MentalCare: Contrastive Disentanglement and Positive Transfer for Generalizable Psychiatric Disorders Detection with A Wrist-worn Device

Yufei Zhang, *Student Member, IEEE*, Wenting Kuang, Kaishun Wu, *Fellow, IEEE*, Zite Huang, Changhe Fan, Yongpan Zou, *Member, IEEE*

***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 system for detecting multiclass psychiatric disorders (PDs) that integrates physiological signals collected from ubiquitous sensors, including photoplethysmography (PPG), skin temperature (SKT), galvanic skin response (GSR), and blood oxygen (OXY). To address physiological heterogeneity between 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 utilizes supervised contrastive learning to facilitate the separation of heterogeneous physiological features between different subjects in the latent subspace. Furthermore, the proposed AFS module identifies source-domain representation that are most compatible with the target input, thereby facilitating cross-subject positive transfer. We collect a real-world dataset comprising 74 participants diagnosed with depression, bipolar disorder, anxiety, or healthy controls. The experimental results demonstrate that CDPT achieves an overall accuracy of 83.34% at the window-level and 91.89% at the subject-level under the leave-one-subject-out cross-validation protocol. At the system level, we further validate the effectiveness and rationality of the proposed *MentalCare* when deployed on edge devices.**

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

Received 30 December 2024; revised 29 June 2025; accepted 28 July 2025. This work was supported in part by China NSFC (62472366), in part by Key Project of Department of Education of Guangdong Province (2024ZDZX1011, 2023KCXTD042, 2024GCZX003, 2024KCXTD008, 2025KCXTD056), in part by Youth S& T Talent Support Programme of GDSTA (SKXRC2025231), in part by Shenzhen Science and Technology Program (JCYJ20250604181429038), in part by the Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (2023B1212010007), in part by 111 Center (D25008), in part by Shenzhen Science and Technology Foundation (ZDSYS20190902092853047). We also extend our sincere gratitude to Prof. Fangshu Chen at Shanghai Polytechnic University for providing additional computing resources during the revision of this work. (*Corresponding Author: Yongpan Zou*)

Approval of all ethical and experimental procedures and protocols was granted by Institutional Review Board (IRB), Faculty of Medicine, Shenzhen University under Application No. 202500165, and performed in line with the Declaration of Helsinki.

Yufei Zhang, Wenting Kuang, Zite Huang, and Yongpan Zou are with the College of Computer Science and Software Engineering, Shenzhen University, 3688 Nanhai Ave, Shenzhen, Guangdong, China. Email: yfyaya@foxmail.com; kuangwenting2023@email.szu.edu.cn; 2023150190@email.szu.edu.cn; yongpan@szu.edu.cn.

Kaishun Wu is with the Information Hub, Hong Kong University of Science and Technology (Guangzhou), 1 Duxue Road, Guangzhou, Guangdong, China. Email: wuks@hkust-gz.edu.cn.

Changhe Fan is with Department of Psychiatry, Affiliated Guangdong Second Provincial General Hospital of Jinan University, Guangzhou 510317, China. Email: changhefan@yahoo.com.

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 cause 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 main approaches to psychiatric assessment [3]. However, as reported by the 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], [12]. However, these modalities are easily affected by the social mask, where patients intentionally conceal or modify their expressions and verbal content to obscure their actual emotional states [13], [14]. This behavior, often driven by shame, fear of stigmatization, or social expectations, leads to the concealment or modification of symptom descriptions, reducing the reliability of the diagnosis [15], [16]. In contrast, physiological signals [17], [18], [19], [20] directly capture variations in physiological responses regulated by the autonomic nervous system or the peripheral nervous system, offering a more objective and reliable biomarker [21].

Nowadays, physiological signals are recognized as a critical modality for intelligent PDs detection systems. Many studies have examined physiological signals, including electroencephalography (EEG) [22], electrocardiography (ECG) [23], and electrooculography (EOG) [24]. Although 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, 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 capacity for continuous, unobtrusive monitoring in daily life, making them particularly suitable for scalable, long-term assessment and intervention in mental health [17]. However, leveraging these signals effectively remains non-trivial. Beyond noise and low signal quality, achieving accurate PDs detection under ***multiclass diagnosis*** and ***domain generalization*** (DG) settings remains a critical challenge.

C1: Generalizing from single-class PD to multi-class PDs diagnosis. Most existing studies focus on single-class PD recognition [17], [19]. However, different PDs often exhibit partially overlapping symptoms and comorbid profiles, and their physiological signatures can share substantial commonalities (*e.g.*, generalized arousal or stress responses). Consequently, a considerable portion of wearable-derived biomarkers are transdiagnostic in the sense that they reflect shared physiological dimensions across disorders rather than disorder-specific signatures, which limits their ability to separate PDs categories.

C2: Enhancing Cross-Subject Generality with DG-based Systems. In PDs detection, individuals exhibit substantial differences in neural dynamics and physiological response patterns. Even under similar clinical symptoms, physiological signals may reveal highly personalized characteristics [17], [25]. Such heterogeneity introduces significant domain shifts across subjects, making models difficult to generalize reliably to unseen subjects. Prior studies have successfully employed domain adaptation (DA) strategies. Zhang et al. [17] apply unsupervised DA for depression recognition and few-shot learning for episode monitoring, while Chen et al. [26] propose GNN-SDA for EEG-based semi-supervised DA in depression detection. These approaches effectively capture subject-specific characteristics and improve cross-domain adaptability. However, DA-based approaches typically depend on target-specific information to align distributions, which compromises subject adaptability and is often impractical for real-world settings (*i.e.*, subject-independent systems) where such information is unavailable [27]. Conversely, although DG is more suitable for cross-user systems, designing an effective method is often more challenging than DA because of its no-target-access property, which prevents the use of any target-domain data for distribution alignment or calibration.

To bridge these gaps under the above two challenges, we propose *MentalCare*, a DG-based ubiquitous system consisting of a smartphone application and a wrist-worn device for the early detection of depression, anxiety, and bipolar disorders. The DG setting enables cold-start deployment to unseen subjects by learning solely from source domains without requiring any target-domain access. In *MentalCare* system, the wristband continuously collects the subject's multimodal physiological signals, including PPG, SKT, GSR, and blood oxygen (OXY). To address the substantial inter-individual variability, we propose a contrastive disentanglement with a positive transfer framework termed CDPT. 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.*, unseen new subjects) [28], [24], [29]. However, in cross-domain scenarios, certain source domains may contain complementary information relevant to the target domain. Using this information can improve the accuracy of the prediction, a phenomenon referred to as positive transfer [30]. In CDPT, the AFS module is designed to implicitly learn the distributional compatibility across source domains. During testing, this mechanism enables the dynamic transfer of complementary heterogeneous representations from source domains to unseen target domains, thereby improving target domain performance. In addition, we employ contrastive learning to encourage the separation of heterogeneous features between subjects during the disentanglement process.

In summary, 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 83.34% cross-subject accuracy at the window-level and 91.89% at the subject-level on our collected dataset.
- We propose CDPT, a novel disentanglement-based framework for multimodal ubiquitous sensing in PDs detection, together with two key enhancements: the SCD and AFS modules.
- We perform interpretability analyses based on channel combinations and attention weights to quantify the contribution of each modality to PDs prediction and provide interpretable physiological findings.
- To promote further research in mobile health, we collect a real-world, ubiquitous, and multimodal physiological dataset from individuals diagnosed with depression, anxiety, and bipolar disorders by certified psychiatrists at our partner hospital[1].

The remainder of the paper is conducted as follows. Sec. II reviews related work on ubiquitous physiological sensing and cross-subject detection of PDs. Subsequently, Sec. III introduces the physiological signal processing pipeline and CDPT framework for PDs detection. Then Sec. IV performs the implementation of our proposals. Sec. V evaluates the effectiveness of *MentalCare*. Finally, Sec. VI and Sec. VII discuss and conclude our work.

II. Related Work

A. Ubiquitous Physiological Sensing for PDs Detection

Pervasive sensing technologies are increasingly being explored for reliable PDs detection based on physiological signals. For example, Ouzar et al. [31] and Shui et al. [13] leverage wearable devices to collect physiological and behavioral data, employing multiple machine learning approaches to

[1]The source codes and test samples are available at *https://github.com/ISAC-GROUP/AFFC-MentalCare*.

detect depression. Hassantabar et al. [32] introduce MHDeep, a framework that integrates wearable sensor data with artificial neural networks to classify mental health disorders. Ahmed et al. [33] propose a multimodal framework that combines convolutional neural networks and random forests to assess affective states and disorders. However, existing approaches employ traditional machine learning or simple neural network techniques [31], [33], which are limited in their ability to uncover complex distributional characteristics between domains. Furthermore, many methods rely solely on a single physiological signal [34], [35], thus missing the complementary information and enhanced discriminative power that can be obtained from integrating multiple peripheral signals. Even if a few studies further improve the diagnostic effectiveness of PDs detection through carefully designed deep learning methods [36], [37]. However, these studies also lack consideration for individual differences, making it difficult to apply models or systems to new subjects. In addition, most of the existing work focuses on a single type of PD diagnosis and rarely generalizes to the detection of multiple mental disorders under more complex conditions [32].

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 proposes MISA [38], which imposes orthogonality constraints and Central Moment Discrepancy (CMD) to guide the separation of factors. Along this line, Yang et al. [39] refine the consistency objective and replace CMD with an adversarial network for more stable disentanglement. Jia et al. [24] propose a multi-level disentanglement model for multimodal physiological sensing in emotion recognition, addressing signal heterogeneity and cross-modal complementarity. Pan et al. [40] propose the Dis^{2}DR framework for depression recognition with missing modalities. It uses the teacher model trained on full multimodal data, while the student model applies randomized masking to simulate modality absence.

Relative to previous work, we present two advances over disentanglement learning. First, we introduce an AFS module that facilitates positive transfer by measuring the distributional consistency between target inputs and source domains. Rather than explicit similarity matching, AFS identifies the most compatible source-specific transformations to project target features into multiple complementary subspaces. This mechanism enables the model to uncover complex physiological heterogeneities and enhances recognition performance in the DG setting. Second, we incorporate contrastive learning to ensure that the heterogeneous features of each subject are sufficiently discriminative. This explicit inter-subject separation within the subject-specific subspace enhances the discriminative quality of the learned representations and enables AFS to reliably identify source subjects that are genuinely similar to the target.

III. Methodology

In this section, we introduce *MentalCare* for multi-class PDs detection. We first formalize the necessary definitions and notations (see Sec. III-A). We then describe the physiological signal processing pipeline used to prepare multimodal inputs (see Sec. III-B). Next, we present the CDPT framework in the order used during training (see Sec. III-C). Finally, we summarize the training and inference procedures of the CDPT framework (see Sec. III-D).

