2026 IEEE International Conference on Pervasive Computing and Communications (PerCom) | 979-8-3315-7613-4/26/$31.00 ©2026 IEEE | DOI: 10.1109/PerCom67906.2026.11524348

MentalCare: Contrastive Disentanglement and Positive Transfer for Psychiatric Disorder Detection with a Wristband–Smartphone System

Yufei Zhang†, Wenting Kuang†, Zite Huang†, Changhe Fan‡, and Yongpan Zou†
† College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
‡ Department of Psychiatry, Affiliated Guangdong Second Provincial General Hospital of Jinan University, Guangzhou, China
yfyaya@foxmail.com; {kuangwenting2023, 2023150190}@email.szu.edu.cn; changhefan@yahoo.com; yongpan@szu.edu.cn

***Abstract*—Low-cost wrist-worn sensing for scalable, low-burden assessment of mental-health symptoms is a central topic in mobile health. In this paper, we propose *MentalCare*, a multi-class psychiatric disorder (PD) detection system that integrates ubiquitous multi-source sensors, including Pulse sensor, LMT70, Grove GSR, and Max30102. These sensors capture physiological signals such as photoplethysmography (PPG), skin temperature (SKT), galvanic skin response (GSR), and blood oxygen (OXY). To address physiological heterogeneity among individuals, we design a contrastive disentanglement with a positive transfer framework (CDPT), which consists of a stacked convolutional neural network module (SCNN), a cross-modal attention interaction module (CMAI), a subject-level contrastive disentanglement module (SCD), and an adaptive feature selector module (AFS). In particular, the SCD module employs supervised contrastive learning to promote the separation of heterogeneous features among different subjects in the latent subspace. Furthermore, the proposed AFS module selects features from the abundant source-domain heterogeneous representations that are most similar to the target domain, thereby facilitating feature alignment and enhancing detection performance. We collect a multi-class PD dataset comprising 74 participants diagnosed with depression, bipolar disorder, anxiety, or verified as healthy controls. The experiment result demonstrates that CDPT achieves 83.96% overall accuracy under the leave-one-subject-out cross-validation protocol, significantly outperforming existing state-of-the-art methods. At the system level, we further validate the effectiveness and rationality of the proposed system when deployed on edge devices.**

***Index Terms*—Ubiquitous Sensing, Mobile Health, Disentanglement Learning, Contrastive Learning, Positive Transfer.**

I. Introduction

Psychiatric Disorders (PDs) emerge as a critical global public health issue, with depression and anxiety experiencing a 25% increase in prevalence in the first year of the COVID-19 pandemic [1]. These disorders can lead to emotional instability, cognitive impairment, and functional decline, posing long-term risks to both individuals and society [2]. Current diagnostic practices remain dependent on clinical interviews and standardized questionnaires, which continue to serve as the mainstream approach in psychiatric assessment [3]. However, as reported by World Health Organization (WHO), the imbalance between the limited number of clinicians and the large patient population creates substantial treatment gaps that restrict access to care and consequently delay diagnosis [1].

With the rapid advancement of artificial intelligence and ubiquitous sensing technologies, new opportunities have emerged for PDs detection. Recent studies demonstrate that PDs detection can leverage a wide range of modalities, including images [4], [5], speech [6], [7], text [8], [9], and beyond [10], [11]. However, these modalities are easily influenced by the social mask, where patients intentionally conceal or modify their expressions and verbal content to obscure their actual emotional states [12], [13]. Such behavior, often driven by shame, fear of stigmatization, or social expectations, leads them to conceal or modify their symptom descriptions, thereby reducing diagnostic reliability [14]. In contrast, physiological signals directly capture variations in physiological responses regulated by the autonomic nervous system or peripheral nervous system, thereby offering a more objective and reliable representation [15].

Nowadays, physiological signals are recognized as a critical modality for enabling intelligent PDs detection. A lot of studies have examined various physiological signals, such as electroencephalography (EEG) [16], electrocardiography (ECG) [17], and electrooculography (EOG) [18], etc. While these signals provide high temporal resolution and clinically precise measurements, their use is limited by high costs, proprietary equipment, and the need for professional operation. In contrast, the ubiquitous physiological signals, including photoplethysmography (PPG), respiration (RESP), skin temperature (SKT), and galvanic skin response (GSR), are typically collected using wearable sensors. These signals are characterized by their affordability, portability, and ability to support continuous and unobtrusive monitoring in daily life, making them particularly suitable for scalable and long-term mental health assessment and intervention [19]. However, leveraging these signals effectively remains non-trivial. Beyond challenges such as noise and low signal quality, ***a more fundamental issue arises from substantial inter-individual variability***. In PDs monitoring, individuals exhibit substantial differences in neural dynamics and physiological response patterns. Even under similar clinical symptoms, physiological signals may reveal highly personalized characteristics [19]. Such heterogeneity introduces significant domain shifts, making models difficult to generalize reliably to unseen subjects.

Prior studies have employed domain adaptation (DA) strategies, Zhang et al. [19] apply unsupervised domain adaptation for depression recognition, as well as few-shot learning for monitoring depressive episodes. Chen et al. [20] propose GNN-SDA, a flexible EEG-based semi-supervised DA framework for depression detection. These approaches effectively capture subject-specific characteristics and improve cross-domain adaptability [13]. However, DA typically depends on target-specific information to align distributions, which compromises subject adaptability and is often impractical for real-world settings (i.e., subject-independent detection systems) where such information is unavailable [21].

In this paper, we propose *MentalCare*, a ubiquitous system consisting of a smartphone application and a wrist-worn device for the early detection of depression, anxiety, and bipolar disorder. The wristband continuously collects the subject's multimodal physiological signals, including PPG, SKT, GSR, and blood oxygen (OXY). To address this core challenge of the substantial inter-individual variability, we propose a contrastive disentanglement with a positive transfer framework (CDPT), which consists of a stacked convolutional neural network module (SCNN), a cross-modal attention interaction module (CMAI), a subject-level contrastive disentanglement module (SCD), and an adaptive feature selector module (AFS). Among these components, the subject feature selector and contrastive disentanglement network play a pivotal role in enhancing cross-subject performance. Recent advanced disentanglement learning methods focus on domain-invariant features from source domains (i.e., training subjects), based on the assumption that these features generalize best to unseen target domains (i.e., new subjects) [18], [22], [23]. However, in cross-domain scenarios, certain source domains may contain complementary information relevant to the target domain. Leveraging this information can improve prediction accuracy, a phenomenon referred to as positive transfer [24]. In CDPT, the AFS module selectively leverages source subjects most similar to the target domain, combining shared features with useful domain-specific information to enhance target-domain performance. Furthermore, we employ contrastive learning to encourage the separation of heterogeneous features across subjects during the disentanglement process.

Overall, our main contributions are summarized as follows:

- We develop *MentalCare*, a multi-class mental disorder analysis system with custom hardware, software, and diagnosis algorithms, achieving 83.96% cross-subject accuracy on our collected dataset.
- We propose a disentanglement representation framework termed as CDPT in multimodal physiological sensing for the detection of PDs. Furthermore, we introduce two noteworthy improvements.
- We propose a SCD network that integrates contrastive learning into the disentanglement process, thereby improving the separation of specific features across subjects.
- We design an AFS module that effectively leverages specific representations from the source domain to enhance mental disorder diagnosis in the target domain.

