{P}_i^{\text{inter}} = \{\mathbf{x}_j^{\text{syn}_p} \in \mathbf{x}_B^{\text{syn}} \mid y_j^{\text{mul}} = y_i^{\text{mul}}, j \neq i\}$, and the contrastive set $\mathbf{K}_i^{\text{inter}}$ is defined as $\mathbf{K}_i^{\text{inter}} = \{\mathbf{x}_{j'}^{\text{syn}_m} \in \mathbf{x}^{\text{comb}} \mid \mathbf{x}_{j'}^{\text{syn}_m} \neq \mathbf{x}_i^{\text{syn}_a}\}$.

Inter-sample contrastive loss encourages the synonymy features of discourse with the same label to be closer to each other in the representation space, while the similarity features of discourse with different labels are pushed farther apart. The cache pool is updated after each iteration using the current batch's $\mathbf{x}_B^{\text{syn}}$ and its true labels $\mathbf{y}_B^{\text{mul}} = \{y_j^{\text{mul}}\}_{j=1}^n$.

3.4. Mixup Mixed Features

To further enhance the generalization ability and robustness of our multimodal classification model, we introduce a Mixup regularization strategy [49] applied to the fused multimodal features. The core of this strategy lies in training the multimodal classifier (MLP^{mul}) to process linearly interpolated samples and their corresponding softened labels. The related loss term is

denoted as L^{mix} , used to quantify the difference between the model’s predictions on these mixed inputs and the interpolated targets.

Specifically, for a batch B containing n samples, the process for calculating L^{mix} is as follows. First, a mixing coefficient λ is sampled from a symmetric Gama distribution, i.e., $\lambda \sim \text{Gama}(\gamma^{\text{mix}}, \gamma^{\text{mix}})$. Here, γ^{mix} is a predefined hyperparameter that controls the interpolation intensity.

Second, a random permutation π of the batch indices $\{1, \dots, n\}$ is generated. This permutation determines which sample pairs within the batch will be mixed. Then, using the sampled coefficient λ and the permutation π , the fused multimodal features $\mathbf{x}_i^{\text{mul}}$ and their corresponding one-hot encoded GT labels $\mathbf{y}_i^{\text{mul}}$ (whose c -th component is $y_{(i,c)}^{\text{mul}}$) are mixed. Here, $\mathbf{x}_i^{\text{mul}}$ is defined by Equation 2 and represents the output of the fusion layer for sample i . The mixed feature $\mathbf{x}_i^{\text{mix}}$ and the ST label $\mathbf{y}_i^{\text{soft}}$ are given by:

$$(\mathbf{x}_i^{\text{mix}}, \mathbf{y}_i^{\text{soft}}) = \lambda(\mathbf{x}_i^{\text{mul}}, \mathbf{y}_i^{\text{mul}}) + (1 - \lambda)(\mathbf{x}_{\pi(i)}^{\text{mul}}, \mathbf{y}_{\pi(i)}^{\text{mul}}) \quad (15)$$

where λ denotes the mixing coefficient sampled from the Gama distribution, π represents a random permutation applied to the batch indices, and the c -th component of the ST label $\mathbf{y}_i^{\text{soft}}$ is given by $y_{(i,c)}^{\text{soft}} = \lambda y_{(i,c)}^{\text{mul}} + (1 - \lambda) y_{(\pi(i),c)}^{\text{mul}}$.

Subsequently, we input the mixed features $\mathbf{x}_i^{\text{mix}}$ into the multimodal classifier MLP^{mul} with shared parameters, obtaining the predicted mixed Logits vector $\hat{\mathbf{u}}_i^{\text{mix}}$. We then convert these mixed Logits into a predicted probability distribution $\hat{\mathbf{p}}_i^{\text{mix}}$ across C classes using the Softmax function. $\hat{\mathbf{u}}_i^{\text{mix}}$ and $\hat{\mathbf{p}}_i^{\text{mix}}$ are given by:

$$\hat{\mathbf{u}}_i^{\text{mix}} = \text{MLP}^{\text{mul}}(\mathbf{x}_i^{\text{mix}}) \quad (16)$$

$$\hat{p}_{i,c}^{\text{mix}} = \frac{\exp(\hat{u}_{i,c}^{\text{mix}})}{\sum_{j=1}^C \exp(\hat{u}_{i,j}^{\text{mix}})} \quad (17)$$

where c denotes the class, and C represents the number of classes. $\hat{p}_{(i,c)}^{\text{mix}}$ denotes the predicted probability of the mixed sample $\mathbf{x}_i^{\text{mix}}$ belonging to class c . As shown in Equation 9, the Mixup loss L^{mix} is finally calculated by soft cross-entropy. By minimizing L^{mix} , the model is explicitly encouraged to exhibit linear behavior in the mapping from the fused feature space to the output Logits, especially between training sample pairs. This property serves as a form of data augmentation and regularization, which helps to form smoother decision boundaries, thereby improving the model’s robustness and its generalization ability to unseen data. (Note: All **non-scalar symbols** in the entire document’s formulas are **bolded**.)

4. Experimental Details

4.1. Datasets

Our proposed model is evaluated on the DAIC-WOZ [50], E-DAIC [51], and CMDC [52] datasets. All three datasets use the 8-item Patient Health Questionnaire (PHQ-8 [6]) for depression assessment. As shown in Table 1, PHQ-8 quantifies the frequency of eight depressive symptoms (0=not at all to 3=nearly every day) over 14 days through self-report, yielding a total

Table 1: PHQ-8 Score ranges with depression severity and state, and sample distribution in DAIC-WOZ, E-DAIC and CMDC datasets.