A. Problem and Notation Definition

To better illustrate our proposal, Table I presents the frequently appearing notations, and the relevant concepts are mathematically defined as follows.

Definition 1 (Cross-subject PDs Detection) We study cross-subject PDs detection using multimodal ubiquitous physiological signals. Therefore, we define that the model is trained on labeled data from multiple source subjects and is generalized to an unseen target subject without access to its labels during training.

Formally, let there be M labeled source subjects. We represent the physiological data of the i-th subject in the t-th sliding window as the 3-tuple $\mathbb{D}_t^i = \left(\{X_j^i\}_{j=1}^N, Y^i, \hat{Y}^i\right)_t$, where $X_j^i \in \mathbb{R}^L$ denotes the physiological time series of the j-th modality, L is the window length, $Y^i \in \{0, \ldots, 3\}$ is the ground truth diagnosis, and $\hat{Y}^i$ is the identity label of the subject. By aggregating all sliding windows of the i-th source subject, we obtain the i-th source dataset as $\mathbb{D}^i = \{\mathbb{D}_t^i\}_{t=1}^{\mathcal{T}^i}$, where $\mathcal{T}^i$ is the number of windows of the i-th subject. Therefore, the overall labeled source set is $\mathbb{D}^S = \{\mathbb{D}^i\}_{i=1}^M$.

Meanwhile, we denote the physiological data of the unseen target subject by $\mathbb{D}_t^T = (\{X_j^T\}_{j=1}^N, Y^T, \hat{Y}^T)_t$, where $X_j^T \in \mathbb{R}^L$ is the physiological time series of the j-th modality, and $\hat{Y}^T$ is the identity label of the subject. $Y^T \in \{0, \ldots, 3\}$ denotes the ground-truth diagnosis. By aggregating all sliding windows of the target subject, we define the target data set as $\mathbb{D}^T = \{\mathbb{D}_t^T\}_{t=1}^{\mathcal{T}}$, where $\mathcal{T}$ is the number of windows for the target subject.

Definition 2 (Domain Shift Across Subject) Physiological heterogeneity induces joint distributions $\mathcal{P}_{X,Y}^i$ such that $\mathcal{P}_{X,Y}^i \neq \mathcal{P}_{X,Y}^j$ for $1 \leq i \neq j \leq M$. The unseen target subject T is drawn from another distribution $\mathcal{P}_{X,Y}^T$ that typically differs from all sources.

In this work, we model this cross-subject domain shift as a *covariate shift*. Concretely, the marginal covariate distributions vary between subjects, *i.e.*, $\mathcal{P}_X^i \neq \mathcal{P}_X^j$ for $1 \leq i \neq j \leq M$, while the conditional distribution $\mathcal{P}_{Y|X}$ is assumed to be invariant between subjects. The unseen target subject T is therefore drawn from a covariate distribution $\mathcal{P}_X^T$ that typically differs from all source marginals $\{\mathcal{P}_X^i\}_{i=1}^M$, but shares the same conditional distribution $\mathcal{P}_{Y|X}$.

Definition 3 (DG-based Prediction Goal) The proposed CDPT framework (see Sec. III-C) learns a mapping $f_\theta : \mathbb{R}^{N \times L} \to \mathbb{R}^Z$ using only $\mathbb{D}^S$, such that the expected target risk is minimized without accessing any (X^T, Y^T) from $\mathbb{D}^T$ during training:

$$\min_{\theta} \ \mathbb{E}_{(X,Y)\sim\mathcal{P}_{X,Y}^T}\left[\ell\big(f_\theta(X), Y\big)\right], \tag{1}$$

where $\ell(\cdot,\cdot)$ is a multi-class loss and Z denotes the number of diagnostic categories.

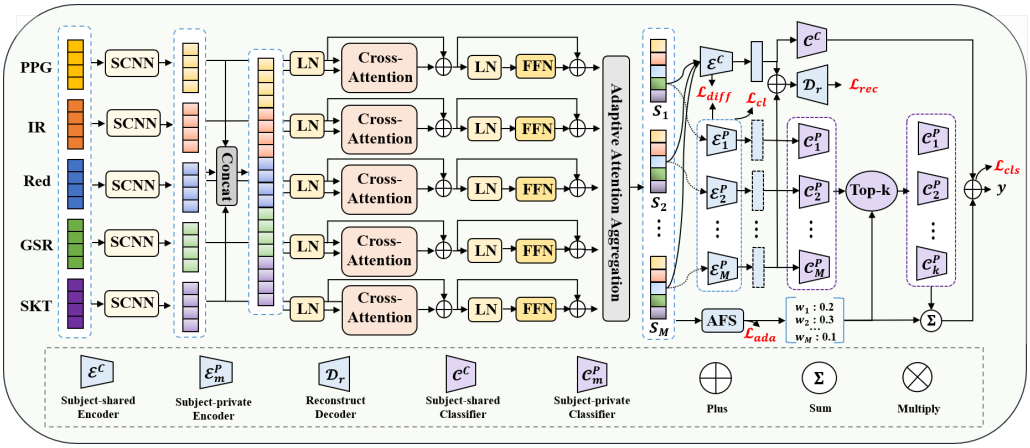


Fig. 1: The structure of CDPT framework.

TABLE I: Notation and the corresponding descriptions.

Symbols	Definitions and Descriptions
M, N, K	Number of source subjects, modalities, and top-k source subjects by specific-feature similarity.
$L, \mathcal{T}$	Window length and number of windows.
$X_j^i,\ X_j^T$	Modality-j time series of i-th source subject and target subject in one window.
Y^C, Y_i^P	Output of shared classifier; Output of i-th subject's private classifier.
$\hat{Y}_i, \tilde{Y}_i, Y_i$	Subject identity label of i-th subject; Ground Truth label of i-th subject; Output of fused final prediction.
F_i, F_g	Encoded feature of modality i; Global combined feature $\mathrm{Concat}(F_1, \ldots, F_N)$.
$\hat{F}_i$	Enhanced modality feature after residual and FFN.
Attn_i	Cross-modal attention output for modality i.
S	Fused multimodal feature.
$\mathcal{E}^C, \mathcal{E}_i^P$	Shared encoder and i-th subject's private encoder.
$\hat{S}, \tilde{S}_i$	Subject-shared (invariant) feature and i-th Subject-private (specific) feature.
$\mathcal{D}_r$	Reconstruction decoder.
$P_i, p_{i,j}$	Adaptive source weights of i-th subject and its j-th element, $\sum_j p_{i,j} = 1$.
$\mathcal{C}^C, \mathcal{C}^S$	General and subject-specific classifiers.
δ	Trade-off between shared and private predictions.

TABLE II: Preprocessing steps for physiological signals.

Physiological Signal	FFT	Filtering Configuration (Hz)	Min-max Normalization
PPG	✓	$0.05 \sim 4$ (Bandpass)	✓
OXY (IR and Red)	✓	$0.1 \sim 4$ (Bandpass)	✓
SKT	✓	50 (Notch)	✓
GSR	✓	$0 \sim 0.3$ (Bandpass)	✓

B. Physiological Signal Processing

First, we conduct a spectral analysis with the Fast Fourier Transform (FFT) to characterize the frequency content of each signal, identify dominant bands, and potential interference. Then, specialized filtering techniques are applied to different modalities. Specifically, a 50 Hz notch filter is applied to the SKT signal to eliminate power-line interference. In contrast, a Butterworth low-pass filter is applied to the remaining signals to retain spectral components within the physiological range and suppress high-frequency noise. Due to the inherent variability of physiological signals, the processing parameters differ for each type of signal, as detailed in Table II. Note that the original OXY saturation signal obtained through the blood oxygen sensor consists of two channels, including the infrared light (IR) signal and the red light (Red) signal.

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

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

C. CDPT Framework

As illustrated in Fig. 1, the CDPT framework comprises four core components—SCNN, CMAI, SCD, and AFS—which operate in synergy to sequentially process and integrate multimodal physiological signals. Specifically, SCNN and CMAI provide the essential processing foundation, focusing on multimodal temporal encoding and cross-modal interaction, respectively. Building upon these representations, SCD and AFS operate in tandem to address the critical challenge of cross-subject distribution shifts and enhance model generalization. We will introduce the core components in their functional order within the training pipeline to illustrate the framework's end-to-end workflow.

1) Modality Encoding: To effectively preserve the modality-specific characteristics of physiological signals, we adopt a channel-independent encoding strategy [41]. 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. The structure of the SCNN module is depicted as Fig. 2. It contains three blocks that progressively enhance temporal representations.

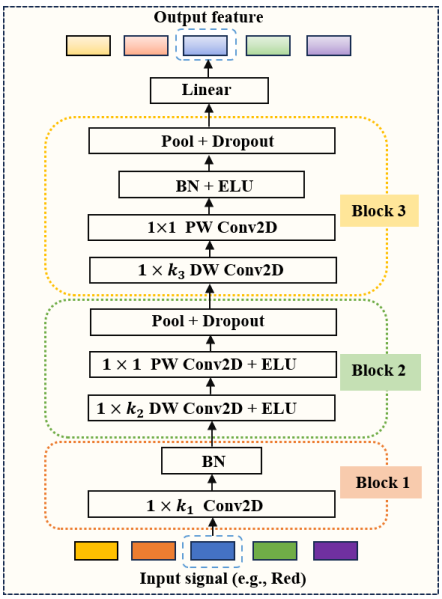


Fig. 2: The structure of SCNN network.

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

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 of each channel. The subsequent pointwise convolution $\text{Conv}^{\text{pw}}_{1\times 1}$ then integrates channel-specific responses, allowing 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, thus 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 applied sequentially as in the previous blocks.

Finally, the output of the third block is flattened and projected into a shared latent space through 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, allowing the model to emphasize the most informative modalities for the downstream task.