II. Related Work

A. Ubiquitous Physiological Sensing for PDs Detection

Pervasive sensing technologies are increasingly explored for reliable PDs detection based on physiological signal. For instance, Ouzar et al. [25] and Shui et al. [12] leverage wearable devices to collect physiological and behavioral data, employing multiple machine learning approaches to detect depression. Hassantabar et al. [26] introduce MHDeep, a framework that integrates wearable sensor data with artificial neural networks to classify mental health disorders. Ahmed et al. [27] propose a multimodal framework that combines convolutional neural networks, and random forests to assess affective states and disorders. However, existing approaches either employ traditional machine learning or simple neural network techniques [25], [27], which are limited in their ability to uncover the complex distributional characteristics across domains, or rely solely on a single physiological signal [28], [29], thereby missing the complementary information and enhanced discriminative power that can be obtained from the integration of multiple peripheral physiological signals. Even if a few studies further enhance the diagnostic effectiveness of PDs detection through carefully designed deep learning methods [30], [31]. However, these studies also lack consideration for individual differences, making it difficult to apply models or systems to new subjects.

B. Disentanglement Representation Learning

Recent work in affective computing has explored disentanglement methods for multimodal signals. Early studies adopted traditional disentanglement frameworks. For example, Hazarika propose MISA [32], which imposes orthogonality constraints and Central Moment Discrepancy (CMD) to guide the separation of factors. Along this line, Yang et al. [33] refine the consistency objective and replace CMD with an adversarial network for more stable disentanglement. Jia et al. [18] propose a multi-level disentanglement model for multimodal physiological sensing in emotion recognition, addressing signal heterogeneity and cross-modal complementarity. Pan et al. [34] propose the Dis^{2}DR framework for depression recognition with missing modalities. It uses the teacher model trains on full multimodal data, while the student model applies randomized masking to simulate modality absence.

Relative to prior works, we present two advances over conventional disentanglement learning. First, we implement cross-domain forward transfer via an AFS module that, at inference time, identifies source-domain heterogeneous features most similar to the target domain and leverages them to improve target recognition. Second, we incorporate contrastive learning to further constrain inter-subject heterogeneity.

III. Methodology

A. Physiological Signal Processing

First, we conduct exploratory spectral analysis with the Fast Fourier Transform (FFT) to characterize each signal's

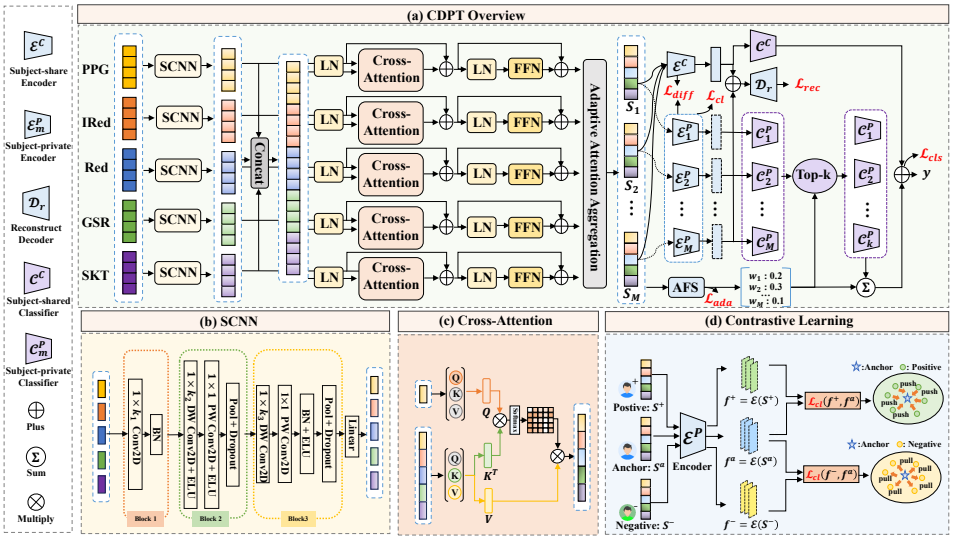


Fig. 1: The structure of CDPT framework.

TABLE I: Preprocessing steps for physiological signals.

Physiological Signal	FFT	Butterworth Bandpass Filter (Hz)	Min-max Normalization
HRV	✓	$0.05 \sim 4$	✓
OXY (IRed and Red)	✓	$0.1 \sim 4$	✓
SKT	✓	$49 \sim 51$	✓
GSR	✓	$0 \sim 0.3$	✓

frequency content, identify dominant bands and potential interference. Then, a Butterworth band-pass filter is applied to retain the effective frequency range while suppressing noise such as power-line interference and frequency-specific motion artifacts. Due to physiological signals vary inherently, the processing parameters differ for each signal type, as detailed in Table I. Note that, the original OXY saturation signal obtained through the blood oxygen sensor consists of two channel: infrared light (IR) signal and red light (Red) signal.

After that, we use the following normalize method to restrict the input value in the ranage of $[-1, 1]$. Finally, we use sliding processing with a window size of 1 and a stride size of 1:

$$X_{\text{norm}} = 2\frac{X - X_{\min}}{X_{\max} - X_{\min}} - 1. \tag{1}$$

B. CDPT Framework

As shown in Fig. 1, CDPT framework consists of several key components that work sequentially to process and integrate physiological signals. We introduce the components sequentially according to the pipeline structure of the framework.

1) Modality Encoding: To effectively preserve the modality-specific characteristics of physiological signals, we adopt a channel-independent encoding strategy [35]. The modality feature encoding module is composed of a set of parallel encoders, where each encoder independently transforms the raw signal of one modality into an aligned feature embedding. Each SCNN module contains three blocks that progressively enhance temporal representations.

Let the input of the i-th modality be denoted as $X_i \in \mathbb{R}^{1\times C\times T}$, where 1 is a expanded convolution dimension, C is the number of channels ($C = 1$), and T is the sequence length within a window.

Block 1 (Short-term convolution). The first block applies a 1D convolution to capture short-term temporal dynamics while preserving the spatial dimension. $\text{Conv}_{1\times k_1}$ denotes a convolution along the temporal axis with kernel size k_1, and BN denotes batch normalization.

Block 2 (Depthwise separable convolution). The second block introduces depthwise separable convolution to enhance feature diversity. The depthwise convolution $\text{Conv}^{\text{dw}}_{1\times k_2}$ with kernel size k_2 first extracts temporal patterns independently from each channel. The subsequent pointwise convolution $\text{Conv}^{\text{pw}}_{1\times 1}$ then integrates channel-specific responses, enabling information exchange across channels. ELU activation is a nonlinear transformation, followed by average pooling for downsampling. Dropout is applied to improve generalization.

Block 3 (Long-term separable convolution). The third block extends this design by enlarging the temporal receptive field through a depthwise convolution $\text{Conv}^{\text{dw}}_{1\times k_3}$ with kernel size k_3, thereby capturing long-range dependencies. The subsequent pointwise convolution $\text{Conv}^{\text{pw}}_{1\times 1}$ consolidates these extended features into high-level representations that preserve discriminative temporal dynamics. Batch normalization, ELU activation, average pooling, and dropout are sequentially applied as in the previous blocks.