PHQ-8 Score	Depression severity	State	DAIC-WOZ	E-DAIC	CMDC
0-4	Not significant	No	86 (45%)	122 (44%)	47 (60%)
5-9	Mild	No	46 (24%)	67 (24%)	6 (8%)
10-14	Moderate	Yes	30 (16%)	43 (16%)	10 (13%)
15-19	Moderately severe	Yes	20 (11%)	33 (12%)	7 (9%)
20-24	Severe	Yes	7 (4%)	10 (4%)	8 (10%)

score ranging from 0 to 24. Using a clinical cutoff (PHQ-8 total score ≥ 10), subjects are classified as depressed (D) or non-depressed (ND). Detailed severity stratification and binary labels (status) are provided in Table 1.

DAIC-WOZ: DAIC-WOZ is part of a large corpus called the "Distress Analysis Interview Corpus (DAIC-WOZ)", which was used in the Audio/Visual Emotion Challenge and Workshop (AVEC) competitions in 2016 and 2017. It contains clinical interviews designed to support the diagnosis of mental distress conditions such as anxiety, depression, and post-traumatic stress disorder. The dataset includes clinical interview recordings of 189 participants, with 56 samples diagnosed with depression by medical experts. The dataset is divided into a training set ($n = 107$), a development set ($n = 35$), and a test set ($n = 47$). All interviews were conducted by an animated virtual interviewer named Ellie and controlled by a human interviewer in another room. The data collected from all interviews is multimodal, including text, video, and audio information.

E-DAIC: E-DAIC is an extended version of DAIC, namely the "Extended Distress Analysis Interview Corpus (E-DAIC)", used in the AVEC 2019 competition. This dataset includes clinical interview recordings of 219 participants (275 interview sessions in total), with 54 samples diagnosed with depression by medical experts. The dataset is divided into a training set ($n = 163$), a development set ($n = 56$), and a test set ($n = 56$). Similar to the DAIC-WOZ dataset, it supports preliminary screening for depression based on semi-structured interviews. Experts used HAMD-17 [58] and self-reported PHQ-8 for assessment. Unlike DAIC-WOZ, its training and development sets are based on clinical interviews. In contrast, the test set consists entirely of autonomous interviews, making the ADD task more challenging than in DAIC-WOZ.

CMDC: is a Chinese depression detection dataset proposed by literature [52]. This dataset was obtained from audio and video recordings of patient clinical interviews in a Chinese context. CMDC pre-sets 12 questions during the interview process and stores the answer data for these 12 questions. In addition, similar to the two datasets above, CMDC also provides self-reported PHQ-8 scores from the interviewees. Likewise, to ensure patient privacy, the dataset only provides processed data for the audio and visual modalities. The dataset contains 78 samples with complete modalities: 26 samples from depressed patients and 52 from the control group. Given the relatively small data volume, the authors provided stratified 5-fold cross-validation to obtain the final performance of the model.

4.2. Experimental Setup

The proposed model was implemented using PyTorch 3 on an NVIDIA GeForce GTX 4090 GPU. Training utilized the Stochastic Gradient Descent (SGD) [53] optimizer with initial learn-

ing rates of 0.001 (DAIC-WOZ, E-DAIC) and 0.0001 (CMDC), decayed to 1% of their initial values via cosine annealing. Models were trained for 60 epochs (DAIC-WOZ, E-DAIC) and 80 epochs (CMDC). We extracted 768-dimensional text embeddings using pre-trained BERT base and 1024-dimensional audio embeddings using pre-trained HuBERT large. The total loss weights were configured as $\alpha = 0.1, \beta = 0.3, \gamma = 0.1$ for the DAIC-WOZ dataset, while for the E-DAIC and CMDC datasets, the weights were set to $\alpha = 0.05, \beta = 0.4, \gamma = 0.05$. For contrastive learning, inner/outer loss weights were $\beta_1 = 1, \beta_2 = 1$, with a temperature of 0.07. The memory bank had a batch size of 256, a maximum capacity of 65,536 samples, a sample update ratio of 0.3, and an update threshold of 0.3. To mitigate class imbalance and ensure consistency across datasets, we followed prior works [14, 24, 52] and adopted utterance-level sampling strategies:

For the DAIC-WOZ dataset, all utterances from depressed participants were retained, and the 50 longest utterances were selected from each non-depressed participant, based on the median utterance length statistics. This resulted in 10,488 utterances in the training set (5,626 non-depressed, 4,862 depressed) and 4,505 in the validation set (1,724 non-depressed, 2,781 depressed).

For the E-DAIC dataset, we selected the 70 longest utterances from each non-depressed participant using the same median-based strategy. This yielded 10,909 utterances in the training set (6,450 non-depressed, 4,459 depressed) and 3,806 in the validation set (2,387 non-depressed, 1,419 depressed).

For the CMDC dataset, we applied data augmentation techniques—including text synonym replacement and audio noise addition (e.g., Gaussian noise)—and followed the 5-fold cross-validation protocol described in the original paper [52]. This yielded 1,224 utterances across all folds, with 618 non-depressed and 606 depressed samples, resulting in a nearly balanced distribution.

4.3. Evaluation Metric

As shown in the following formulas, we use commonly used metrics in classification tasks to facilitate comparison with other methods, including accuracy, precision, recall, and F1-Score:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (18)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (19)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (20)$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (21)$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. In our context, for example, TP refers to the count of depressed individuals correctly classified as depressed, while TN indicates the number of non-depressed individuals accurately identified as non-depressed.

For depression detection on the DAIC-WOZ, E-DAIC and CMDC datasets, we aimed for depression classification. Consistent with previous studies [14, 17, 54], we focused on participant dialogues, excluding interviewer utterances. Utterances were extracted from transcriptions using timestamps, processed, classified, and aggregated via majority voting for subject-level predictions, informing the final evaluation.