Cross-modal interaction phase. As shown in Fig. 3, 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 dimension of the feature, 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:

$$F_g = \text{Concat}(F_1, F_2, \ldots, F_N), \tag{3}$$

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 [42]:

$$K = W_K F_g, \quad V = W_V F_g, \quad Q_i = W_Q^{(i)} F_i, \tag{4}$$

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:

$$\text{Attn}_i = \text{Softmax}\left(Q_i K^\top\right) V, \tag{5}$$

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

$$\hat{F}_i = \text{FFN}(\text{LN}(F_i + \text{Attn}_i)) + (F_i + \text{Attn}_i). \tag{6}$$

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

$$\text{FFN}(x) = W_2\,\sigma(W_1 x + b_1) + b_2, \tag{7}$$

where x denotes the derived feature $\text{LN}(\text{Attn}_i + F_i)$, 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 modality feature is first scored through a shared attention mechanism, resulting in μ_i:

$$\mu_i = q^\top \tanh\left(W_a^{(i)} \hat{F}_i + b_a^{(i)}\right), \tag{8}$$

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

$$\psi_i = \frac{\exp(\mu_i)}{\sum_{j=1}^{M} \exp(\mu_j)}. \tag{9}$$

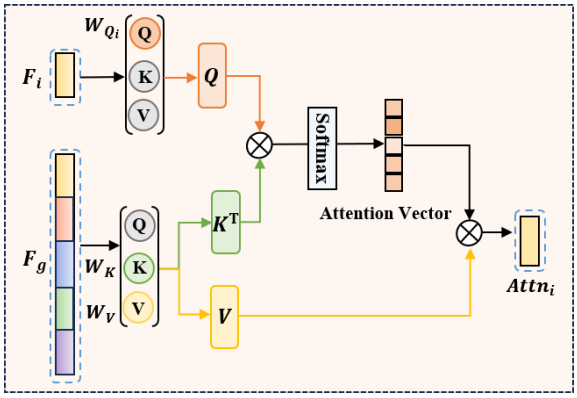


Fig. 3: The procedure of cross attention mechanism.

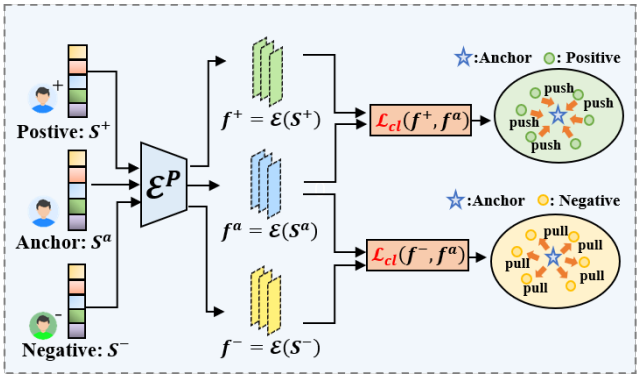


Fig. 4: The procedure of supervised contrastive learning.

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

$$S = \sum_{i=1}^{M} \psi_i \cdot \hat{F}_i. \tag{10}$$

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.

3) Subject-level Contrastive Disentanglement Module: An efficient disentangled representation is defined as one that separates the distinct, independent, and informative generative factors underlying the data [43]. To achieve this, we design the SCD module to separate consistent and specific features among subjects from complex multimodal fusion representations, thereby facilitating cross-subject PDs detection.

Disentangle Encoders: 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 $\hat{S}$. 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) features $\tilde{S}_i$.

$$\hat{S} = \mathcal{E}^C(S, \theta^C), \tag{11}$$

$$\tilde{S}_i = \mathcal{E}_i^P\left(S_i, \theta_i^P\right) (i = 1, 2, \ldots, M), \tag{12}$$

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

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 features:

$$\mathcal{L}_{diff} = \left\| \tilde{S}^\top \hat{S} \right\|_F^2, \tag{13}$$

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.

Reconstruct Supervision: Essentially, reconstruct supervision and orthogonality separation are a pair of mutually constraining components. During the feature disentanglement process, essential information can be lost when separating private and shared components. To explicitly constrain such information loss, we employ a reconstruction decoder $\mathcal{D}_r$ to ensure that the combination of private and shared features preserves the original and useful information.

$$\mathcal{L}_{rec} = \frac{\left\| S - \mathcal{D}_r(\hat{S} + \tilde{S}) \right\|_2^2}{d}, \tag{14}$$

where d is the dimension of S and $\|\cdot\|_2^2$ denotes the squared norm $\mathcal{L}_2$. The reconstruction decoder is implemented as a linear layer that maps the 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.

Supervised Contrastive Learning: Considering that the loss of orthogonal difference acts on both shared and private features, the representations of private features of different subjects may still exhibit certain entanglement. To further promote robust disentangled learning, we incorporate contrastive learning [44] into the disentanglement of private features. Specifically, we adopt an instance-level supervised contrastive learning (SCL) method. The structure of SCL is depicted in Fig. 4. 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 $\hat{Y}_i$. For a given anchor $\tilde{S}_i$, we define the set of positive samples as $\{\tilde{S}_j \,|\, \hat{Y}_j = \hat{Y}_i,\, j \neq i\}$, namely, the features originating from the same subject but in different instances. The remaining samples in the batch are considered negative samples, denoted as $\{\tilde{S}_j \,|\, \hat{Y}_j \neq \hat{Y}_i\}$. The supervised contrastive loss is defined as follows:

$$\mathcal{L}_{cl} = -\frac{1}{L} \sum_{i=1}^{L} \log \frac{\sum_{j:\hat{Y}_j = \hat{Y}_i, j \neq i} \exp\left(\frac{\text{sim}(\tilde{S}_i, \tilde{S}_j)}{\tau}\right)}{\sum_{k \neq i} \exp\left(\frac{\text{sim}(\tilde{S}_i, \tilde{S}_k)}{\tau}\right)}, \tag{15}$$

where $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity function and τ

is a temperature parameter. This contrastive objective makes private features from the same subject cluster tight while pushing those from different subjects apart, improving diagnostic robustness.

4) Positive Transfer for Cross-subject Augmentation: In this part, we design an AFS module to integrate information from source subjects that are most compatible with the target representation. 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 outputs a weight vector $P_i \in \mathbb{R}^M$:

$$P_i = \{p_{i,1},\ p_{i,2}, ...,\ p_{i,M}\}\ s.t.\ \ 0 \leq p_{i,j} \leq 1,\ \sum_{j=1}^{M} p_{i,j} = 1. \tag{16}$$

Here, the selector outputs a probability vector P_i over all source domains. $p_{i,j}$ reflects how the input feature responds to the distribution of source subject j in the learned representation space, rather than a strict pairwise similarity measure between features. This design avoids computing feature similarity explicitly and instead captures the relative compatibility between the input and different source domains.

Although the selector is trained with a subject classification objective, this objective encourages the model to learn discriminative representations that implicitly encode distributional patterns of different subjects. As a result, samples with similar physiological characteristics tend to produce similar weighting distributions across source subjects.

It should be noted that, since CDPT is designed for DG rather than DA methods, the private encoder of the target domain does not exist. Therefore, the input to the AFS module is the fused feature S rather than the specific feature $\hat{S}$.

To train the adaptive feature selector, we define the following loss function:

$$\mathcal{L}_{ada} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{M} \hat{Y}_{i,j}\log(p_{i,j}), \tag{17}$$

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.

During inference, the output probability vector can be interpreted as measuring the degree of consistency between the target feature and each source domain in the learned feature space. Based on this, AFS selects the top-K source subjects whose representations are more compatible with the current input for subsequent feature modeling.

Although the selected private encoders are trained in a subject-specific manner, they are not restricted to modeling individual identities. Under the constraints of subject-level contrastive learning and orthogonality regularization, these encoders are encouraged to capture diverse and complementary variations across subjects, thereby spanning multiple representation subspaces. Each private encoder can thus be viewed as a specialized nonlinear transformation that emphasizes a particular type of subject-dependent variation. By applying multiple selected encoders to the target feature, the model projects the input into different subspaces and extracts complementary heterogeneous representations, which helps compensate for the absence of a target-specific encoder in the DG setting.

5) 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 $\mathcal{C}^C$ and the domain-specific classifier $\mathcal{C}^S$. In conjunction with the weight derived by the AFS module, we extract the Top-K weights from the output vector of the selector and use them to weight the prediction results of the private classifiers.

$$Y_i^P = \sum_{j=1}^{K} p_{i,j}\mathcal{C}_j^S(\tilde{S}_i + \hat{S}_i; \theta_j^{\mathcal{C}^S}), \tag{18}$$

$$Y^C = \mathcal{C}^C(\hat{S}; \theta^{\mathcal{C}^C}), \tag{19}$$

where Y_i^P and Y^C denote the results of the 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:

$$Y_i = (1-\delta)Y_i^P + \delta Y^C. \tag{20}$$

Accordingly, the classification process can be constrained by

$$\mathcal{L}_{cls} = -\frac{1}{M}\sum_{i=1}^{M} Y_i log(\tilde{Y}_i), \tag{21}$$

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

6) Loss constraint of CDPT: The overall objective of the proposed CDPT contains several types of constraints as follows:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1\mathcal{L}_{diff} + \lambda_2\mathcal{L}_{rec} + \lambda_3\mathcal{L}_{cl} + \lambda_4\mathcal{L}_{ada}, \tag{22}$$

where λ_1, λ_2, ..., and λ_4 serve as trade-off coefficients that regulate the balance among the constraints.

D. Algorithm for CDPT

For a better understanding of the process of CDPT training and inference, Algorithm 1 and Algorithm 2 summarize the procedure of the proposed CDPT framework for cross-subject PDs detection. The parameter details of each module in the CDPT framework, as well as the analysis of spatial and temporal complexity, can be found in the supplementary material.

IV. IMPLEMENTATION AND EXPERIMENTS

A. Hardware–software System

Fig. 5 illustrates the integrated hardware–software design and experimental setup. In Fig. 5(a), we develop a wear-

Algorithm 1 Training Procedure of CDPT

Input: Labeled sources $\mathbb{D}^S = \{\mathbb{D}^i\}_{i=1}^M$; epochs epc; temperature τ; top-K; weights $\lambda_{1\ldots4}$; fusion factor δ; optimizer Adam
Output: The trained framework of CDPT.