Finally, the output of the third block is flattened and projected into a shared latent space via a linear layer, ensuring dimensional consistency across modalities. The encoding module converts each raw physiological signal into a modality-aligned embedding $\{F_1, F_2, \ldots, F_N\}$ that captures multi-scale temporal dynamics, which are then integrated by the cross-modal attention fusion module.

2) Cross-Modal Attention Interaction Module: After obtaining modality-specific embeddings, we employ a CMAI

module to synthesize these physiological signals. This module is designed as a two-phase process that first facilitates explicit interactions among modalities and then adaptively aggregates the enhanced features by assigning learnable attention weights, enabling the model to emphasize the most informative modalities for the downstream task.

Cross-modal interaction phase. The encoded features from different modalities are denoted as $\{F_1, F_2, \ldots, F_N\}$, where each $F_i \in \mathbb{R}^H$ corresponds to the representation of the i-th modality, H is the feature dimension, N is the number of modalities. These modality-specific embeddings are concatenated to construct a global context representation, which serves as the common source of information for cross-modal exchange:

$$F_g = \text{Concat}(F_1, F_2, \ldots, F_N), \tag{2}$$

where $F_g \in \mathbb{R}^{N \cdot H}$ denote the combined global features.

To enable interaction, the global representation F_g is projected into key and value spaces, while each modality feature F_i is transformed into a modality-specific query [36]

$$K = W_K F_g, \quad V = W_V F_g, \quad Q_i = W_Q^{(i)} F_i, \tag{3}$$

where $K \in \mathbb{R}^{N \times H}$ and $V \in \mathbb{R}^{N \times H}$ are the global key and value matrices, and $Q_i \in \mathbb{R}^{1 \times H}$ is the query vector. The projections use $W_K, W_V \in \mathbb{R}^{(N \cdot H) \times (N \cdot H)}$ and $W_Q^{(i)} \in \mathbb{R}^{H \times H}$.

Cross-modal attention is then performed by matching the query from modality i with the keys of all modalities, and using the resulting attention weights to compute a weighted combination of the value:

$$\text{Attn}_i = \text{Softmax}\big(Q_i K^\top\big)\, V, \tag{4}$$

The interaction outcome is added back to the original modality feature through a residual connection and further refined by a position-wise feed-forward network (FFN), which consists of two linear layers with a nonlinear activation in between:

$$\hat{F}_i = \text{FFN}(\text{LN}(F_i + \text{Attn}_i)) + (F_i + \text{Attn}_i). \tag{5}$$

Here, LN denotes Layer Normalization, which normalizes the hidden features across the feature dimension for each sample. Meanwhile, the FFN is defined as

$$\text{FFN}(x) = W_2\, \sigma(W_1 x + b_1) + b_2, \tag{6}$$

where W_1 and $W_2 \in \mathbb{R}^{H \times H}$ denote learnable weight matrices, b_1 and $b_2 \in \mathbb{R}^H$ represent learnable bias vectors, and $\sigma(\cdot)$ is ReLU nonlinear activation.

Adaptive aggregation phase: In this phase, the enhanced features $\{\hat{F}_1, \hat{F}_2, \ldots, \hat{F}_N\}$ are combined into a single representation. Each μ_i is first scored through a shared attention mechanism that evaluates its relative importance:

$$\mu_i = q^\top \tanh\Big(W_a^{(i)} \hat{F}_i + b_a^{(i)}\Big), \tag{7}$$

where $q \in \mathbb{R}^H$ is a learnable attention vector. The scores are normalized across modalities to obtain attention weights:

$$\psi_i = \frac{\exp(\mu_i)}{\sum_{j=1}^{M} \exp(\mu_j)}. \tag{8}$$

Finally, the fused multimodal physiological representation $S \in \mathbb{R}^H$ is computed as a weighted combination of the modality-enhanced features:

$$S = \sum_{i=1}^{M} \psi_i \cdot \hat{F}_i. \tag{9}$$

This two-phase design allows the CMAI module to explicitly model inter-modal dependencies while adaptively emphasizing the most informative modalities, thereby yielding a compact and discriminative representation.

C. Subject-level Contrastive Disentanglement Module

1) Disentangled Encoders: An efficient disentangled representation is defined as one that separates the distinct, independent, and informative generative factors underlying the data. To achieve this, we design the SCD module to separate consistent and specific features among subjects from complex multi-modal fusion representations, thereby facilitating cross-subject PDs detection. Specifically, we implement a set of private encoders $\{\mathcal{E}_i^P\}_{i=1}^M$ as well as a shared encoder $\mathcal{E}^C$. On the one hand, the shared encoder $\mathcal{E}^C$ maps the fused features S of all subjects into a common subspace while extracting cross-domain consistent features $\widehat{S}$. On the other hand, each private encoder $\mathcal{E}_i^P$ projects the fused features of its corresponding subject i into an independent subspace to capture subject-specific (or heterogeneous) feature $\widetilde{S}_i$.

$$\widehat{S} = \mathcal{E}^C(S, \theta^c), \tag{10}$$

$$\widetilde{S}_i = \mathcal{E}_i^P\left(S_i, \theta_i^P\right) (i = 1, 2, \ldots, M), \tag{11}$$

where $\widehat{y}_i \in \{1, ..., M\}$ is the subject identity label of the i-th sample, θ^C and $\theta_{\widehat{y}_i}^P$ denote the parameters of the shared and private encoders belongs to $\widehat{y}_i$-th subject, respectively. In implementation, both the shared encoder and the private encoders are implemented with two fully connected layers.

2) Disparity Constraint: To encourage the disentangled process as well as ensure subject-invariant features and subject-consistent features remain mutually independent, we introduce an orthogonality constraint to maximize the difference between the two feature:

$$\mathcal{L}_{diff} = \left\| \tilde{S}^\top \hat{S} \right\|_F^2, \tag{12}$$

where $\|\cdot\|_F^2$ denotes the Frobenius norm. This constraint minimizes overlap between invariant and specific representations, encouraging each to focus on distinct factors of variation.

3) Reconstruct Supervision: During the feature disentanglement process, essential information may be lost when separating private and shared components. To address this, we employ a reconstruction decoder $\mathcal{D}_r$ to ensure that the combination of private and shared features preserves the original and useful information.

$$\mathcal{L}_{rec} = \frac{\left\| S - \mathcal{D}_r(\hat{S} + \tilde{S}) \right\|_2^2}{d}, \tag{13}$$

where d is the dimension of h and $\|\cdot\|_2^2$ denotes the squared $\mathcal{L}_2$ norm. The reconstruction decoder is implemented as a linear layer that maps the concatenated features back to the hidden space of the same dimension. Reconstruct supervision encourages the model to disentangle meaningful subject-specific and cross-subject components while maintaining the fidelity of the original representations.