Table 2: Performance comparison of other methods on DAIC-WOZ dataset. D stands for depression, and ND stands for non-depression. T, A, and V denote Text, Audio, and Vision respectively. † Denotes results on the validation set. * Denotes reproduced results from open-source implementations. Best in **bold**.

Modality	Method	Model	Precision (D/ND)	Recall (D/ND)	F1-Score (D/ND)	F1-Score (mean)
T	Sun et al. [55] (2017)†	Random Forest	0.62 (0.74)	0.42 (0.87)	0.50 (0.80)	0.65
	Soliman et al. [56] (2021)	CNN+LSTM	0.50 (0.74)	0.29 (0.88)	0.36 (0.81)	0.58
	Zou et al. [52] (2023)	BiLSTM	0.91 (–)	0.81 (–)	0.86 (–)	–
A	Huang et al. [57] (2020)†	CNN	–	–	0.46 (0.82)	0.64
	Sardari et al. [24] (2022)	CNN+LSTM*	0.71 (0.70)	0.68 (0.72)	0.70 (0.71)	0.70
	Chen et al. [14] (2023)†	Transformer *	–	–	–	0.70
	Zou et al. [52] (2023)	BiLSTM	0.88 (–)	0.70 (–)	0.77 (–)	–
	Ioannides et al. [58] (2024)	Transformer	–	–	0.60 (0.84)	0.72
V	Valstar et al. [19] (2016)	SVM	0.46 (0.93)	0.77 (0.79)	0.58 (0.85)	0.71
	Li et al. [59] (2023)†	GNN	0.79 (–)	0.80 (–)	0.79 (–)	–
	Zou et al. [52] (2023)	BiLSTM	0.94 (–)	0.74 (–)	0.83 (–)	–
	Zhang et al. [60] (2024)†	XGBoost	(–)	(–)	0.64 (0.83)	0.73
T+A	Zou et al. [52] (2023)	Transformer+CNN	0.88 (–)	0.81 (–)	0.84 (–)	–
	Yang et al. [15] (2024)	BiLSTM	0.76 (–)	0.76 (–)	0.76 (–)	–
V+A	Wei et al. [54] (2022)	Attention+ConvBiLSTM *	0.78 (–)	0.50 (–)	0.61 (–)	–
	Zou et al. [52] (2023)	Transformer+CNN	0.91 (–)	0.75 (–)	0.82 (–)	–
	Yang et al. [15] (2024)	BiLSTM	0.70 (–)	0.70 (–)	0.70 (–)	–
T+V	Zou et al. [52] (2023)	Transformer+CNN	0.97 (–)	0.84 (–)	0.90 (–)	–
	Yang et al. [15] (2024)	BiLSTM	0.78 (–)	0.78 (–)	0.78 (–)	–
T+A+V	Wei et al. [54] (2022)	Sub-Attention+ConvBiLSTM *	0.89 (–)	0.57 (–)	0.70 (–)	–
	Zou et al. [52] (2023)	Transformer+CNN	0.94 (–)	0.81 (–)	0.87 (–)	–
	Jung et al. [32] (2024)	Transformer+Cross-Attention	–	–	–	0.79
	Yang et al. [15] (2024)	BiLSTM	0.80 (–)	0.81 (–)	0.80 (–)	–
T+A	IICL-DTA(Ours)	Transformer+CL	0.87 (0.97)	0.93 (0.94)	0.90 (0.95)	0.93

5. Results and Discussion

In this section, we compare our architecture with state-of-the-art (SOTA) methods, including both unimodal and multimodal approaches. The unimodal methods include Random Forest [55], CNN+LSTM [56], BiLSTM [52], FVTC-CNN [57], CNN AE+SVM [24], SpeechFormer++ [14], DAAM-Audio Transformer [58], SVM [19], DSE-HGAT [59], MADAV [60], Lexical Availability [61], CID+Referee Network [62].

The multimodal methods include MulT [52], RLKT-MDD [15], Sub-atten ConvBiLSTM [54], HiQuE [32] CID+Referee Network [62], [62], Non-Verbal Transformer [16], IIFDD [5]. For methods with publicly available code, we reproduced the results using their original experimental settings and strictly followed the evaluation metrics reported in the corresponding papers. For methods without released implementations, or with incomplete code repositories lacking key files for reproduction, we directly cited the performance reported in the original publications. The distinction between reproduced and cited results is clearly annotated in Table 2 and Table 3.

5.1. Comparison with the SOTA Methods

Tables 2, 3, and 4 present the performance of State-of-the-Art (SOTA) methods and IICL-DTA on the DAIC-WOZ, E-DAIC, and CMDC datasets, including comparative results for uni-

Table 3: Performance comparison of other methods on E-DAIC dataset. D stands for depression, and ND stands for non-depression. T, A, and V denote Text, Audio, and Vision respectively. † Denotes results on the validation set. * Denotes reproduced results from open-source implementations. Best in **bold**.