Initialize Θ randomly
for epoch $= 1$ to epc **do**
  $F_1, \ldots, F_N \leftarrow \mathrm{SCNN}(X_1, \ldots, X_N; \theta^{\mathrm{SCNN}})$ ▷ temporal encoding
  $\hat{F}_1, \ldots, \hat{F}_N \leftarrow \mathrm{CMAI}(F_1, \ldots, F_N; \theta^{\mathrm{CMAI}})$ ▷ cross-modal interaction and fusion
  $S = Concat(\hat{F}_1, \ldots, \hat{F}_N)$ ▷ concatenation
  $\hat{S} \leftarrow \mathcal{E}^C(S; \theta^C), \quad \tilde{S} \leftarrow \mathcal{E}_i^P(S; \theta_i^P)$ ▷ disentangle shared/private features
  $P = \{p_{i,1}, \ldots, p_{i,M}\} \leftarrow \mathrm{Softmax}(\mathrm{Linear}(S; \theta^{\mathrm{AFS}}))$ ▷ adaptive source similarity
  $Y^P \leftarrow \sum_{j \in \mathrm{TopK}(P)} p_{i,j}\, \mathcal{C}_j^S(\tilde{S} + \hat{S})$ ▷ weighted private (Top-K)
  $Y^C \leftarrow \mathcal{C}^C(\hat{S}; \theta^C)$ ▷ domain-shared prediction
  $\hat{Y} \leftarrow (1-\delta)\, Y^P + \delta\, Y^C$ ▷ fusion of predictions
  Compute $\mathcal{L}_{diff}, \mathcal{L}_{rec}, \mathcal{L}_{cl}, \mathcal{L}_{ada}, \mathcal{L}_{cls}$ ▷ by Eqs. (13), (14), (15), (17), and (21)
  Compute weighted loss $\mathcal{L}$ ▷ by Eq. (22)
  $\Theta \leftarrow \mathrm{Adam}(\Theta, \nabla_\Theta \mathcal{L})$
end for
return The trained framework of CDPT

Algorithm 2 Inference Procedure of CDPT

Input: Trained parameters Θ; test signals $\{X_1, \ldots, X_N\}$; fusion factor δ; Top-K
Output: Predicted label Y.

$F_1, \ldots, F_N \leftarrow \mathrm{SCNN}(X_1, \ldots, X_N; \theta^{\mathrm{SCNN}})$ ▷ temporal encoding for each modality
$\hat{F}_1, \ldots, \hat{F}_N \leftarrow \mathrm{CMAI}(F_1, \ldots, F_N; \theta^{\mathrm{CMAI}})$ ▷ cross-modal interaction and fusion
$S = Concat(\hat{F}_1, \ldots, \hat{F}_N)$ ▷ concatenation
$\hat{S} \leftarrow \mathcal{E}^C(S; \theta^C)$ ▷ domain-shared feature (shared encoder)
$P = \{p_1, \ldots, p_M\} \leftarrow \mathrm{Softmax}(\mathrm{Linear}(S; \theta^{\mathrm{AFS}}))$ ▷ adaptive source similarity
$Y^P \leftarrow \sum_{j \in \mathrm{TopK}(P)} p_j\, \mathcal{C}_j^S\big(\mathcal{E}_j^P(S; \theta_j^P) + \hat{S}\,; \theta_j^{\mathcal{C}^S}\big)$ ▷ weighted private prediction over Top-K
$Y^C \leftarrow \mathcal{C}^C(\hat{S}; \theta^{\mathcal{C}^C})$ ▷ domain-shared prediction
$Y \leftarrow (1-\delta)\, Y^P + \delta\, Y^C$ ▷ fusion of predictions
return Y

able 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. In Fig. 5(b), 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* 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 [45] with informed consent. As shown in Fig. 5(c), the patient group data are collected in the psychiatry department of Guangdong Second Provincial General 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 are recruited from our university and screened for mental stability using the SDS, SCL-90, and SAS scales. The control cohort includes not only students but also faculty, administrative staff, and cleaning staff, thereby diversifying age, occupation, and lifestyle backgrounds. Each participant first learns about the functionality and usage of the *MentalCare* system before wearing a custom wristband while resting. It should be noted that, to minimize the influence of medication on physiological signals, data were collected when patients had not taken any medication on the day of acquisition. During the 10-minute recording, participants are asked to remain still to minimize motion artifacts, while five physiological signals, namely PPG, IR, Red, GSR, and SKT, are continuously collected at sampling rates of 400 Hz, 400 Hz, 400 Hz, 200 Hz, and 100 Hz, respectively.

C. Baselines

We select representative DG-based methods, including the latest disentangled representation approaches, as competitive baselines to evaluate the performance superiority of CDPT.

- **SVM** [46] classifies samples by learning a maximum-margin decision boundary, where basic statistical descriptors from each physiological modality are used as input features.
- **CNN** [47] is a classical convolutional neural network architecture that extracts features through 1D convolution.
- **MSTGCN** [48] is a spatial-temporal graph generalization model that incorporates adversarial domain generalization to confuse the domain discriminator.
- **TSception** [49] is a novel spatial-temporal convolutional network that employs multi-scale spatial-temporal convolution kernels to capture multiple granularity features.

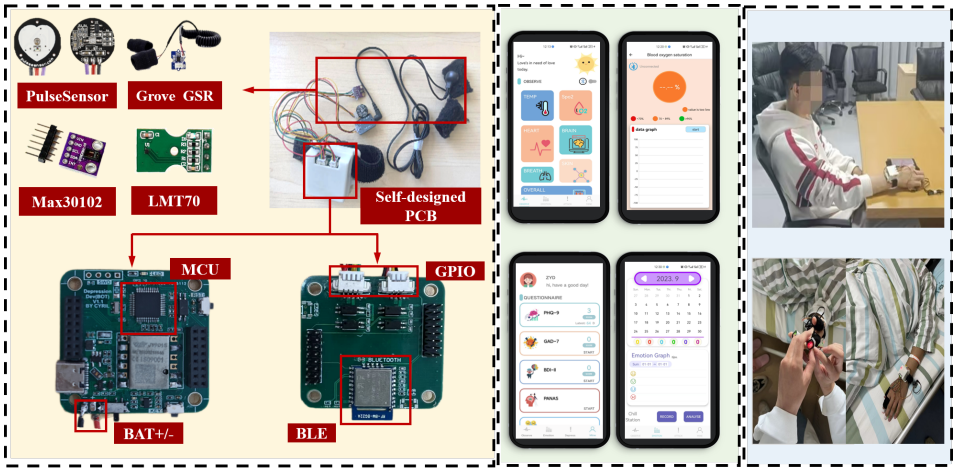


Fig. 5: The system prototype and experiment setup.

- **ManyDG** [50] is a novel DG model that utilizes mutual reconstruction and orthogonal projection to eliminate domain features.
- **MDNet** [24] 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** [51] 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** [29] is a novel causal disentanglement framework. It disentangles causal and domain-related representations through separate encoders.

D. Implementation Details

The experiments adopt the 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. Specifically, each test fold contains a single subject and therefore only one type of 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, recall, and F1 on a per-fold basis. Instead, we compute these metrics from a confusion matrix obtained by merging the predictions from all folds. Additionally, varying sample sizes across subjects can skew performance reports in favor of those with more windows. To avoid bias in window-level evaluation, we preserve the full 10-minute recording for each participant and apply a 1-second non-overlapping sliding window, so every subject contributes exactly 600 window samples. This setting differs from our previous conference version [52], where low-quality segments were discarded before evaluation. In this work, we retain all windows after preprocessing, which provides a more conservative and fair evaluation for continuous wearable sensing, since noisy or less stable segments may also be included. This protocol ensures that the window-level metrics are not dominated by subjects with more retained segments, and may partly explain the slight decrease in window-level accuracy.

All performance metrics used for evaluation are defined in Eq. (23)–Eq. (26). These metrics are primarily computed at the window-level by aggregating predictions over all test samples.

$$\text{Accuracy (Acc)} = \frac{\sum_{i=1}^{C} TP_i}{N}. \tag{23}$$

$$\text{Macro-Recall (Rec)} = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FN_i}. \tag{24}$$

$$\text{Macro-Precision (Pre)} = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FP_i}. \tag{25}$$

$$\text{Macro-F1 (F1)} = \frac{1}{C}\sum_{i=1}^{C}\frac{2TP_i}{2TP_i + FP_i + FN_i}. \tag{26}$$

In the above equations, C denotes the total number of classes, and N denotes the total number of test samples. For class i, TP_i, FP_i, and FN_i represent the numbers of true positives, false positives, and false negatives, respectively.

It is worth noting that all subjects in the dataset have identical recording durations and therefore an equal number of windows per subject. Consequently window-level evaluation does not introduce bias toward any specific individual. As a standardized complement to cross subject generalization assessment we report subject-level performance within the leave one subject out cross validation LOSO CV framework. The subject-level metric is computed by aggregating all window-wise predictions for a given subject via majority voting to obtain a single subject specific label after which overall accuracy is calculated across all subjects.

Window-level evaluation offers the advantage of sensitively capturing fine grained performance variations across temporal segments whereas subject-level evaluation eliminates redundant intra individual fluctuations by aggregating all windows

TABLE III: 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 the 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	Window-level				Window-level Overall				Subject-level Overall			
	Normal	Depression	Anxiety	Bipolar								
	Acc	Acc	Acc	Acc	Acc	Rec	Pre	F1	Acc	Rec	Pre	F1
SVM [46]	59.72%	78.03%	28.32%***	75.58%***	62.65%***	60.83%	62.09%	59.72%	72.97%	70.72%	73.32%	67.22%
CNN [47]	49.40%***	52.63%***	61.82%***	23.40%***	48.26%***	50.25%	49.78%	48.22%	51.35%	49.51%	49.46%	48.29%
MSTGCN [48]	55.13%***	43.44%***	52.42%***	47.66%***	49.51%***	49.46%	49.39%	49.24%	66.22%	67.03%	66.34%	66.33%
TSception [49]	80.56%	73.14%*	84.63%	81.25%	80.11%*	79.61%	79.57%	79.53%	87.84%	87.66%	88.10%	87.78%
BioDG [51]	80.86%	76.70%	83.09%	80.36%	80.41%*	80.26%	80.18%	80.19%	87.84%	87.66%	87.71%	87.65%
ManyDG [50]	72.97%**	77.54%	81.10%*	81.99%	78.45%***	78.41%	78.43%	78.35%	89.19%	89.65%	89.84%	89.40%
MDNet [24]	40.94%***	80.34%	**89.71%**	43.94%***	65.67%***	63.73%	66.60%	63.11%	74.32%	72.22%	79.13%	71.31%
CDDG [29]	73.29%***	77.34%	81.28%*	74.46%*	76.91%***	76.59%	76.57%	76.58%	89.19%	88.70%	89.39%	88.68%
CDPT(Ours)	**84.14%**	**81.22%**	85.33%	**82.09%**	**83.34%**	**83.20%**	**83.28%**	**83.20%**	**91.89%**	**92.59%**	**91.82%**	**92.04%**
Δ over Others	20.03% ↑	11.33% ↑	15.03% ↑	18.51% ↑	15.60% ↑	15.81% ↑	15.45% ↑	16.33% ↑	14.53% ↑	15.95% ↑	13.91% ↑	16.21% ↑
Δ over Best Baseline	3.28% ↑	0.88% ↑	4.38% ↓	0.10% ↑	2.93% ↑	2.94% ↑	3.10% ↑	3.01% ↑	2.70% ↑	2.94% ↑	1.98% ↑	2.64% ↑

for a given subject thereby providing a more robust reflection of the model's overall discriminative capacity for each independent individual. Unless explicitly stated otherwise accuracy in the subsequent experimental results is reported at the window-level by default.