4) Supervised Contrastive Learning: Considering that the orthogonal difference loss acts on both shared and private features, the private feature representations of different subjects may still exhibit certain entanglement. To further promote robust disentangled learning, we incorporate contrastive learning [37] into the disentanglement of private features. Specifically, we adopt an instance-level supervised contrastive learning strategy. Formally, given a batch of feature representations $\{\tilde{S}_i\}_{i=1}^{N}$ extracted from the private encoders, each feature $\tilde{S}_i$ is associated with a subject label $\widehat{y}_i$. For a given anchor $\tilde{S}_i$, we define the set of positive samples as $P(i) = \{\tilde{S}_j \,|\, \widehat{y}_j = \widehat{y}_i,\, j \neq i\}$, namely the features originating from the same subject but different instances. The remaining samples in the batch are regarded as negative samples, denoted as $N(i) = \{\tilde{S}_j \,|\, \widehat{y}_j \neq \widehat{y}_i\}$. The supervised contrastive loss is defined as follows:

$$\mathcal{L}_{cl} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\sum_{j:\widehat{y}_j=\widehat{y}_i, j\neq i}\exp\left(\frac{\mathrm{sim}(\tilde{S}_i,\tilde{S}_j)}{\tau}\right)}{\sum_{k\neq i}\exp\left(\frac{\mathrm{sim}(\tilde{S}_i,\tilde{S}_k)}{\tau}\right)}, \quad (14)$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes the cosine similarity function and τ is a temperature parameter. This contrastive objective makes private features from the same subject cluster tightly while pushing those from different subjects apart, improving diagnostic robustness.

D. Positive Transfer for Cross-subject Augmentation

In this part, we design an AFS module to integrate information from source subjects that are most similar to the target subject. This mechanism, termed positive transfer, enhances the prediction process and facilitates the exploration of physiological heterogeneity in the target domain.

During training, for the source subject i, AFS employs a linear layer and a Softmax function with the fused features S as input. This selector return a set of weighted vector $P_i \in \mathbb{R}^M$:

$$P_i = \{p_{i,1},\ p_{i,2}, ...,\ p_{i,M}\}\ s.t.\ \ 0 \leq p_{i,j} \leq 1,\ \sum_{j=1}^{M} p_{i,j} = 1. \quad (15)$$

Here, the selector outputs a probability vector P_i over all source domains. $p_{i,j}$ denotes the similarity probability between subject feature i and feature j, which lies within the interval $[0, 1]$ and satisfies that the sum of all elements in the set p_i equals 1. It should be noted that, since CDPT is designed for DG rather than DA method, the private encoder of the target domain does not exist. Therefore, the input to the AFS module is fused feature S rather than specific feature $\hat{S}$.

To achieve training of the adaptive feature selector, we define the following loss function

$$\mathcal{L}_{ada} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{M}\hat{y}_{i,j}\log(p_{i,j}), \quad (16)$$

where $\hat{y}_i$ denotes a one-hot ground-truth vector, indicating the correct source subject. The adaptive selector is trained to maximize the likelihood of the correct source subject.

E. Weighted Fusion Classification

After obtaining the disentangled private and shared features, we design a weighted fusion strategy with domain classifiers to perform PDs prediction for each subject. Specifically, each disentanglement encoder is followed by an independent classifier, namely the domain-shared classifier $\mathcal{C}^G$ and the domain-specific classifier $\mathcal{C}^C$. In conjunction with the weight derived by AFS module, we extract the Top-K weights from the output vector of selector, and use them to weight the prediction results of the private classifiers.

$$y_i^P = \sum_{j=1}^{K} p_{i,j}\mathcal{C}_j^S(\tilde{S}_i + \hat{S}_i; \theta_j^{\mathcal{C}^S}), \quad (17)$$

$$y^C = \mathcal{C}^C(\tilde{S}; \theta^{\mathcal{C}^C}), \quad (18)$$

where y_i^P and y^C denote the results of i-th subject's private classifier and the shared classifier, respectively.

To obtain the target domain prediction, we integrate the outputs of the domain-specific and domain-invariant classifiers. A trade-off parameter δ is introduced to balance the contributions of these two classifiers. Accordingly, the final probability distribution for a sample X_i from the i-th source subject can be expressed as follows:

$$y_i = (1-\delta)y_i^P + \delta y^C. \quad (19)$$

The classification process can be constrained by

$$\mathcal{L}_{cls} = -\frac{1}{M}\sum_{i=1}^{M} y_i log(\tilde{y}_i), \quad (20)$$

where $\tilde{y}_i$ and y_i denote the ground-truth and predicted label of i-th subject.

F. Loss constraint of CDPT

The overall objective of the proposed CDPT contains several kinds of constraints as follows

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1\mathcal{L}_{diff} + \lambda_2\mathcal{L}_{rec} + \lambda_3\mathcal{L}_{cl} + \lambda_4\mathcal{L}_{ada}, \quad (21)$$

where λ_1 to λ_4 serve as trade-off coefficients that regulate the balance among the constraints.

IV. IMPLEMENTATION AND EXPERIMENTS

A. Hardware–software System

Fig. 2 illustrates the integrated hardware–software design and experimental setup. We develop a wearable device that integrates a custom circuit board, a central microcontroller, a Bluetooth BLE 5.0 module, and multiple physiological sensors, including PulseSensor, Max30102, Grove GSR, and LMT70. These sensors collect PPG, OXY, GSR, and SKT signals, which the microcontroller processes via GPIO interfaces and transmits to a software terminal through Bluetooth. The device measures $52 \times 45 \times 30$ mm and weighs 111.8 g while being designed for continuous wear and incorporating a SY8089 DC-DC buck converter with a TP4054 charging circuit that supports up to 500 mA to ensure stable power management. On the software side, real-time physiological data are received, analyzed, and used for biosignal assessment and psychiatric screening with our model, while the system also manages sessions, records logs, and visualizes results on a smartphone. The application interface is deployed within a mobile application and evaluated on a VIVO iQOO 11S smartphone, equipped with a Snapdragon 8 Gen 2 processor, a 4700 mAh battery, and 12 GB of memory. Through the combination of hardware resources and the developed software, *MentalCare* system is able to comprehensively capture both physiological and psychological states of the participants.

B. Data Collection

We devoted nearly three months to participant recruitment and data collection. The dataset contains 74 participants (40 males and 34 females), aged 14 to 60, all enrolled under an approved Institutional Review Board protocol [38] with informed consent. The patient group data is collected in the psychiatry department of our collaborative hospital, which comprises 18 individuals with depression, 22 with anxiety disorders, and 16 with bipolar disorder, with diagnoses confirmed through standardized clinical scales and psychiatrist evaluations to ensure label reliability. An additional 18 healthy controls were recruited from Shenzhen University and screened for mental stability using the SDS, SCL-90, and SAS scales. The control group include not only students but also faculty, administrative, and cleaning staff, thereby diversifying age, occupation, and lifestyle backgrounds.

As shown in Fig. 2(c), each participant first learns about the functionality and usage of the *MentalCare* system and is guided to remain in a calm state before wearing a custom wristband while resting. During the 10-minute recording, participants are first confirmed by the clinician through brief conversation to be in a calm baseline state, and then instructed to stay as still as possible throughout, so as to minimize motion artifacts, while five physiological signals namely PPG, OXY, GSR, and SKT, are continuously collected at sampling rates of 400 Hz, 400 Hz, 200 Hz, and 100 Hz, respectively.

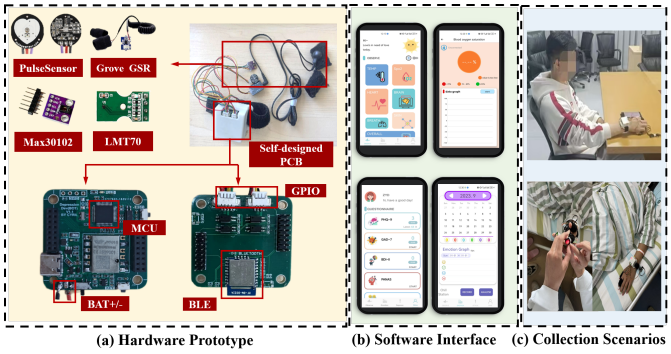


Fig. 2: The system prototype and experiment setup.