Modality	Method	Model	Precision (D/ND)	Recall (D/ND)	F1-Score (D/ND)	F1-Score (mean)
T	Villatoro-Tello et al. [61] (2021)†	Logistic Regression	–	–	0.38 (0.90)	0.64
A	Cheong et al. [62] (2024)	Transformer	0.50 (–)	0.51 (–)	0.50 (–)	–
	Gimeno-Gómez et al. [16] (2024)	Transformer *	0.71 (–)	0.52 (–)	0.51 (–)	–
V	Gimeno-Gómez et al. [16] (2024)	Transformer *	0.58 (–)	0.49 (–)	0.50 (–)	–
	Cheong et al. [62] (2024)	Transformer	0.53 (–)	0.58 (–)	0.58 (–)	–
V+A	Cheong et al. [62] (2024)	Transformer	0.56 (–)	0.62 (–)	0.60 (–)	–
V+A	Gimeno-Gómez et al. [16] (2024)	Transformer *	0.59 (–)	0.58 (–)	0.56 (–)	–
T+A+V	Jung et al. [32] (2024)	Transformer+Cross-Attention	–	–	–	0.70
T+A	IICL-DTA(Ours)	Transformer+CL	0.84 (0.86)	0.76 (0.91)	0.80 (0.89)	0.84

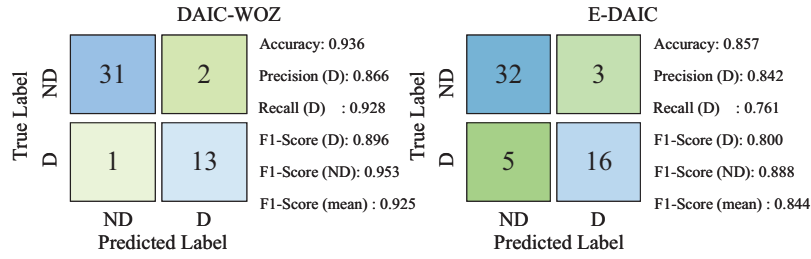


Figure 4: Confusion matrices of the proposed IICL-DTA were generated on the test sets of DAIC-WOZ and E-DAIC. ND stands for nondepression and D stands for depression.

modal and multimodal data inputs. Herein, IICL-DTA is the model designed in this paper. As shown in Table 2, on the DAIC-WOZ dataset, IICL-DTA outperforms all baseline methods, achieving an F1-score of 0.90 for the positive class (D), 0.95 for the negative class (ND), and an average F1-score of 0.93, with an accuracy of 0.94 (detailed metrics and the confusion matrix are presented in Figure 4), reaching the SOTA level. As indicated in the tables, IICL-DTA exhibits a high recall for the positive class (D), suggesting it can detect more cases of depression. In comparison with unimodal methods, IICL-DTA demonstrates competitiveness in terms of average F1-score. Since the dataset used for tri-modal fusion constitutes only a subset, we also compared the performance of other methods using text and audio features, visual and audio features, and text and visual features from this subset as inputs (corresponding to the T+A, V+A, and T+V subsets in Tables 2, 3, and 4). As can be seen, when compared with both bimodal and trimodal methods, our method using only T+A bimodal inputs achieved the highest results. As shown in Table 3, when experimenting on the E-DAIC dataset, which exhibits greater noise issues, the performance of all models is lower than on the DAIC-WOZ dataset. Nevertheless, IICL-DTA still achieved the best performance, yielding an F1-score of 0.80 for the positive class (D), 0.89 for the negative class (ND), and an average F1-score of 0.84. As shown in Table 4, on the Chinese CMDC dataset, our method also achieved the best results, with a precision of 0.94 for the positive class (D), a recall of 1.00 for the positive class (D), and an F1-score of 0.97 for

Table 4: Performance comparison of other methods on CMDC dataset. Results are reported for the depressed (D) class based on 5-fold cross-validation [52]. † Denotes results on the validation set. * Denotes reproduced results from open-source implementations. Best in **bold**.

Modality	Method	Model	Precision	Recall	F1-Score
T	Zou et al. [52] (2023)†	BiLSTM *	0.87	0.90	0.88
	Chen et al. [5] (2024)	BERT	0.82	0.79	0.80
A	Zou et al. [52] (2023)†	BiLSTM *	1.00	0.83	0.91
	Chen et al. [5] (2024)	VGG	0.96	0.89	0.92
V	Zou et al. [52] (2023)†	BiLSTM *	1.00	0.71	0.83
	Chen et al. [5] (2024)	Transformer	0.98	0.87	0.89
T+A	Zou et al. [52] (2023)†	Transformer+CNN *	0.97	0.91	0.94
	Chen et al. [5] (2024)	Transformer	0.96	0.93	0.95
V+A	Zou et al. [52] (2023)†	Transformer+CNN *	0.87	0.87	0.87
	Chen et al. [5] (2024)	Transformer	0.95	0.90	0.91
T+V	Zou et al. [52] (2023)†	Transformer+CNN *	1.00	0.83	0.91
	Chen et al. [5] (2024)	Transformer	0.98	0.87	0.90
T+A+V	Zou et al. [52] (2023)†	Transformer+CNN *	0.97	0.85	0.91
	Chen et al. [5] (2024)	Transformer	0.95	0.93	0.93
T+A	IICL-DTA(Ours)	Transformer+CL	0.94±0.07	1.00±0.00	0.97±0.04

the positive class (D).

Overall, the above results demonstrate the effectiveness of IICL-DTA. Furthermore, we observed similar findings across the three datasets: (1) After integrating two or more modalities, the model performance is significantly superior to unimodal methods. The results indicate the effectiveness of our IICL-DTA for multimodal ADD. (2) Even when using only text and audio (two modalities), our proposed model can still be comparable to trimodal methods, demonstrating the strong performance of our Transformer-based model and contrastive learning approach.

5.2. Ablation Study

We conducted ablation experiments on the DAIC-WOZ and E-DAIC datasets to evaluate the impact of the three key modules in our IICL-DTA model. Table 5 shows the results under different settings. Our model includes three core modules: the Unimodal Feature Decomposition module (Uni), the Intra-Inter Contrastive Learning module (CL), and the Mixup Feature Augmentation module (Mixup). We evaluated the contribution of each module by adding them one at a time.

