# Highlights

## DrowsyDG-Phys: Generalizable Driver Drowsiness Estimation in Conditional Automated Vehicles Using Physiological Signals

Jiyao Wang,Wenbo Li,Zhenyu Wang,Suzan Ayas,Birsen Donmez,Dengbo He,Kaishun Wu

- A generalizable driver drowsiness detection framework based on physiological signals is proposed
- Time and frequency domain features of signals are explicitly extracted and fused
- Three loss functions are introduced to promote generalizability and robustness of the model
- The first multi-domain generalization benchmark for driver drowsiness detection

# DrowsyDG-Phys: Generalizable Driver Drowsiness Estimation in Conditional Automated Vehicles Using Physiological Signals

Jiyao Wang[a], Wenbo Li[b], Zhenyu Wang[a], Suzan Ayas[c], Birsen Donmez[c], Dengbo He[a,*] and Kaishun Wu[d]

[a]Systems Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, Guangdong, China
[b]College of Electrical Engineering, Sichuan University, Chengdu, Sichuan, China
[c]Mechanical and Industrial Engineering Department, University of Toronto, Toronto, Ontario, Canada
[d]Information Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, Guangdong, China

ARTICLE INFO

*Keywords*:
Driver drowsiness estimation
Physiological signals
Domain generalization
Neural network

ABSTRACT

Driver drowsiness is one of the leading causes of crashes, injuries, and fatalities on the road. Traditional drowsiness detection models relied on manually extracted physiological features processed through machine learning algorithms. However, these methods lacked flexibility and robustness across diverse real-world conditions. Although recent advances in deep learning have improved detection accuracy through automated feature extraction based on larger learnable parameter space, the generalization of existing models is still limited due to domain shifts. In this study, we proposed **DrowsyDG-Phys**, a novel domain generalization (DG) framework for driver drowsiness detection using three physiological signals (i.e., electrocardiogram, electrodermal activity, and respiration signals) that can be measured by in-vehicle or wearable sensors. Our approach introduced a backbone network for explicit time and frequency domain feature learning. In addition, our approach integrated three novel loss functions: a prior knowledge-based contrastive regularization for robustness, a feature centralization loss to promote generalization in heterogeneities, and a novel loss function to align drowsiness assessment criteria. Finally, we established a multi-source DG benchmark and evaluated our model on three existing datasets and a self-collected dataset involving 60 participants in a simulated SAE Level-3 driving scenario. Our proposed DrowsyDG-Phys achieves 78.5% accuracy on the DG protocol, as well as 88.4% accuracy on the cross-subject protocol. Experimental results demonstrated that DrowsyDG-Phys outperformed baseline methods, and improved generalization and robustness of physiological signal-based drowsiness monitoring.

## 1. Introduction

Driver drowsiness in safety-critical environments can lead to fatal consequences. In the United States alone, each year, an estimated 328,000 crashes are attributed to drowsy drivers, resulting in approximately 109,000 injuries and 6,400 fatalities (Tefft, 2012). Drowsiness in driving is defined as a state of impaired consciousness where a driver is more inclined to sleep than to stay awake (Slater, 2008). This state can lead to slower reaction times, poor decision-making abilities, and overall diminished responsiveness (Ashraf, Hur, Shafiq and Park, 2019; Khushaba, Kodagoda, Lal and Dissanayake, 2010). With advancements in technology, data-driven methods have been developed to detect drowsiness using images, physiological signals, and vehicle sensors (Saleem, Siddiqui, Raza, Rustam, Dudley and Ashraf, 2023; Wang, Yang, Wang, Wei, Wang, He and Wu, 2024e). However, not all measures are suitable for deployment in vehicles, especially in vehicles with driving automation, where drivers no longer need to continuously control the vehicle, making vehicle-based measures less relevant. Further, associated privacy concerns can also decrease the acceptance of image-based drowsiness detection. Thus, monitoring drowsiness through physiological signals, due to its comparatively less private data collection compared to in-car cameras, has gained increasing attention from researchers (Kakhi, Jagatheesaperumal, Khosravi, Alizadehsani and Acharya, 2024).