C. Baselines

We select representative and SOTA DG-based methods, including latest disentanglement approaches as competitive baselines to evaluate performance superiority of CDPT.

- **CNN** [39] is a classical convolutional neural network architecture that extracts features through 1D convolution.
- **MSTGCN** [40] is a spatial-temporal graph generalization model, which incorporates adversarial domain generalization to confuse the domain discriminator.
- **TSception** [41] is a novel spatial-temporal CNN, which employs multi-scale spatial/temporal convolution kernels to capture multiple granularity features.
- **ManyDG** [42] is a novel DG model which utilizes mutual reconstruction and orthogonal projection to eliminate domain features.
- **MDNet** [18] is a novel multimodal disentanglement learning model. It utilizes the idea of disentangling representation to learn the characteristics of consistency and heterogeneity between modalities.
- **BioDG** [43] is a single-source domain generalization framework that aggregates multi-layer 1D CNN features via lightweight concentration pipelines to extract domain-invariant biosignal representations.
- **CDDG** [23] is a novel causal disentanglement framework. It disentangle causal and domain-related representations through separate encoders.

D. Implementation Details

The experiments adopt Leave-One-Subject-Out Cross-Validation (LOSO-CV) protocol to evaluate cross-subject performance. In each fold, one subject is held out for testing, while the remaining 73 subjects are split into training and validation sets in an $8:2$ ratio. It should be notice that, each test fold contains a single subject and therefore only one ground-truth class. As a result, precision for the remaining classes cannot be meaningfully estimated within that fold. For statistical rigor, we refrain from reporting precision and F1 at the per-fold level.

The training process employs the Adam optimizer with a learning rate of $1e{-}4$, a batch size of 256, a dropout rate of 0.3, and up to 200 epochs to ensure robust convergence. The temperature parameter ϵ is fixed at 0.1. To stabilize

TABLE II: Model performance comparison under LOSO-CV setting. The best prediction results are highlighted in **bold**, while the second best results are underlined. For the result of Wilcoxon signed-rank test: p-value of the improvement of CDPT over the baseline method: * indicating ($p < 0.05$), ** indicating ($p < 0.01$), *** indicating ($p < 0.001$).

Model	Normal	Depression	Anxiety	Bipolar	Overall
CNN [39]	46.78***	64.05***	69.85***	18.44***	51.78***
MSTGCN [40]	55.08***	58.80***	56.58***	72.66***	60.30***
TSception [41]	78.71***	73.97*	80.27**	78.44*	78.35***
BioDG [43]	78.92***	78.06	84.98*	81.34	81.06***
ManyDG [42]	70.01**	78.97	80.14**	80.44*	77.56***
MDNet [18]	52.59***	**79.01**	79.33**	44.96***	65.49***
CDDG [23]	72.40**	74.19*	82.23*	73.47***	76.04***
CDPT(Ours)	**85.61**	78.51	**86.61**	**84.71**	**83.97**

learning and balance the loss components, the weights are set as $\lambda_1 = 5\text{e}{-1}$, $\lambda_2 = 1\text{e}{-1}$, $\lambda_3 = 5\text{e}{-2}$, and $\lambda_4 = 2\text{e}{-2}$. The classification weight parameter δ is set to 0.9. Based on validation performance, early stopping is applied with a patience of 35 epochs. Final results are reported as the average accuracy across all test folds.

V. Results

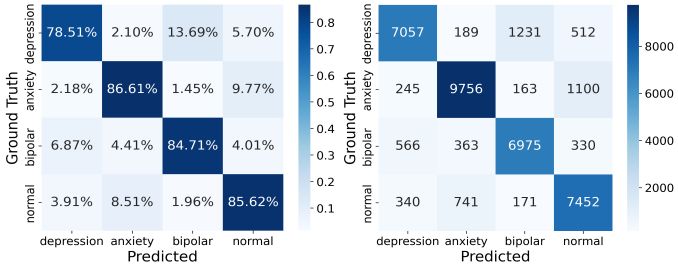
A. Cross-validation Performance with Statistical Test

As show in Table. III, we present classification results for the overall dataset as well as for each label category. To assess the reliability of our findings, we employ the Wilcoxon signed-rank test, a statistical analysis method commonly applied in time series forecasting [44]. Our proposed method achieves an optimal overall accuracy of 83.97%, surpassing the second-best method BioDG by 2.81% and demonstrating statistically significant superiority ($p < 0.001$) over all baseline methods.

For the classification results of individual label categories, the CDPT model achieves the best performance in distinguishing healthy individuals, anxiety disorders, and bipolar disorder. Note that in the cross-subject setting, establishing statistical significance for a single label becomes more difficult because the number of paired samples decreases from the full dataset to only those within that class. This reduction, together with strong physiological heterogeneity across individuals, limits the ability of a stronger model to show consistent and significant improvements at the class level. Even so, statistical comparisons of classification results show that our recognition performance reaches significance across all categories except for bipolar disorder, where the comparison does not achieve significance ($p = 0.058$ for CDPT vs. BioDG). For depression group, our method achieves the third-best performance, falling below MDNet and ManyDG by 0.46% and 0.5%, respectively. Statistical tests indicate that neither difference is statistically significant ($p = 0.865$ for CDPT vs. BioDG and $p = 0.831$ for CDPT vs. ManyDG). In addition, cross-subject experiments reveal a distinct characteristic: DG techniques significantly outperform traditional deep learning methods.

B. Confusion Matrix Analysis

After completing all folds of LOSO-CV, we concatenate the predicted labels and corresponding ground-truth labels across folds to construct an overall confusion matrix.



(a) Probability Matrix (b) Count Matrix

Fig. 3: Confusion matrices for multi-class PDs detection.

TABLE III: Ablation study results. All notation and symbols are consistent with Table II.

Model	Normal	Depression	Anxiety	Bipolar	Overall
CDPT w/o CL	80.53*	**81.53**	85.54	82.73*	82.77*
CDPT w/o AFS	81.73*	81.15	85.42	81.73*	82.92*
CDPT w/o SCD	80.98*	77.48	86.39	82.72*	81.54*
CDPT w/o SCNN	77.51**	77.17	85.10*	82.02*	80.70***
CDPT(Ours)	**85.61**	78.51	**86.61**	**84.71**	**83.97**

As shown in the Fig. 3, the values on the main diagonal represent the accuracy for each class, while the off-diagonal entries reveal two major confusion patterns. First, depression and bipolar exhibit the strongest bidirectional confusion. Specifically, 1231 depression samples (13.69%) are misclassified as bipolar disorder, while 566 bipolar samples (6.87%) are predicted as depression, together accounting for nearly one-fifth of both cohorts. This confusion is consistent with the overlapping clinical symptoms of these disorders, such as persistent low mood, reduced energy, and cognitive impairments, which can make them challenging to distinguish. Besides, anxiety receives the most misclassified samples, with a total of 1293 instances incorrectly assigned to this category. These include 189 depression samples (2.10%), 741 normal samples (8.51%), and 363 bipolar samples (4.41%). This pattern indicates that our model has limitations in distinguishing anxiety from the other categories, particularly when features of different disorders partially overlap. Finally, normal samples are most often confused with anxiety, accounting for 741 samples (8.51%), followed by depression with 340 samples (3.91%). Mis-classification into bipolar disorder is the least frequent, with only 171 samples (1.96%). This aligns with the fact that bipolar disorder presents distinctive episodic patterns (e.g., manic phases), which are easier for the model to separate from normal mental states compared to depression or anxiety.