The following conclusions can be drawn from Table 5 : (1) All proposed modules are effective. Adding any of them improves performance. (2) The Uni module significantly boosts performance on both datasets. This shows that decomposing unimodal features helps the multimodal ADD task and enhances contrastive learning. (3) The CL module helps the model learn differences between depressed and non-depressed samples. This improves its ability to distinguish hard cases and raises overall performance. (4) The Mixup module enriches the input data. With this module, the model achieves its best performance. This confirms the benefit of incorporating mixed features during training.

Table 5: Results of ablation studies on the two datasets, including ablation of three methods.

Method			DAIC-WOZ				E-DAIC			
Uni	CL	Mixup	Accuracy	F1-Score (D)	F1-Score (ND)	F1-Score (mean)	Accuracy	F1-Score (D)	F1-Score (ND)	F1-Score (mean)
✗	✗	✗	0.77	0.70	0.81	0.75	0.70	0.64	0.74	0.69
✓	✗	✗	0.89	0.81	0.93	0.87	0.84	0.76	0.88	0.82
✗	✓	✗	0.85	0.77	0.89	0.83	0.77	0.70	0.81	0.75
✗	✗	✓	0.85	0.72	0.90	0.81	0.77	0.72	0.80	0.76
✓	✓	✓	0.94	0.90	0.95	0.93	0.86	0.80	0.89	0.84

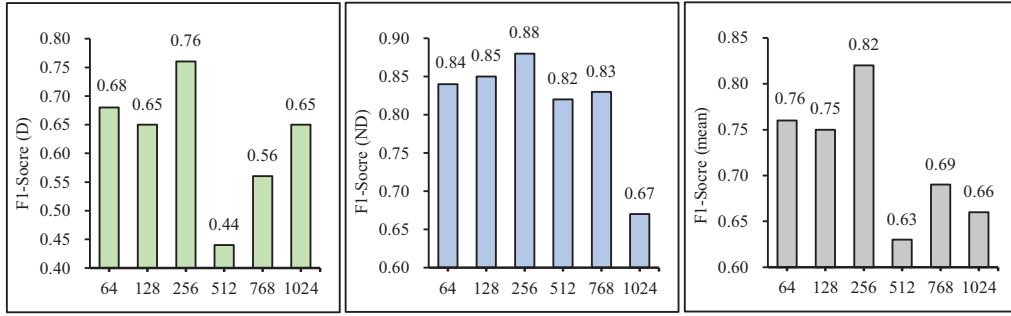


Figure 5: Effect of feature decomposition dimension on the performance of the Uni module on the E-DAIC dataset.

5.3. Parameter Sensitivity Analysis

In this section, we further analyze the parameter sensitivity of the three proposed modules, including the feature decomposition dimension in Uni, the negative sample selection parameter in the CL module, and the mixing coefficient in Mixup.

5.3.1. Feature Decomposition Dimension

To investigate the impact of feature decomposition dimension on model performance, we conducted a sensitivity analysis within the Unimodal Feature Decomposition (Uni) module. As shown in Figure 5, the model achieves the best results when the decomposition dimension is set to $d = 256$, yielding F1-scores of 0.76 and 0.88 for depressed (D) and non-depressed (ND) samples, respectively, and an average F1-score of 0.82.

When d is too small (e.g., $d = 64$ or $d = 128$), the representation capacity is insufficient to capture discriminative features. Conversely, excessively large dimensions (e.g., $d = 512$ and above) may introduce redundancy and increase the risk of overfitting, ultimately degrading performance. Based on these observations, we selected $d = 256$ as the default configuration in all subsequent experiments.

5.3.2. Negative Sample Update Sensitivity

The update ratio controls the proportion of features replaced in the memory bank each iteration, while the update threshold determines the minimum similarity score for a sample to be enqueued. To investigate the impact of these memory update parameters on model performance, we conducted a sensitivity analysis in the memory-based contrastive learning (CL) module. For

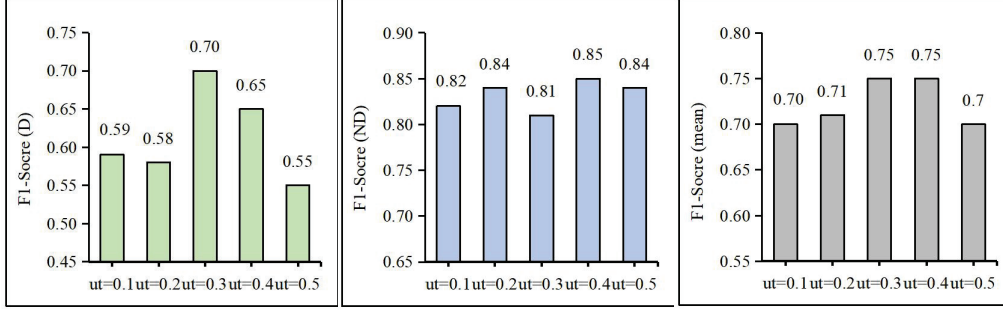
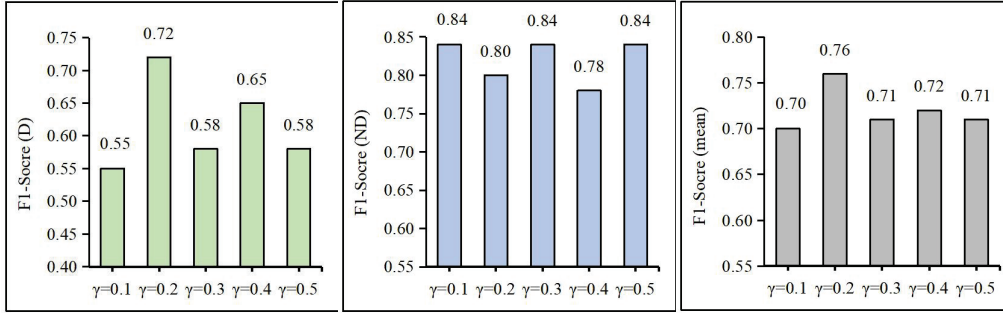


Figure 6: Effect of update ratio and threshold on the performance of the CL module on the E-DAIC Dataset.