All models are implemented in PyTorch. Before training, for models capable of handling asynchronous multimodal inputs (e.g., MDNet), no resampling is applied. For the remaining baselines, all modalities are upsampled to a unified frequency of 400 Hz to ensure consistent input formats. During training, shared hyperparameters across all models are kept consistent to ensure a fair comparison. Specifically, all models are optimized using the Adam optimizer with a learning rate of $1e{-}4$, a batch size of 100, and a maximum of 300 epochs. Early stopping is applied based on validation performance with a patience of 45 epochs. For model-specific hyperparameters, we adopt the configurations suggested in the original papers and further tune them on the validation set to better adapt to our dataset. For the proposed CDPT method, the temperature parameter τ is fixed at 0.1. To stabilize training and balance different loss terms, the weighting coefficients are set to $\lambda_1 = 5e{-}1$, $\lambda_2 = 1e{-}1$, $\lambda_3 = 5e{-}2$, and $\lambda_4 = 2e{-}2$, while the classification weight δ is set to 0.9.

V. Results

A. Cross-validation Performance with Statistical Test

As shown in Table III, we report the classification results on the entire dataset and for each category, with statistical significance evaluated using the Wilcoxon signed-rank test.

CDPT achieves the best overall performance at both the window-level and subject-level across all evaluation metrics. At the window-level, it attains an accuracy of 83.34%, recall of 83.20%, precision of 83.28%, and F1 score of 83.20%. The performance gains of CDPT at the window-level are statistically significant compared with all baseline methods, with all p-values below 0.05. At the subject-level, the corresponding metrics reach 91.89%, 92.59%, 91.82%, and 92.04%, respectively. In addition, BioDG and ManyDG achieve the second-best performance at the window-level and subject-level, respectively.

For window-level classification across individual categories, CDPT achieves the highest accuracy in healthy control, depression, and bipolar disorder groups, reaching 84.14%, 81.22%, and 82.09%, respectively. In the anxiety category, CDPT achieves the second-best performance, with an accuracy 4.38% lower than that of MDNet. However, MDNet shows limited capability in distinguishing healthy controls from bipolar disorder patients. Furthermore, statistical analysis indicates that CDPT achieves significant improvements in part of comparisons. It is worth noting that statistical testing is conducted within each category, resulting in a limited number of paired samples. Combined with substantial inter-subject variability, this makes it challenging to obtain consistent statistical significance across all categories. Finally, the cross-subject results show that SVM outperforms CNN and MSTGCN, demonstrating a certain degree of generalization ability; however, this ability largely depends on the effort devoted to handcrafted feature engineering.

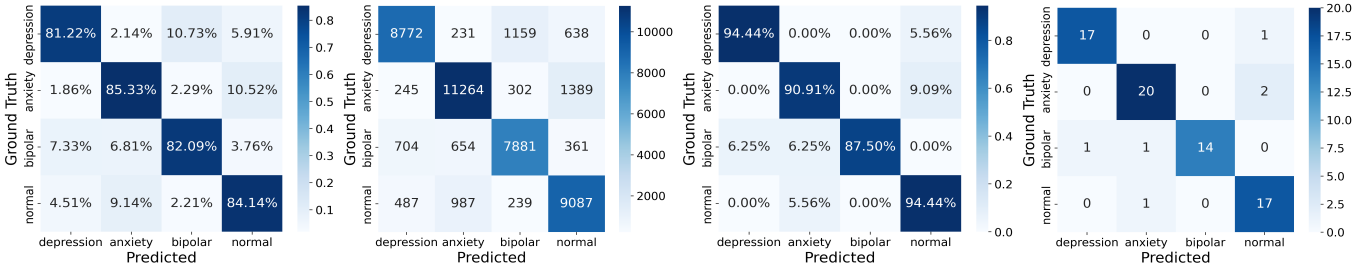
B. Confusion Matrix Analysis

After completing all folds of the LOSO-CV procedure, we aggregate the predicted labels and their corresponding ground-truth across folds to derive a global confusion matrix.

On one hand, Fig. 6(a)-(b) illustrate the window-level classification results, where the main diagonal entries represent the classification accuracy of each individual class, while the off-diagonal entries indicate misclassification patterns.

First, depression and bipolar disorder exhibit the highest bidirectional misclassification proportion among all category pairs. Specifically, 1159 depression samples (10.73%) are misclassified as bipolar, while 704 bipolar samples (7.33%) are predicted as depression. Collectively, these errors account for nearly one-fifth of the combined cohorts.

This confusion is consistent with overlapping clinical symptoms between the two disorders, such as persistent low mood, reduced energy, and cognitive impairments, which can make them challenging to distinguish. Second, healthy controls are most often confused with anxiety, accounting for 987 samples (9.14%), followed by depression at 487 samples (4.51%). Misclassification into bipolar is the least frequent, representing only 239 samples (2.21%). Finally, the healthy



(a) Window-level rate matrix (b) Window-level count matrix (c) Subject-level rate matrix (d) Subject-level count matrix

Fig. 6: Confusion matrices for multi-class PDs detection.

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

Model	Window-level Normal	Window-level Depression	Window-level Anxiety	Window-level Bipolar	Window-level Overall				Subject-level Overall			
	Acc	Acc	Acc	Acc	Acc	Rec	Pre	F1	Acc	Rec	Pre	F1
CDPT w/o SCNN	74.77%*	75.83%*	81.12%	72.74%*	76.48%***	76.12%	76.21%	76.15%	86.49%	85.92%	86.90%	86.11%
CDPT w/o CMAI	79.24%*	77.99%	84.81%	82.08%	81.21%*	81.04%	81.20%	81.08%	**91.89%**	**92.59%**	**91.82%**	**92.04%**
CDPT w/o SCD	80.05%*	77.76%	83.86%	78.96%*	80.39%*	80.16%	80.21%	80.15%	90.54%	90.26%	91.58%	90.55%
CDPT w/o AFS	82.90%*	81.15%	**85.86%**	79.02%*	81.83%*	81.61%	81.74%	81.59%	90.54%	90.44%	91.49%	90.71%
CDPT(Ours)	**84.14%**	**81.22%**	85.33%	**82.09%**	**83.34%**	**83.20%**	**83.28%**	**83.20%**	**91.89%**	**92.59%**	**91.82%**	**92.04%**
Δ over Others	4.90% ↑	3.04% ↑	1.42% ↑	3.90% ↑	3.36% ↑	3.47% ↑	3.44% ↑	3.46% ↑	2.03% ↑	2.79% ↑	1.37% ↑	2.19% ↑
Δ over Best Baseline	1.24% ↑	0.07% ↑	0.53% ↓	0.01% ↑	1.51% ↑	1.59% ↑	1.54% ↑	1.61% ↑	–	–	–	–

control category receives the largest number of misclassified samples, with a total of 2388 instances incorrectly assigned to this category. These include 638 depression samples (5.91%), 1389 anxiety samples (10.52%), and 361 bipolar samples (3.76%). This pattern indicates that our model has limitations in distinguishing healthy controls from the other categories.

On the other hand, Fig. 6(c)-(d) illustrate the subject-level classification results. A notable observation from the aforementioned findings is that the most frequent subject-level misclassification occurs between anxiety and healthy controls, with three subjects being incorrectly classified. Consistent with the window-level results, misclassification between anxiety and healthy controls also accounts for the largest number of misclassified samples and the second-highest misclassification rate. Specifically, 987 samples from healthy controls are misclassified as anxiety, while 1389 anxiety samples are misclassified as healthy controls.

C. Ablation Study

To further validate the effectiveness of each module in the proposed CDPT framework, we construct four ablation models: **w/o SCNN**, **w/o CMAI**, **w/o SCD**, and **w/o AFS**. These variants independently evaluate the contribution of stacked convolutional neural networks, attention mechanism, disentanglement representation learning, and adaptive feature selection to PDs detection. In the **w/o SCNN** model, we remove the three stacked convolutional blocks and only keep the linear layers for feature alignment. Meanwhile, in the **w/o SCD** model, we replace the SCD network with a two-layer multilayer perceptron to ensure that the framework retains a classifier. The results are reported in the Table IV.

For the window-level classification across individual categories, all ablation models exhibit varying degrees of performance degradation across different metrics. For instance, compared with the full CDPT model, the variants **w/o SCNN**, **w/o CMAI**, **w/o SCD**, and **w/o AFS** reduce overall accuracy by 6.86% ($p = 0.000034$), 2.13% ($p = 0.0049$), 2.95% ($p = 0.0033$), and 1.51% ($p = 0.045$), respectively. Statistical tests confirm that these performance drops are significant for all variants ($p < 0.05$). Regarding subject-level classification, we observe a distinct trend for **w/o CMAI**. Notably, although **w/o CMAI** yields lower window-level accuracy than the full CDPT framework, its subject-level performance remains identical to that of CDPT after aggregating window predictions via majority voting. For the remaining ablation models, a greater number of subjects are misclassified, leading to varying degrees of performance degradation at the subject-level.

There are some conclusions that can be drawn from the generalization performance reported in Table IV. One key observation is that removing the modality encoder leads to the most substantial performance decrease, as the absence of early stage modality-specific multi-scale temporal dynamics precludes the learning of discriminative representations in downstream layers. The **w/o SCD** and **w/o CMAI** variants also incur considerable declines. Removing the SCD network weakens the extraction of domain-invariant features across subjects, whereas eliminating CMAI undermines the effectiveness of adaptive multimodal interaction and fusion. In contrast, the **w/o AFS** model exhibits relatively minor degradation. This is because the adaptive feature selection mechanism acts as an auxiliary optimization component. It is designed to enhance the separation of private representations and to leverage source-domain private features for target-domain prediction. Nevertheless, statistical testing further validates the effectiveness of these optimizations.

Further analysis of per-category classification performance

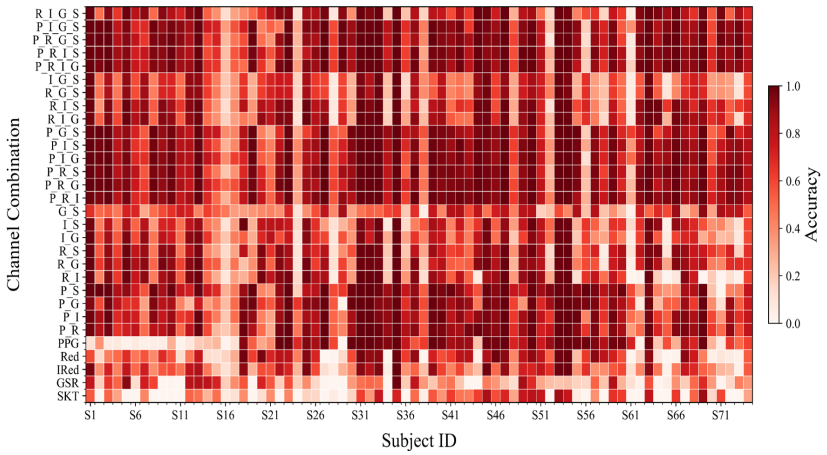


Fig. 7: The accuracy distribution results of different channel combinations for various subjects.