C. Ablation Study

To further validate the effectiveness of each module in the proposed CDPT framework, we construct four ablation models: w/o CL, w/o AFS, w/o SCD, and w/o SCNN. These variants independently evaluate the contribution of contrastive learning, adaptive feature selection, decoupled representation learning, and SCNN to mental disorder prediction. In the

w/o SCD model, we replace the SCD network with a two-layer multilayer perceptron to ensure the framework retains a classifier. The results are reported in the Table II. All ablation models show different levels of performance degradation. Compared with the full CDPT, w/o CL, w/o AFS, w/o SCD, and w/o SCNN reduce accuracy by 1.2% ($p = 0.0061$), 1.05% ($p = 0.018$), 2.43% ($p = 0.0028$), and 3.27% ($p = 0.0007$), respectively. Statistical tests confirm that the generalization degradation in all ablation models is significant since all p values are below 0.05. Among these models, removing the modal encoder produces the largest performance drop because downstream layers fail to extract effective discriminative features when early-stage independent multi-scale temporal dynamics are not captured for each physiological signal. The w/o SCD variant also causes a considerable decline since the absence of the SCD network weakens the extraction of domain-invariant features across subjects in downstream tasks. In contrast, the w/o CL and w/o AFS models exhibit relatively smaller degradation because contrastive learning and adaptive feature selection act as auxiliary optimizers. Their functions are to enhance the separation of private representations and to exploit source-domain private features for prediction on the target domain. Nevertheless, statistical tests still validate the effectiveness of both optimization designs.

Further analysis of classification performance across label categories reveals a pattern consistent with previous observations. The complete CDPT framework achieves the highest recognition accuracy for healthy individuals, anxiety disorders, and bipolar disorder when compared with ablation models. Its advantage over all ablation models is statistically significant ($p < 0.05$) for both anxiety disorder and bipolar disorder recognition. For depression classification, the w/o CL and w/o AFS variants obtain slightly better predictive results, yet this improvement is not statistically significant.

D. Analysis of Performance Influencing Factors

1) Comparison of Channel Combinations: Different types of physiological signals play a synergistic or complementary role in PDs detection. To investigate the impact of these signals, we conduct channel ablation experiments. Detailed results are presented in Fig. 4, where the horizontal axis represents anonymized subject IDs and the vertical axis denotes different channel combinations. Starting from the four-channel combination, the recognition performance gradually decreases as the number of channels is reduced, with accuracies of 80.11%, 76.8%, 72.19%, and 67.38%, respectively. This validates that incorporating multiple physiological signals facilitates the recognition of mental disorders.

In addition, there are some important results can be observed. Owing to substantial physiological heterogeneity among subjects, the model fails to achieve consistently high recognition accuracy across all individuals. Among these subjects, S16, S24, and S61 consistently exhibit poor detection performance across almost all channel combinations (accuracy lower than 20%). In contrast, the model achieves superior performance for many other subjects, with accuracies exceeding

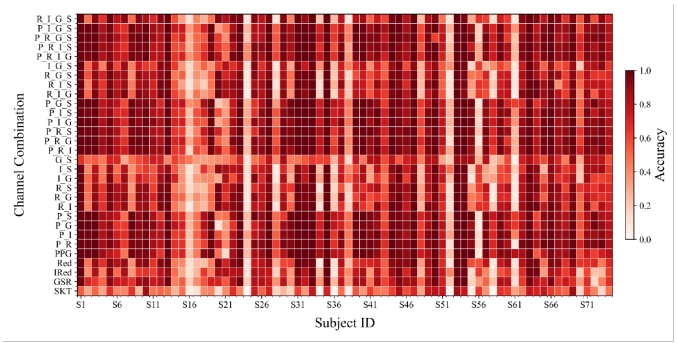


Fig. 4: The accuracy distribution results of different channel combinations for various subjects.

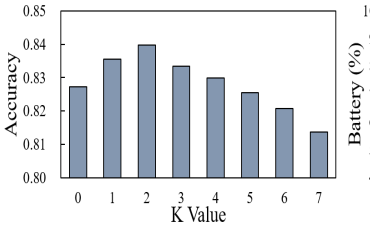


Fig. 5: Performance variation of the AFS Module with different K values.

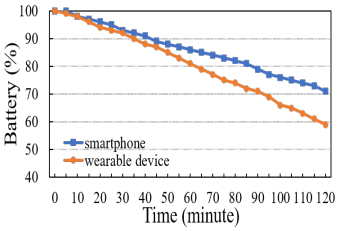


Fig. 6: Power trends of smartphone and wearable devices under continuous usage.

90%. Furthermore, within single-modality analysis, we find that using only PPG signal achieves an accuracy of 78.1%, which is not only substantially higher than other unimodal physiological signals (Red: 68.94%, IRed: 62.57%, GSR: 67.16%, and SKT: 46.82%) but also exceeds certain two-channel combinations that exclude PPG signal(e.g., R_S, R_G, and R_I). This finding demonstrates a stronger association between PPG signal and PDs prediction, and it is consistent with observations reported in existing studies [19], [45].

2) Top-K Value: Positive transfer is a critical mechanism in our framework for improving cross-subject recognition performance. However, incorporating heterogeneous features from different numbers of source-domain subjects (i.e., the Top-K value) influences the performance. To investigate this effect, we evaluate the model under varying values of K. The experimental results are illustrated in Fig. 5, where $K = 0$ indicates that no source-domain heterogeneous features are used. The results clearly show that as K increases, the recognition performance first improves and then declines. The best accuracy of 83.97% is achieved at $K = 2$. When K continues to increase beyond 2, performance gradually decreases, although it still remains higher than the baseline case of $K = 0$. Starting from $K = 5$, the recognition accuracy drops to 82.55%, which is lower than the 82.72% obtained at $K = 0$. This indicates that incorporating an excessive number of heterogeneous features introduces substantial discrepancies between source and target domains, thereby impairing recognition. Consequently, maintaining a relatively small K is more suitable for positive transfer learning in this framework.

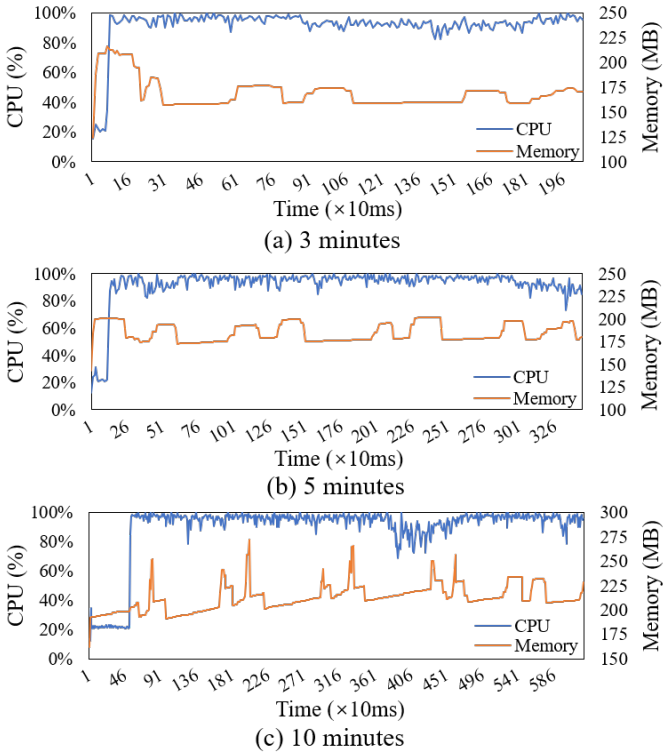


Fig. 7: CPU and memory utilization of mobile applications for predicting physiological signals with different durations.