 Figure 7: Effect of the Mixup mixing coefficient (γ^{mix}) on the performance of the Mixup module on the E-DAIC dataset.

simplicity, we denote this pair of parameters as ut , where both values are set to the same number (e.g., $ut = 0.1$ means both the update ratio and update threshold are set to 0.1).

As shown in Figure 6, the model achieved the best performance when $ut = 0.3$, reaching an average F1-score of 0.75, with 0.70 for depressed samples and 0.81 for non-depressed samples. Setting it too low (e.g., 0.1 or 0.2) limited the diversity of negative samples in the memory bank, thereby reducing contrastive effectiveness. Conversely, setting it too high (e.g., 0.5) introduced excessive noise, which disrupted stable representation learning and led to performance degradation.

5.3.3. Mixup Mixing Coefficient

We conducted a sensitivity analysis on the key parameter γ^{mix} in the Mixup module to evaluate its impact on model performance. As shown in Figure 7, the model’s performance on the E-DAIC dataset varies significantly with different mixing levels. When $\gamma^{\text{mix}} = 0.2$, the model achieves F1 scores of 0.72 and 0.80 for depressed and non-depressed samples, respectively, with an average F1 of 0.76, representing the best performance. Increasing or decreasing this parameter results in a decline in F1 scores, especially with larger fluctuations in the F1 score for depressed samples, indicating the model is sensitive to the mixing degree. These results suggest that a moderate level of sample mixing ($\gamma^{\text{mix}} = 0.2$) effectively enhances classification ability and optimizes overall performance.

Notably, the performance trend does not follow a monotonic increase with larger mixing ra-

tios. For instance, while $\gamma^{\text{mix}} = 0.2$ yields the best average performance, further increasing the coefficient to $\gamma^{\text{mix}} = 0.3$ or 0.5 leads to degraded results. This observation suggests that excessive interpolation may blur class-specific features and introduce distributional noise, especially for minority classes such as depressed utterances. Conversely, very low values like $\gamma^{\text{mix}} = 0.1$ may not provide sufficient regularization and thus underexploit the benefits of sample mixing. Therefore, these results support the conclusion that there exists an optimal range of mixup strength, and blindly increasing the mixing coefficient does not guarantee performance improvement.

6. Conclusion

In this paper, we propose HICL-DTA, a supervised multimodal contrastive learning framework for automatic depression detection. The framework integrates text (BERT-based) and audio (SpeechFormer++-based) feature extraction modules, alongside a projection module designed to generate synonymous and discriminative features. Through designed intra- and inter-modal contrastive learning loss functions, the model effectively distinguishes similarities from dissimilarities in multimodal features, thereby enhancing its sensitivity to depressive signals. Furthermore, a multimodal auxiliary loss, leveraging Mixup-augmented text and audio features, was developed to mitigate challenges associated with data scarcity. We evaluated our model on three public datasets: English DAIC-WOZ, E-DAIC, and Chinese CMDC. The results demonstrate that HICL-DTA achieved F1-scores (Depressed class) of 0.90, 0.80, and 0.97, respectively, outperforming several state-of-the-art methods and thus validating its effectiveness and generalization capabilities.

In future work, we will focus on advancing depression detection technology through several avenues: (1) Leveraging Large Language Models (LLMs) for data augmentation to alleviate low-resource constraints and improve generalization capabilities; (2) Introducing multi-task learning, incorporating depression severity assessment as an auxiliary task to enhance the precision of symptom recognition; and (3) Optimizing multimodal fusion strategies to reduce computational costs while maintaining performance stability. These endeavors aim to propel the development of automatic depression detection technology, contributing to the creation of more precise and efficient computer-aided diagnostic tools for mental health.

Author Contributions

Dongfang Han: Conceptualization, Investigation, Methodology, Software, Writing - original draft. **Guo-Xing Xiang:** Writing - review & editing. **Yi Liang:** Data curation, Investigation. **Bo Kong:** Data curation, Software. **Yuanyuan Liao:** Writing - review & editing. **Askar Hamdulla:** Supervision, Validation, Writing – review & editing. **Turdi Tohti:** Resources, Supervision, Validation, Writing – review & editing.

Acknowledgements

This work was supported by the Tianshan Talents Cultivation Program – Leading Talents for Scientific and Technological Innovation under Grant No. 2024TSYCLJ0002.