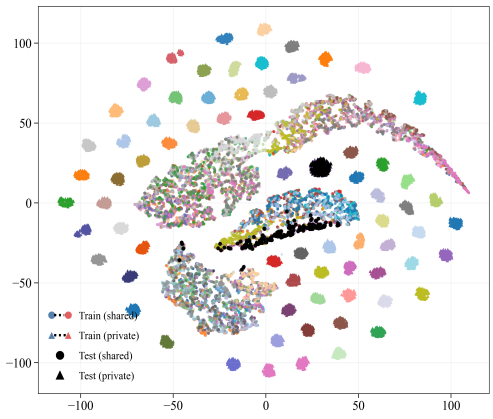


Fig. 8: Disentangled feature visualization.

reveals a pattern consistent with the observations in Table IV. The complete CDPT framework achieves the highest recognition accuracy for healthy controls, anxiety disorders, and bipolar disorders compared with the ablation models. It significantly outperforms the ablation variants in most comparisons for healthy controls and bipolar disorder ($p < 0.05$). For anxiety classification, the **w/o AFS** variant yields slightly better predictive results. However, this improvement is not statistically significant.

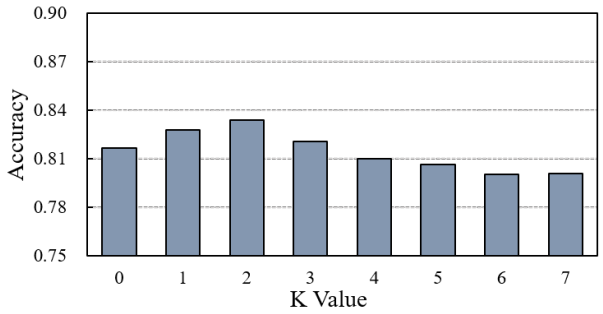


Fig. 9: Performance variation of the AFS Module with different K values.

D. Analysis of Influencing Factors

1) Comparison of Channel Combinations: Different physiological signals play complementary roles in PD detection. To investigate their contributions, we conduct channel ablation experiments. Detailed results are shown in Fig. 7, where the horizontal axis represents anonymized subject IDs and the vertical axis denotes different channel combinations. Starting from the full four-channel combination, the performance consistently degrades as the number of channels decreases, with accuracies of 79.88%, 75.85%, 67.57%, and 47.70%, respectively. The clear performance gap between multimodal and unimodal settings indicates that incorporating multiple physiological signals substantially improves PDs detection.

Beyond the overall performance, we highlight some additional observations. Due to substantial inter-subject physiological variability, the model does not achieve uniformly high accuracy across all subjects. Specifically, S16, S24, and S61 consistently exhibit poor performance across nearly all channel combinations (accuracy below 20%), whereas many other subjects achieve accuracies above 90%. In contrast, the model achieves superior performance for many other subjects, with accuracies exceeding 90%. Furthermore, under the unimodal setting, the PPG signal alone achieves an accuracy of 60.99%, outperforming other individual modalities (Red: 53.78%, IR: 54.12%, GSR: 36.09%, and SKT: 31.99%), and even approaching the performance of certain two-channel combinations without PPG (*e.g.*, R_G: 58.91%, I_S: 60.20%, and G_S: 53.97%). This suggests a stronger association between PPG signals and PD detection, which is consistent with findings reported in prior studies [17], [53].

2) Top-K Value: Positive transfer plays a critical role in improving cross-subject recognition in our framework. However, the number of source-domain subjects used to construct heterogeneous features (i.e., the Top-K value) affects the detection performance. To investigate this effect, we evaluate the model under different values of K. Fig. 9 presents the experimental results, where $K = 0$ indicates that no heterogeneous features are incorporated from the source domains. The results show that the recognition performance first improves and then declines as K increases. The highest accuracy of 83.34% is achieved at $K = 2$. As K further increases, the performance gradually decreases, although it remains higher than that at $K = 0$ for moderate values of K. When K reaches 5, the accuracy drops to 81.01%, which is lower than the 81.68% obtained at $K = 0$. This suggests that incorporating excessive heterogeneous features introduces substantial discrepancies between the source and target domains, thereby degrading performance. Therefore, a small K is more suitable for achieving effective positive transfer in this framework.

3) Hidden Size of Disentangle Encoders: In the CDPT framework, disentangled representation learning plays a crucial role in extracting domain-invariant physiological features. The hidden dimension (H) of the disentangled encoder de-

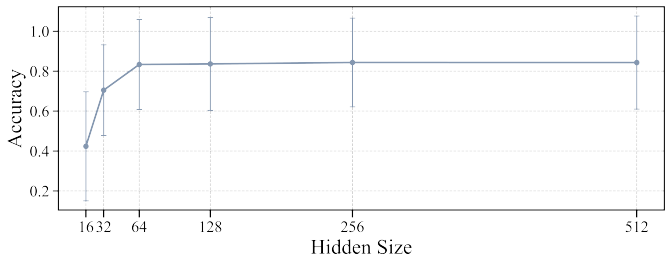


Fig. 10: The accuracy trend across different hidden sizes of disentangled encoders.

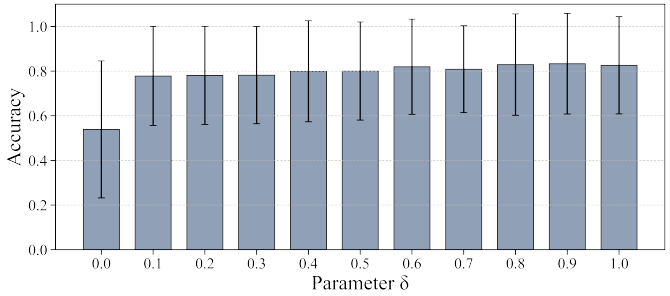


Fig. 11: The accuracy trend across different parameter δ.

termines the model capacity and complexity. We therefore evaluate the detection performance under different values of H. As shown in Fig. 10, when $H = 16$, the detection accuracy is 42.36%. As H increases, the performance improves significantly, reaching 70.48% at $H = 32$ and 83.34% at $H = 64$. Further increasing H leads to limited improvement; the accuracies at $H = 128$, 256, and 512 are 83.62%, 84.34%, and 84.31%, respectively. To balance model complexity and performance, we set $H = 64$ in the disentangled encoder.

4) Weighted Fusion Parameter δ: The parameter δ controls the relative contributions of the domain-shared and domain-specific classifiers, thereby affecting the overall classification performance. A larger δ places greater emphasis on the domain-shared output. We evaluate the performance of CDPT under different values of δ. As shown in Fig. 11, the highest accuracy of 83.34% is achieved at $\delta = 0.9$. As δ increases, the performance first improves and then gradually declines. The accuracy surpasses that of $\delta = 0$ when $\delta \geq 0.8$. When the prediction relies entirely on the domain-specific classifier ($\delta = 0$), the model yields the lowest accuracy of 53.90%. These results suggest that appropriately combining domain-shared and domain-specific features is beneficial for improving classification performance.

5) Different Number of Source Subjects: The number of training subjects significantly affects cross-subject recognition performance. Increasing the number of source subjects introduces greater inter-subject variability, which can improve robustness. However, excessive diversity may introduce inconsistencies that hinder the learning of stable domain-invariant features. We vary the number of labeled source subjects while maintaining a balanced class distribution. Specifically, the same number of participants is sampled from each of the four categories, ensuring a constant per-class sample size. Since the smallest class contains 16 participants, the source set can include at most 60 subjects while preserving at least one target subject under the cross-subject protocol.

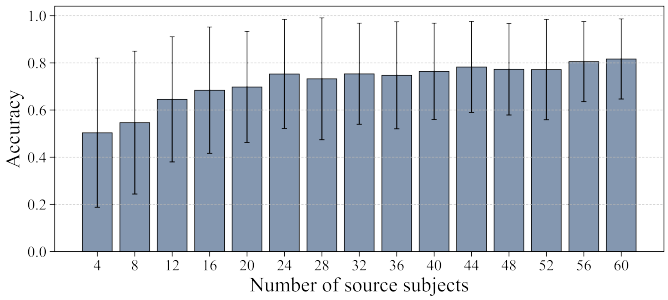


Fig. 12: The mean accuracy with different numbers of source subjects.

Fig. 12 shows the performance of CDPT under different source-set sizes. The target accuracy increases substantially as the number of source subjects grows from 4 to 24, rising from 50.39% to 75.31%. Further increasing the source size from 24 to 52 leads to moderate improvements with fluctuations, reaching 77.18%. When the number reaches 60, the accuracy further improves to 81.66%.

Taken together, the results indicate that increasing the number of source subjects generally improves cross-subject recognition. When the number of source subjects increases from 4 to 24, the accuracy rises sharply from 50.39% to 75.31%, suggesting that adding subjects at a small scale introduces rich physiological variability and helps the model learn more stable cross-subject representations. As the source set becomes larger, the performance continues to improve, but the increase rate is lower than that in the early-stage expansion from 4 to 24 subjects. Specifically, the accuracy reaches 77.18% with 52 source subjects and further increases to 81.66% with 60 source subjects, indicating that additional subjects can still provide useful information for cross-subject generalization. At the same time, newly added subjects may also contain physiological patterns that partially overlap with those already covered by the existing source subjects, so the performance improvement may not grow proportionally with the number of subjects. These results suggest that source-subject scale is an important factor for cross-subject PDs detection, and increasing the number of subjects can further enhance generalization while also introducing more complex inter-subject physiological variations.

E. Visualization and Explainability

1) Disentangled Feature Visualization: We visualize the learned representations using t-SNE by projecting them into a two-dimensional space. The first 73 subjects are used for training, and the last subject is used for testing. The model produces subject-level shared and private features, which are jointly visualized to evaluate disentanglement.

As shown in Fig. 8, the learned representations exhibit a clear separation between shared and private components. The shared features form dense clusters within a compact region, indicating stable patterns across subjects. In contrast, the

private features are distributed in a more dispersed subspace, forming multiple distinct clusters that reflect subject-specific variability. This observation suggests that the disentanglement objective effectively separates common physiological structures from subject-specific characteristics. Moreover, the private embeddings of the test subject show noticeable overlap with those of subject 65 (both diagnosed with bipolar disorder), indicating similar physiological patterns and supporting the effectiveness of the AFS mechanism.