E. System Evaluations

1) CPU and Memory Load: This experiment evaluates CPU and memory usage during data prediction on an edge device. To ensure precise measurements, all background applications and services are temporarily disabled throughout testing. Real-time CPU and memory utilization is continuously monitored for 3, 5, and 10 minutes using Android Profiler in USB debugging mode. As shown in Fig. 7, the pipeline has two stages. In data loading, latency rises with input length: 80 ms at 3 min, 130 ms at 5 min, and 540 ms at 10 min, while CPU use stays low (22% to 28%). The second stage covers preprocessing and prediction; preprocessing takes 200/340/900 ms for 3/5/10 min, much shorter than prediction at 1.76/2.87/4.81 s. Even so, 10 min of data are processed end to end in under 7 s, enabling edge deployment. This stage is CPU intensive (94% to 100%) because both steps are compute heavy. Memory grows with sequence length: 170 MB at 3 min, 185 MB at 5 min, and 203 MB at 10 min, which remains well within modern smartphone RAM.

2) Battery Consumption: We assess the system's feasibility for extended use by measuring power consumption of the wearable and the smartphone during continuous operation after fully charging both devices. Participants wear the wristband and launch the custom application while closing nonessential apps. After tapping Start, power use is logged via the Android Debug Bridge. We simulate a session of two hours in which the wristband continuously acquires physiological signals and streams them to the smartphone for real-time visualization while the screen stays on.

Fig. 6 shows power consumption for both devices, with an approximately linear decline. On average, the smartphone depletes 15% of battery per hour and the wearable 21%. Extrapolating indicates about five hours of continuous collection, meeting long-duration requirements.

VI. Discussion and Conclusion

In this work, we present *MentalCare*, an all-in-one PDs system unifying AI model, wearable sensing hardware and edge mobile software. While *MentalCare* is authentically user-independent system, however, wearable physiological signals still reflect cohort-specific factors (e.g., medication and lifestyle). Therefore, the current results should be interpreted as a feasibility study rather than purely PD-specific signatures. We control the acquisition as much as possible by using identical wristband hardware, a unified mobile app, and a standardized recording protocol, but residual confounding and limited representativeness may remain. Concretely, given the limited dataset size, our study cannot guarantee effective generalization across all real-world conditions, such as varying medication, lifestyle patterns, age groups, and comorbid physical illnesses. That said, our current model demonstrates reasonable robustness to basic demographic variations (e.g., age and gender), as these factors did not introduce significant distribution shifts in our cohort. However, it is important to highlight that changes in medication status and the presence of comorbidities can substantially affect the system's recognition performance. In this study, we restricted participants to those who had not taken psychiatric medication on the day of data collection and excluded individuals with known physiological disorders (e.g., cardiovascular diseases) to avoid large confounding shifts. This limitation reflects a core obstacle in physiological-signal mental health research [46], where reliable detection under medication or comorbidity remains scarcely studied and a critical open challenge for deployment.

In the future, we will recruit community controls from the same catchment area, broaden sociodemographic coverage, and include participants with common physical comorbidities to evaluate the robustness of MentalCare in more realistic and clinically heterogeneous settings.

VII. Acknowledgement

This work was supported in part by China NSFC (62172286, U2001207) ; in part by key program of DEGP (2024ZDZX1011); in part by Shenzhen Science and Technology Program (JCYJ20250604181429038); in part by Shenzhen Science and Technology Foundation (ZDSYS20190902092853047); in part by Youth S&T Talent Support Programme of GDSTA (SKXRC2025231); in part by Guangdong Yiyang Healthcare Charity Foundation (JZ2022001-3); in part by Talents' plan Foundation of Guangdong Second Provincial General Hospital (2024C003). Yongpan Zou is the corresponding author.

REFERENCES

[1] W. H. Organization, "World mental health report: Transforming mental health for all," 2022.

[2] M. Guha, "Diagnostic and statistical manual of mental disorders: Dsm-5," *Reference Reviews*, vol. 28, no. 3, pp. 36–37, 2014.

[3] T. G. Westerman and G. E. Dear, "The need for culturally valid psychological assessment tools in indigenous mental health," *Clinical Psychologist*, vol. 27, no. 3, pp. 284–289, 2023.

[4] Y. Ding, Y. Dai, X. Wang, L. Feng, L. Cao, and H. Zhang, "Integrating content-semantics-world knowledge to detect stress from videos," in *Proceedings of the ACM MM*, pp. 10373–10381, 2024.

[5] H. Liu, C. Li, X. Zhang, F. Zhang, W. Wang, F. Ma, H. Chen, H. Yu, and X. Zhang, "Depression detection via capsule networks with contrastive learning," in *Proceedings of the AAAI*, vol. 38, pp. 22231–22239, 2024.

[6] J. Yu and H. Kaya, "Using emotionally rich speech segments for depression prediction," in *Proceedings of the ICASSP*, pp. 1–5, IEEE, 2025.

[7] S. S. Leal, S. Ntalampiras, and R. Sassi, "Speech-based depression assessment: A comprehensive survey," *IEEE TAFFC*, 2024.

[8] K. Yang, T. Zhang, Z. Kuang, Q. Xie, J. Huang, and S. Ananiadou, "Mentallama: interpretable mental health analysis on social media with large language models," in *Proceedings of the ACM WWW*, pp. 4489–4500, 2024.

[9] M. Yang, E. C. Ngai, X. Hu, B. Hu, J. Liu, E. Gelenbe, and V. C. Leung, "Digital phenotyping and feature extraction on smartphone data for depression detection," *Proceedings of the IEEE*, 2025.

[10] J. Cheong, S. Kalkan, and H. Gunes, "Fairrefuse: Referee-guided fusion for multi-modal causal fairness in depression detection," in *Proceedings of the IJCAI*, 2024.

[11] Y. Ren, Y. Chen, M. C. Chuah, and J. Yang, "User verification leveraging gait recognition for smartphone enabled mobile healthcare systems," *IEEE TMC*, vol. 14, no. 9, pp. 1961–1974, 2015.

[12] X. Shui, H. Xu, S. Tan, and D. Zhang, "Depression recognition using daily wearable-derived physiological data," *Sensors*, vol. 25, no. 2, p. 567, 2025.

[13] W. Kuang, S. Jin, D. Wang, Y. Zheng, Y. Zou, and K. Wu, "Emotracer: A user-independent wearable emotion tracer with multi-source physiological sensors based on few-shot learning," in *Proceedings of the BIBM*, pp. 2680–2687, 2024.

[14] K. Prizeman, C. McCabe, and N. Weinstein, "Stigma and its impact on disclosure and mental health secrecy in young people with clinical depression symptoms: A qualitative analysis," *Plos one*, vol. 19, no. 1, p. e0296221, 2024.