References

- [1] C. Otte, S. M. Gold, B. W. Penninx, C. M. Pariante, A. Etkin, M. Fava, D. C. Mohr, and A. F. Schatzberg, "Major depressive disorder," *Nature Reviews Disease Primers*, vol. 2, no. 1, pp. 1–20, 2016.
- [2] Q. Liu, H. He, J. Yang, X. Feng, F. Zhao, and J. Lyu, "Changes in the global burden of depression from 1990 to 2017: Findings from the global burden of disease study," *Journal of Psychiatric Research*, vol. 126, pp. 134–140, 2020.
- [3] K. Yuan, Y.-B. Zheng, Y.-J. Wang, Y.-K. Sun, Y.-M. Gong, Y.-T. Huang, X. Chen, X.-X. Liu, Y. Zhong, S.-Z. Su *et al.*, "A systematic review and meta-analysis on prevalence of and risk factors associated with depression, anxiety and insomnia in infectious diseases, including covid-19: a call to action," *Molecular Psychiatry*, vol. 27, no. 8, pp. 3214–3222, 2022.
- [4] M. Du, S. Liu, T. Wang, W. Zhang, Y. Ke, L. Chen, and D. Ming, "Depression recognition using a proposed speech chain model fusing speech production and perception features," *Journal of Affective Disorders*, vol. 323, pp. 299–308, 2023.
- [5] 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," *Information Fusion*, vol. 102, p. 102017, 2024.
- [6] Z. Ding, J. Chen, B.-L. Zhong, C.-L. Liu, and Z.-T. Liu, "Emotional stimulated speech-based assisted early diagnosis of depressive disorders using personality-enhanced deep learning," *Journal of Affective Disorders*, vol. 376, no. 1, pp. 177–188, 2025.
- [7] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. Williams, J. T. Berry, and A. H. Mokdad, "The phq-8 as a measure of current depression in the general population," *Journal of Affective Disorders*, vol. 114, no. 1–3, pp. 163–173, 2009.
- [8] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. D. la Torre, "Detecting depression from facial actions and vocal prosody," in *Proc. Humaine Assoc. Conf. Affective Comput. Intell. Interact. (ACII)*, 2009, pp. 1–7.
- [9] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency, "Automatic nonverbal behavior indicators of depression and ptsd: Exploring gender differences," in *Proc. Humaine Assoc. Conf. Affective Comput. Intell. Interact. (ACII)*, 2013, pp. 147–152.
- [10] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency, "Automatic behavior descriptors for psychological disorder analysis," in *Proc. Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–8.
- [11] Z. Yu, S. Scherer, D. Devault, J. Gratch, G. Stratou, L.-P. Morency, and J. Cassell, "Multi-modal prediction of psychological disorders: Learning verbal and nonverbal commonalities in adjacency pairs," in *Proc. Workshop on the Semantics and Pragmatics of Dialogue (Sem-Dial)*, 2013, pp. 160–169.
- [12] A. Haque, M. Guo, A. S. Miner, and L. Fei-Fei, "Measuring depression symptom severity from spoken language and 3d facial expressions," *arXiv preprint arXiv:1811.08592*, 2018.

- [13] T. A. Hanai, M. M. Ghassemi, and J. R. Glass, “Detecting depression with audio/text sequence modeling of interviews,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, 2018, pp. 1716–1720.
- [14] W. Chen, X. Xing, X. Xu, J. Pang, and L. Du, “Speechformer++: A hierarchical efficient framework for paralinguistic speech processing,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 775–788, 2023.
- [15] S. Yang, L. Cui, L. Wang, T. Wang, and J. You, “Enhancing multimodal depression diagnosis through representation learning and knowledge transfer,” *Heliyon*, vol. 10, no. 4, p. e25959, 2024.
- [16] D. Gimeno-Gómez, A.-M. Bucur, A. Cosma, C.-D. Martínez-Hinarejos, and P. Rosso, “Reading between the frames: Multi-modal depression detection in videos from non-verbal cues,” in *Proc. European Conf. Inf. Retr. (ECIR)*, 2024, pp. 191–209.
- [17] M. Fang, S. Peng, Y. Liang, C.-C. Hung, and S. Liu, “A multimodal fusion model with multi-level attention mechanism for depression detection,” *Biomedical Signal Processing and Control*.
- [18] H. Ma, J. Wang, H. Lin, B. Zhang, Y. Zhang, and B. Xu, “A transformer-based model with self-distillation for multimodal emotion recognition in conversations,” *IEEE Trans. Multimedia*, vol. 26, pp. 776–788, 2023.
- [19] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, “Avec 2016: Depression, mood, and emotion recognition workshop and challenge,” in *Proc. Int. Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2016, pp. 3–10.
- [20] J. R. Williamson, E. Godoy, M. Cha, A. Schwarzentruher, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, and T. F. Quatieri, “Detecting depression using vocal, facial and semantic communication cues,” in *Proc. Int. Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2016, pp. 11–18.
- [21] Y. Gong and C. Poellabauer, “Topic modeling based multi-modal depression detection,” in *Proc. Int. Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2017, pp. 69–76.
- [22] Z. Huang, J. Epps, D. Joachim, B. Stasak, J. R. Williamson, and T. F. Quatieri, “Domain adaptation for enhancing speech-based depression detection in natural environmental conditions using dilated cnns,” in *Proc. Annual Conf. of the Int. Speech Commun. Assoc. (Interspeech)*, 2020, pp. 4561–4565.
- [23] A. Vázquez-Romero and A. Gallardo-Antolín, “Automatic detection of depression in speech using ensemble convolutional neural networks,” *Entropy*, vol. 22, no. 6, p. 688, 2020.
- [24] S. Sardari, B. Nakisa, M. N. Rastgoo, and P. Eklund, “Audio based depression detection using convolutional autoencoder,” *Expert Systems with Applications*, vol. 189, p. 116076, 2022.