2) Adaptive Attention Mechanism: To quantify the extent to which Adaptive Attention Aggregation in the CDPT framework depends on each physiological modality (PPG, IR, Red, GSR, SKT), we visualize the learned adaptive fusion weights, thereby characterizing both the average contribution and the stability of each modality. Specifically, we group the data by ground-truth within each class, and then select the three most accurate subjects (Top-3) as the targets for visualization. For each selected subject, we aggregate the adaptive fusion weights across all test samples and compute the mean and standard deviation (SD) for each modality. A blue solid curve represents the per-modality mean fusion weights for the subject. Two orange dashed contours depict the upper and lower bounds given by the mean plus one standard deviation and the mean minus one standard deviation.

Fig. 13 shows that across subjects, the model assigns higher adaptive fusion weights to PPG and the smallest weights to SKT. This pattern aligns with physiological knowledge. PPG and its spectral channels capture peripheral blood volume pulse as well as heart rate and heart rate variability (HRV), which are sensitive to autonomic shifts and therefore provide richer discriminative information for affective and psychiatric states. In contrast, SKT primarily reflects slow thermoregulatory processes, has lower temporal resolution, and is easily influenced by ambient temperature and baseline differences, which reduces its discriminative power over short time windows. At the group level, the depressive group shows a markedly stronger reliance on PPG, consistent with alterations in sympathetic and parasympathetic balance and cardiovascular dynamics that make pulse wave features more informative. This physiological finding is consistent with prior research [54], [17]. The bipolar disorder group exhibits elevated weights on both PPG and IR, indicating that the near-infrared channel and conventional PPG jointly capture complementary information related to hemoglobin absorption and pulse morphology that is particularly useful for this group. The anxiety group and the normal group display more balanced allocations across the five modalities, suggesting that their discriminative cues are distributed among multiple physiological channels and that the model integrates them rather than depending on a single source.

Overall, the attention pattern is consistent with modality-specific physiological interpretations, indicating that the adaptive fusion mechanism selectively leverages complementary signals.

Fig. 13: Modality Attention Aggregation Weight ψ. Group A: Depressive Disorder; Group B: Anxiety Disorder; Group C: Bipolar Disorder; Group D: Normals.

F. System-level Evaluation

1) CPU, Memory Load, and Running Time: 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 the testing process. 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. 14, the pipeline consists of two stages. During data loading, latency increases with input length: 80 ms at 3 minutes, 130 ms at 5 minutes, and 540 ms at 10 minutes, while the CPU utilization ratio remains low, ranging from 22% to 28%. The second stage includes preprocessing and prediction; preprocessing takes 200/340/900 ms for 3/5/10 minutes, which is much shorter than the prediction time of 1.76/2.87/4.81 s. Even so, 10 minutes of data are processed end-to-end in under 7 s, enabling deployment on an edge device. This stage is CPU intensive (94% to 100%) because both steps are compute-heavy. Memory increases with sequence length: 170 MB at 3 minutes, 185 MB at 5 minutes, and 203 MB at 10 minutes, remaining within on-device memory constraints for commodity smartphones.

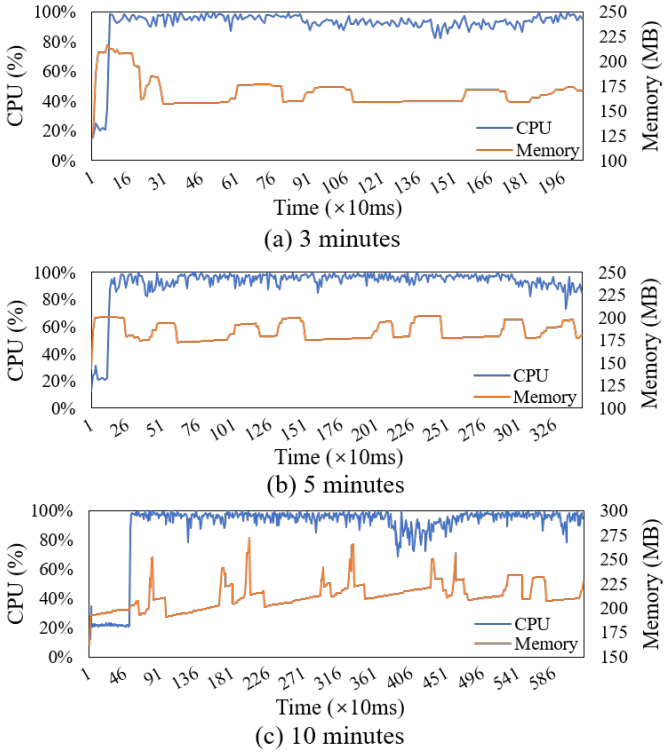


Fig. 14: CPU and memory utilization of mobile applications for predicting physiological signals with different durations.

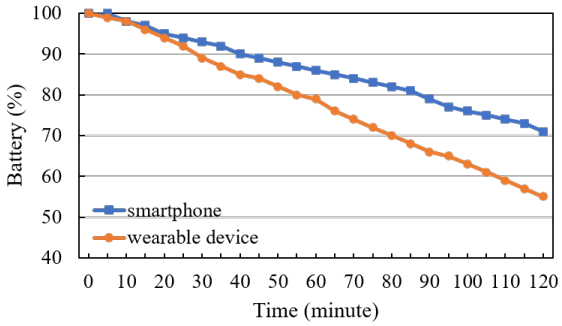


Fig. 15: Power trends of smartphone and wearable devices under continuous usage.

2) Battery Consumption: We assess the feasibility of *Mentalcare* for extended use by measuring the 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 2-hour session in which the wristband continuously acquires physiological signals and streams them to the smartphone for real-time visualization, with the screen kept on throughout. Fig. 15 shows power consumption for both devices, with an approximately linear decline. On average, the smartphone depletes 15% of its battery per hour, and the wearable 21%. Extrapolating indicates about five hours of continuous collection, meeting long-duration requirements.

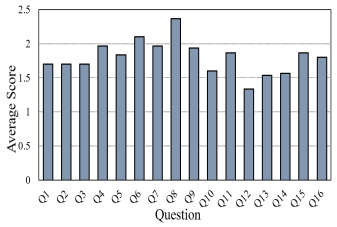


Fig. 16: Score of each question.

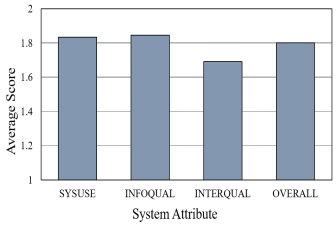


Fig. 17: Score of system attribute.

3) User Study: We evaluate usability using the Post-Study System Usability Questionnaire (PSSUQ) [55]. The instrument contains 16 items scored on a seven-point scale from "strongly agree" to "strongly disagree." Items $1 \sim 6$ measure System Usefulness (SYSUSE), items $7 \sim 12$ measure Information Quality (INFOQUAL), and items $13 \sim 16$ measure Interface Quality (INTERQUAL). Lower scores indicate better perceived usability.

A total of 30 participants (20 males and 10 females) took part in the evaluation, aged 19–56 years (mean = 33.57, SD = 13.04). Among them, 13 were recruited from the university community and 17 from the general public. Within the university group, 9 participants had computer-related backgrounds, including 4 laboratory members who were not involved in the system development.

As shown in Fig. 16 and Fig. 17, the overall mean score is 1.8. Across the three PSSUQ system attributes, the system attains mean scores of 1.83 in SYSUSE, 1.84 in INFOQUAL, and 1.69 in INTERQUAL. From the per-item results, question Q12 (The information on the system screen is well organized) records the lowest score of 1.33, indicating that users are highly satisfied with the clarity and organization of the system interface. In contrast, question Q8 (Easy and quick error recovery) received the highest score of 2.37, suggesting that the system can be further improved in supporting efficient error recovery. Overall, the results indicate favorable user satisfaction with the system.

VI. Discussion and Future Works

In this section, we discuss key aspects of the study to provide a comprehensive evaluation of its strengths and limitations, and outline directions for future research.

A. MentalCare System for PDs Detection

The subject-independent design enables reliable cross-subject operation without personalization, supporting generalization and scalability. However, many existing studies either overlook cross-subject settings or assume access to target-domain information during training. A large body of work adopts domain adaptation, which requires target data or meta-calibration and thus limits deployment when such information is unavailable. Moreover, most prior research focus on binary classification (*e.g.*, depression/healthy), ignoring the clinical heterogeneity of PDs and limiting real-world applicability. In contrast, we consider a four-class setting, including depression, anxiety, bipolar disorder, and healthy controls.

TABLE V: Performance comparison for 3-class classification under LOSO-CV after removing site-related confounding factors.

Model	Window-level			Window-level Overall				Subject-level Overall			
	Depression	Anxiety	Bipolar								
	Acc	Acc	Acc	Acc	Rec	Pre	F1	Acc	Rec	Pre	F1
BioDG	82.28%	93.11%	83.35%	86.85%	86.25%	86.19%	86.22%	94.64%	93.98%	94.27%	94.07%
ManyDG	83.56%	94.05%	81.86%	87.20%	86.49%	86.53%	86.51%	94.64%	93.98%	94.27%	94.07%
CDPT	**85.44%**	**96.17%**	**84.15%**	**89.32%**	**90.66%**	**88.37%**	**88.56%**	**96.43%**	**95.83%**	**96.67%**	**96.02%**

To address inter-subject variability in the DG setting, we propose a novel framework termed CDPT based on contrastive disentanglement and positive transfer. Under the LOSO-CV protocol on a four-class dataset with 74 participants, CDPT achieves accuracies of 83.34% and 91.89% at the window-level and subject-level, respectively. Statistical tests further confirm that CDPT significantly outperforms strong DG baselines, with all p-values below 0.05.

At the system level, we integrate CDPT into *MentalCare*. A wristband streams multimodal signals to a smartphone, where preprocessing and inference are performed on-device. System evaluations show that *MentalCare* processes a 10-minute recording in under 7 seconds on a Snapdragon 8 Gen 2 device, enabling near real-time feedback. The subject-independent pipeline operates reliably within the resource constraints of mobile sensing. Nevertheless, we acknowledge a scalability limitation when the number of training subjects becomes very large. Since CDPT employs subject-specific private encoders and classifiers, memory consumption grows roughly linearly with the number of source subjects. Thus, the current implementation is better suited to moderate-scale cohorts than massive populations. A promising direction is to introduce an aggregation mechanism before the subject-private encoders, allowing subjects with similar physiological patterns to share compact prototypes or grouped transformations. In addition, although the system is designed for continuous and long-term monitoring, the current experiments use 10-minute recordings per participant. Such short-duration data may not fully capture long-term temporal variations within individuals [56]. Future work will investigate longitudinal data collection across multiple sessions to evaluate cross-session generalization and robustness to temporal variability in physiological patterns.