[15] S. Saganowski, B. Perz, A. G. Polak, and P. Kazienko, "Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review," *TAFFC*, vol. 14, no. 3, pp. 1876–1897, 2022.

[16] Y. Ding, N. Robinson, C. Tong, Q. Zeng, and C. Guan, "Lggnet: Learning from local-global-graph representations for brain–computer interface," *IEEE TNNLS*, vol. 35, no. 7, pp. 9773–9786, 2024.

[17] L. Zhao, R. Lyu, H. Lei, Q. Lin, A. Zhou, H. Ma, J. Wang, X. Meng, C. Shao, Y. Tang, *et al.*, "Airecg: Contactless electrocardiogram for cardiac disease monitoring via mmwave sensing and cross-domain diffusion model," *Proceedings of the ACM IMWUT*, vol. 8, no. 3, pp. 1–27, 2024.

[18] Z. Jia, F. Zhao, Y. Guo, H. Chen, T. Jiang, and B. Center, "Multi-level disentangling network for cross-subject emotion recognition based on multimodal physiological signals," in *Proceedings of the IJCAI*, pp. 3069–3077, 2024.

[19] Y. Zhang, S. Jin, W. Kuang, Y. Zheng, Q. Song, C. Fan, Y. Zou, V. C. Leung, and K. Wu, "Depguard: Depression recognition and episode monitoring system with a ubiquitous wrist-worn device," *IEEE TMC*, vol. 25, no. 1, pp. 197–214, 2026.

[20] T. Chen, Y. Guo, S. Hao, and R. Hong, "Semi-supervised domain adaptation for major depressive disorder detection," *IEEE TMM*, vol. 26, pp. 3567–3579, 2023.

[21] W. Kuang, Y. Zhang, Y. Cheng, K. Wu, C. Fan, and Y. Zou, "Generalizable psychiatric screening from wearable: Causal disentanglement of multi-channel biosignals with interpretability analysis," in *Proceedings of the BIBM*, pp. 3745–3750, 2025.

[22] W. Deng, S. Zhong, R. Lu, and Y. Wang, "Identity-domain removal for robust eeg-based emotion recognition," in *Proceedings of the ICMR*, pp. 183–191, 2025.

[23] L. Jia, T. W. Chow, and Y. Yuan, "Causal disentanglement domain generalization for time-series signal fault diagnosis," *Neural Networks*, vol. 172, p. 106099, 2024.

[24] Y. Yao and G. Doretto, "Boosting for transfer learning with multiple sources," in *CVPR*, pp. 1855–1862, IEEE, 2010.

[25] Y. Ouzar, C. Nineuil, F. Boualeb, E. Pierson, A. Amad, and M. Daoudi, "Wearable-derived behavioral and physiological biomarkers for classifying unipolar and bipolar depression severity," in *In Proceedings of the IEEE FG*, pp. 1–10, 2025.

[26] S. Hassantabar, J. Zhang, H. Yin, and N. K. Jha, "Mhdeep: Mental health disorder detection system based on wearable sensors and artificial neural networks," *ACM TECS*, vol. 21, no. 6, pp. 1–22, 2022.

[27] A. Ahmed, J. Ramesh, S. Ganguly, R. Aburukba, A. Sagahyroon, and F. Aloul, "Evaluating multimodal wearable sensors for quantifying affective states and depression with neural networks," *IEEE Sensors Journal*, vol. 23, no. 19, pp. 22788–22802, 2023.

[28] M. Naeem, S. A. Fawzi, H. Anwar, and A. S. Malek, "Wearable ecg systems for accurate mental stress detection: a scoping review," *Journal of Public Health*, vol. 33, no. 6, pp. 1181–1197, 2025.

[29] C. Shende, S. Sahoo, S. Sam, P. Patel, R. Morillo, X. Wang, S. Ware, J. Bi, J. Kamath, A. Russell, *et al.*, "Predicting symptom improvement during depression treatment using sleep sensory data," *Proceedings of the ACM IMWUT*, vol. 7, no. 3, pp. 1–21, 2023.

[30] D. Sethia, S. Indu, *et al.*, "St-cirl: a reinforcement learning-based feature selection approach for enhanced anxiety classification," *Physiological Measurement*, vol. 13, no. 2, p. 025009, 2025.

[31] M. Li, J. Li, Y. Chen, and B. Hu, "Stress severity detection in college students using emotional pulse signals and deep learning," *IEEE TAFFC*, 2025.

[32] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the ACM MM*, pp. 1122–1131, 2020.

[33] D. Yang, S. Huang, H. Kuang, Y. Du, and L. Zhang, "Disentangled representation learning for multimodal emotion recognition," in *Proceedings of the ACM MM*, pp. 1642–1651, 2022.

[34] Y. Pan, J. Jiang, K. Jiang, and X. Liu, "Disentangled-multimodal privileged knowledge distillation for depression recognition with incomplete multimodal data," in *Proceedings of the ACM MM*, pp. 5712–5721, 2024.

[35] L. Han, H.-J. Ye, and D.-C. Zhan, "The capacity and robustness trade-off: Revisiting the channel independent strategy for multivariate time series forecasting," *IEEE TKDE*, vol. 36, no. 11, pp. 7129–7142, 2024.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[37] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *NIPS*, vol. 33, pp. 18661–18673, 2020.

[38] C. Grady, "Institutional review boards: Purpose and challenges," *Chest*, vol. 148, no. 5, pp. 1148–1155, 2015.

[39] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 2002.

[40] Z. Jia, Y. Lin, J. Wang, X. Ning, Y. He, R. Zhou, Y. Zhou, and L.-w. H. Lehman, "Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification," *IEEE TNSRE*, vol. 29, pp. 1977–1986, 2021.

[41] Y. Ding, N. Robinson, S. Zhang, Q. Zeng, and C. Guan, "Tsception: Capturing temporal dynamics and spatial asymmetry from eeg for emotion recognition," *IEEE TAFFC*, vol. 14, no. 3, pp. 2238–2250, 2023.

[42] C. Yang, M. B. Westover, and J. Sun, "Manydg: Many-domain generalization for healthcare applications," in *Proceedings of the ICLR*, 2023.

[43] A. Ballas and C. Diou, "Towards domain generalization for ecg and eeg classification: Algorithms and benchmarks," *IEEE TETCI*, vol. 8, no. 1, pp. 44–54, 2023.

[44] F. Chen, Y. Zhang, L. Chen, X. Meng, Y. Qi, and J. Wang, "Dynamic traveling time forecasting based on spatial-temporal graph convolutional networks," *FCS*, vol. 17, no. 6, p. 176615, 2023.

[45] P. Kumar, S. Misra, Z. Shao, B. Zhu, B. Raman, and X. Li, "Multimodal interpretable depression analysis using visual, physiological, audio and textual data," in *Proceedings of the WACV*, pp. 5305–5315, IEEE, 2025.

[46] Z. Zheng, Y. Liang, R. Lyu, J. Bao, Y. Huang, A. Zhou, H. Ma, J. Wang, X. Meng, C. Shao, *et al.*, "Bp3: Improving cuff-less blood pressure monitoring performance by fusing mmwave pulse wave sensing and physiological factors," in *Proceedings of the Sensys*, pp. 730–743, 2024.