- [25] H. Tian, Z. Zhu, and X. Jing, “Deep learning for depression recognition from speech,” *Mobile Networks and Applications*, pp. 1–16, 2023.
- [26] A. Salekin, J. W. Eberle, J. J. Glenn, B. A. Teachman, and J. A. Stankovic, “A weakly supervised learning framework for detecting social anxiety and depression,” *ACM Transactions on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 2, pp. 1–26, 2018.
- [27] Y. Zhao, Y. Xie, R. Liang, L. Zhang, L. Zhao, and C. Liu, “Detecting depression from speech through an attentive lstm network,” *IEICE Trans. Inf. Syst.*, vol. 104, no. 11, pp. 2019–2023, 2021.
- [28] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, “Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech,” *Biomedical Signal Processing and Control*, vol. 71, p. 103107, 2022.
- [29] K. Mao, W. Zhang, D. B. Wang, A. Li, R. Jiao, Y. Zhu, B. Wu, T. Zheng, L. Qian, and W. Lyu, “Prediction of depression severity based on the prosodic and semantic features with bidirectional lstm and time distributed cnn,” *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2251–2265, 2022.
- [30] X. Zhang, X. Zhang, W. Chen, C. Li, and C. Yu, “Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments,” *Scientific Reports*, vol. 14, no. 1, p. 9543, 2024.
- [31] V. Ravi, J. Wang, J. Flint, and A. Alwan, “Enhancing accuracy and privacy in speech-based depression detection through speaker disentanglement,” *Computer Speech and Language*, vol. 86, p. 101605, 2024.
- [32] J. Jung, C. Kang, J. Yoon, S. Kim, and J. Han, “Hique: Hierarchical question embedding network for multimodal depression detection,” in *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, 2024, pp. 1049–1059.
- [33] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2006, pp. 1735–1742.
- [34] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.
- [35] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.
- [36] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proc. Int. Conf. on Machine Learning (ICML)*, 2020, pp. 1597–1607.
- [37] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9729–9738.

- [38] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. Guo, M. G. Azar *et al.*, “Bootstrap your own latent—a new approach to self-supervised learning,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 271–21 284, 2020.
- [39] A. Saeed, D. Grangier, and N. Zeghidour, “Contrastive learning of general-purpose audio representations,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3875–3879.
- [40] R. Qian, T. Meng, B. Gong, M.-H. Yang, H. Wang, S. Belongie, and Y. Cui, “Spatiotemporal contrastive video representation learning,” in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6964–6974.
- [41] Z. Qi, S. Wang, C. Su, L. Su, Q. Huang, and Q. Tian, “Self-regulated learning for egocentric video activity anticipation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6715–6730, 2021.
- [42] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, “Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 206–24 221, 2021.
- [43] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, “Supervised contrastive learning,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 661–18 673, 2020.
- [44] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.
- [45] Z. Wu, S. Wang, J. Gu, M. Khabsa, F. Sun, and H. Ma, “Clear: Contrastive learning for sentence representation,” *arXiv preprint arXiv:2012.15466*, 2020.
- [46] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, “Unsupervised data augmentation for consistency training,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 6256–6268, 2020.
- [47] A. Kumar, S. Bhattamishra, M. Bhandari, and P. Talukdar, “Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation,” in *Proc. Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019, pp. 3609–3619.
- [48] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance discrimination,” in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3733–3742.
- [49] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “Mixup: Beyond empirical risk minimization,” *arXiv preprint arXiv:1710.09412*, 2017.
- [50] 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 *Proc. Int. Conf. on Language Resources and Evaluation (LREC)*, vol. 14, 2014, pp. 3123–3128.

- [51] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet *et al.*, “Simsensei kiosk: A virtual human interviewer for health-care decision support,” in *Proc. Int. Conf. on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2014, pp. 1061–1068.
- [52] B. Zou, J. Han, Y. Wang, R. Liu, S. Zhao, L. Feng, X. Lyu, and H. Ma, “Semi-structural interview-based chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders,” *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 2823–2838, 2022.
- [53] L. Bottou, “Large-scale machine learning with stochastic gradient descent,” in *Proc. Int. Conf. on Computational Statistics (COMPSTAT)*, 2010, pp. 177–186.
- [54] P.-C. Wei, K. Peng, A. Roitberg, K. Yang, J. Zhang, and R. Stiefelhagen, “Multi-modal depression estimation based on sub-attentional fusion,” in *Proc. Eur. Conf. on Computer Vision (ECCV)*, 2022, pp. 623–639.
- [55] B. Sun, Y. Zhang, J. He, L. Yu, Q. Xu, D. Li, and Z. Wang, “A random forest regression method with selected-text feature for depression assessment,” in *Proc. Annual Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2017, pp. 61–68.
- [56] H. Solieman and E. A. Pustozarov, “The detection of depression using multimodal models based on text and voice quality features,” in *Proc. IEEE Conf. of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*, 2021, pp. 1843–1848.
- [57] Z. Huang, J. Epps, and D. Joachim, “Exploiting vocal tract coordination using dilated cnns for depression detection in naturalistic environments,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6549–6553.
- [58] G. Ioannides, A. Kieback, A. Chadha, and A. Elkins, “Density adaptive attention-based speech network: Enhancing feature understanding for mental health disorders,” *arXiv preprint arXiv:2409.00391*, 2024.
- [59] M. Li, X. Sun, and M. Wang, “Detecting depression with heterogeneous graph neural network in clinical interview transcript,” *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 1, pp. 1315–1324, 2023.
- [60] W. Zhang, K. Mao, and J. Chen, “A multimodal approach for detection and assessment of depression using text, audio and video,” *Phenomics*, vol. 4, no. 3, pp. 234–249, 2024.
- [61] E. Villatoro-Tello, G. R. de-la Rosa, D. Gática-Pérez, M. Magimai.-Doss, and H. Jiménez-Salazar, “Approximating the mental lexicon from clinical interviews as a support tool for depression detection,” in *Proc. Int. Conf. on Multimodal Interaction (ICMI)*, 2021, pp. 557–566.
- [62] J. Cheong, S. Kalkan, and H. Gunes, “Fairrefuse: Referee-guided fusion for multimodal causal fairness in depression detection,” in *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 2024, pp. 7224–7232.