B. Dataset

A recent survey of 88 psychiatric datasets [57] highlights a lack of diversity: only 17 include multiple PDs, and just two cover three disorders. Furthermore, most physiological datasets rely solely on single-modality EEG signals. Against this background, we collected multimodal physiological data via wearable devices from 74 participants across four mutually exclusive cohorts: depression, anxiety, bipolar disorder, and healthy controls. Our sample size exceeds several widely used benchmarks, such as MODMA [58] (53 subjects) and WESAD [59] (15 subjects). This multi-cohort setting is highly challenging, as distributional overlap and increased inter-subject variability among diagnostic groups typically degrade detection accuracy.

Beyond diagnostic heterogeneity, our protocol may introduce environmental and pharmacological confounds. First, restricting data collection to hospital and campus settings means models might capture environment-related differences (e.g., ambient noise or activity patterns) rather than pure physiological responses. However, such confounds are inherently difficult to eliminate in realistic mobile health applications. We conduct an experiment to evaluate the impact of removing site-related confounding factors on model performance. The LOSO-CV results of the proposed CDPT framework and the baselines BioDG and ManyDG for three-category mental disorder classification are listed in Table V. The experimental results show that CDPT attains accuracies of 89.32% and 96.43% at the window-level and subject-level, respectively, and outperforms BioDG and ManyDG across all disorder categories. Besides, all models achieve higher performance in the three-class setting than in the four-class setting reported in Table III. This stability indicates the model captures discriminative physiological patterns rather than site-specific noise, which would otherwise degrade single-site evaluation.

Second, medication also poses significant confounding effects on physiological signals. Although we required patients to be in an unmedicated state on the day of acquisition, at most a 24-hour window is insufficient to wash out chronic medication-induced physiological changes (e.g., altered HRV from beta-blockers) and may even introduce short-term withdrawal effects. To completely eliminate these chronic effects, a multi-week wash-out period is typically required. However, such a prolonged cessation of treatment is deemed clinically unfeasible due to the severe ethical risks to patient stability. Therefore, our real-world dataset inevitably contains persistent medication footprints.

Finally, aligning with similar studies [60], we excluded subjects with physical comorbidities that cause pronounced vital sign shifts. For instance, cardiovascular disease and anxiety both present tachycardia and reduced PPG amplitude, producing morphological overlap that could severely confound classifiers. Future work will broaden sociodemographic coverage and include participants with common physical comorbidities to assess how well *MentalCare* distinguishes PDs from non-psychiatric physiological alterations.

VII. Conclusion

In this work, we present *MentalCare*, an all-in-one PDs system unifying an AI model, wearable sensing hardware, and edge mobile software. At the algorithmic level, we introduce a DG-based CDPT that effectively exploits positive transfer and contrastive disentanglement to enhance the performance of

target-subject diagnostics. The extensive experiments demonstrate that the proposed framework achieves significant improvements over SOTA deep learning models and disentanglement representation learning approaches. At the system level, we evaluate the key performance indicators of the algorithm when deployed on mobile devices, thereby demonstrating the practicality and feasibility of applying the proposed framework on edge devices.

REFERENCES

[1] W. H. Organization, "World mental health report: Transforming mental health for all," Geneva, 2022. [Online]. Available: https://www.who.int/publications/i/item/9789240049338

[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*, 2024, pp. 10 373–10 381.

[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, no. 20, 2024, pp. 22 231–22 239.

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

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

[8] V. Tejaswini, K. Sathya Babu, and B. Sahoo, "Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 1, pp. 1–20, 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] T. He, J. Chen, C. Lin, and W. Wang, "Mobile phone-based digital biomarkers empowered by knowledge distillation for diagnosis of parkinson's disease," *IEEE TMC*, pp. 1–16, 2025.

[12] 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.

[13] 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.

[14] 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 IEEE BIBM*. IEEE, 2024, pp. 2680–2687.

[15] L. Dolezal, "Shame anxiety, stigma and clinical encounters," *Journal of evaluation in clinical practice*, vol. 28, no. 5, pp. 854–860, 2022.

[16] R. I. Lanyon and R. E. Wershba, "The effect of underreporting response bias on the assessment of psychopathology." *Psychological Assessment*, vol. 25, no. 2, p. 331, 2013.

[17] 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.

[18] K. Shao, R. Wang, Y. Hao, L. Hu, M. Chen, and H. Arno Jacobsen, "Multimodal physiological signals representation learning via multiscale contrasting for depression recognition," in *Proceedings of the ACM MM*, 2024, pp. 5692–5701.

[19] Q. Zhao, L. Zhang, H. Zhang, H. Jiang, K. Cui, Z. Wu, J. Liu, M. Zhao, F. Tian, and B. Hu, "Lsnn model: a lightweight spiking neural network-based depression classification model for wearable eeg sensors," *IEEE TMC*, 2025.

[20] Y. Guo, T. Zhang, and W. Huang, "Emotion recognition based on physiological signals multi-head attention contrastive learning," in *Proceedings of the IEEE BIBM*. IEEE, 2023, pp. 1929–1934.

[21] 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.

[22] 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.

[23] 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.

[24] 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*, 2024, pp. 3069–3077.

[25] W. Kuang, Y. Zhang, Y. Cheng, K. Wu, and Y. Zou, "Generalizable psychiatric screening from wearable: Causal disentanglement of multi-channel biosignals with interpretability analysis," in *Proceedings of the IEEE BIBM*. IEEE, 2025, pp. 1–6.

[26] 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.

[27] P. Singhal, R. Walambe, S. Ramanna, and K. Kotecha, "Domain adaptation: challenges, methods, datasets, and applications," *IEEE Access*, vol. 11, pp. 6973–7020, 2023.

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

[29] 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.

[30] Y. Yao and G. Doretto, "Boosting for transfer learning with multiple sources," in *Proceedings of the IEEE/CVF CVPR*. IEEE, 2010, pp. 1855–1862.

[31] 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 *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*, 2025, pp. 1–10.

[32] 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 Transactions on Embedded Computing Systems*, vol. 21, no. 6, pp. 1–22, 2022.

[33] 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. 22 788–22 802, 2023.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

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

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

[40] 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*, 2024, pp. 5712–5721.

[41] 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.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[43] X. Wang, H. Chen, S. Tang, Z. Wu, and W. Zhu, "Disentangled representation learning," *IEEE TPAMI*, vol. 46, no. 12, pp. 9677–9696, 2024.

[44] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *NeurIPS*, vol. 33, pp. 18 661–18 673, 2020.

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

[46] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[47] 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.

[48] 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.

[49] 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.

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

[51] 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.

[52] Y. Zhang, W. Kuang, Z. Huang, C. Fan, and Y. Zou, "Mentalcare: Contrastive disentanglement and positive transfer for psychiatric disorder detection with a wristband–smartphone system," in *PerCom*. IEEE, 2026, pp. 1–10.

[53] 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 *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 5305–5315.

[54] R. Hartmann, F. M. Schmidt, C. Sander, and U. Hegerl, "Heart rate variability as indicator of clinical state in depression," *Frontiers in Psychiatry*, vol. 9, p. 735, 2019.

[55] J. R. Lewis, "Psychometric evaluation of the post-study system usability questionnaire: The pssuq," in *Proceedings of the Human Factors Society Annual Meeting*, vol. 36, no. 16. Sage Publications Sage CA: Los Angeles, CA, 1992, pp. 1259–1260.

[56] D. Duan, H. Yang, G. Lan, T. Li, X. Jia, and W. Xu, "Emgsense: A low-effort self-supervised domain adaptation framework for emg sensing," in *PerCom*. IEEE, 2023, pp. 160–170.

[57] A. Mandal, P. K. Adhikary, H. Arnaout, I. Gurevych, and T. Chakraborty, "A comprehensive review of datasets for clinical mental health ai systems," *arXiv preprint arXiv:2508.09809*, 2025.

[58] H. Cai, Z. Yuan, Y. Gao, S. Sun, N. Li, F. Tian, H. Xiao, J. Li, Z. Yang, X. Li *et al.*, "A multi-modal open dataset for mental-disorder analysis," *Scientific Data*, vol. 9, no. 1, p. 178, 2022.

[59] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *Proceedings of the ICMI*, 2018, p. 400–408.

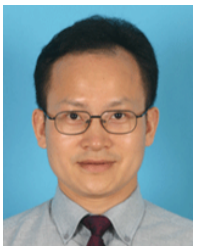
[60] Z. Zheng, Y. Liang, R. Lyu, J. Bao, Y. Huang, A. Zhou, H. Ma, J. Wang, X. Meng, C. Shao *et al.*, "Bp3: Improving cuff-less blood pressure monitoring performance by fusing mmwave pulse wave sensing and physiological factors," in *Proceedings of the ACM Sensys*, 2024, pp. 730–743.

Yufei Zhang (Student Member, IEEE) is currently pursuing the PhD degree in the College of Computer Science and Software Engineering, Shenzhen University. His research interests include mobile computing, affective computing and spatial-temporal data analysis. He was honored with the Best Paper Nomination at the IEEE PerCom 2026.

Wenting Kuang is currently working toward the master's degree with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. Her research interests covers affective computing and intelligent sensing.

Kaishun Wu (Fellow, IEEE) received the PhD degree from the Hong Kong University of Science and Technology, Hong Kong, in 2011. He is currently a professor in information hub with the Hong Kong University of Science and Technology (Guangzhou). His research interests include wireless communications and mobile computing. He won several best paper awards of international conferences, such as IEEE Globecom 2012 and IEEE MASS 2014.

Zite Huang is currently pursuing a bachelor's degree in the College of Computer Science and Software Engineering at Shenzhen University, China. His research interests include intelligent sensing and affective computing.



Changhe Fan received the PhD degree in Psychiatry from the Institute of Mental Health, Xiangya School of Medicine, Central South University, in 2002. is currently a chief physician from the Department of Psychiartry, Guangdong Second Provincial General Hospital. His research interests include biological psychiatry and behavioral medicine.

Yongpan Zou (Member, IEEE) received the PhD degree from the Department of Computer Science and Engineering (CSE), Hong Kong University of Science and Technology, in 2017. He is currently an associate professor with the College of Computer Science and Software Engineering, Shenzhen University. His research interests include ubiquitous sensing, mobile computing, and human-computer interaction.