Traditional physiological-based models for driver drowsiness monitoring have mainly relied on the manual extraction of physiological features, such as low frequency (LF) and heart rate (HR) from electrocardiogram (ECG) signals. These models often utilized machine learning (ML) techniques, including LightGBM and SVM (Cheon and Kang, 2017; Chowdhury, Shankaran, Kavakli and Haque, 2018; Zhou, Alsaid, Blommer, Curry, Swaminathan,

*Corresponding author, email address: dengbohe@hkust-gz.edu.cn
ORCID(s): 0000-0002-0743-0121 (J. Wang); 0000-0002-8257-5806 (D. He)

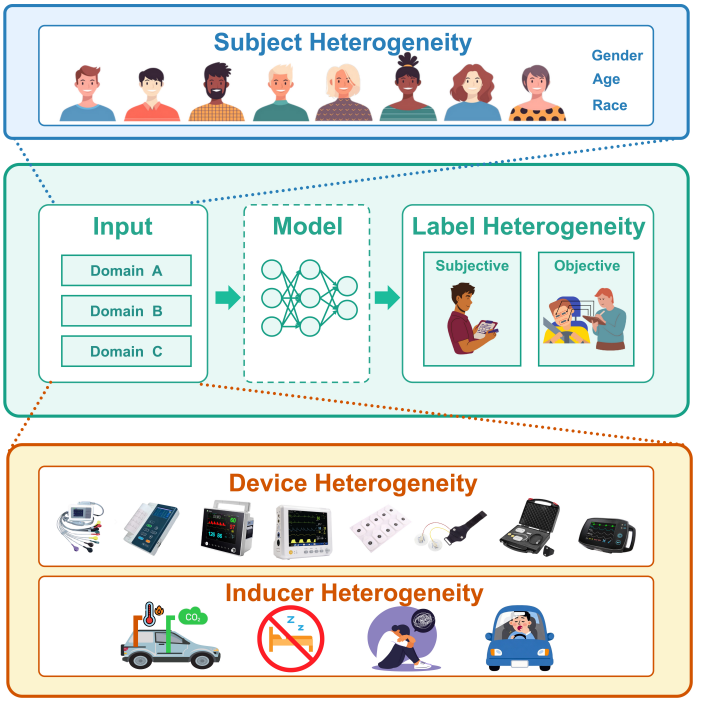


**Figure 1:** Heterogeneities when training drowsiness detection model with multi-source domain data.

Kochhar, Talamonti and Tijerina, 2022). While past research has achieved satisfactory accuracy, the manual extraction process of these low-level features can be computationally intensive and may lead to a loss of information contained in the raw signals. With the development of deep learning (DL), an increasing number of works have attempted advanced DL models based on raw physiological signals without the need for hand-crafted features to monitor drowsiness (Hultman, Johansson, Lindqvist and Ahlström, 2021; Fan, Peng, Peng, Zhang, Wu and Kwong, 2021; Alguindigue, Singh, Narayan and Samuel, 2024). The DL-based approach, compared to traditional methods, removes the need for time-consuming manual feature selection and offers an automated end-to-end solution. Additionally, the DL approach demonstrates more robust performance in complex environments with diverse populations (Zhang, Bengio, Hardt, Recht and Vinyals, 2022).

Although DL methods have shown promising results in detecting driver drowsiness, their effectiveness in real-world situations remains limited (as shown in Fig. (1)) (Wang, Ayas, Zhang, Wen, He and Donmez, 2025a). Many previous detection methods (Cheon and Kang, 2017; Hultman et al., 2021) were trained and tested on the within-dataset protocol, which means all data is from one specific experiment where drowsiness was induced by specific procedures and assessed by particular criteria. This within-dataset protocol cannot guarantee the generalizability of the model when applied in new environments (Kim, Han, Jeong and Lee, 2025). For example, in real life, various drowsiness inducers, such as low arousal, mental fatigue, and sleep deprivation (Ayas, Donmez and Tang, 2024a), may lead to different physiological responses. Besides, there are heterogeneities in how drowsiness was assessed across studies. Several criteria were adopted for driver drowsiness assessment (Kundinger, Mayr and Riener, 2020a), including: (1) Subjective assessments, which relied on self-assessment questionnaires, such as the Karolinska Sleepiness Scale (KSS) (Kaida, Takahashi, Åkerstedt, Nakata, Otsuka, Haratani and Fukasawa, 2006), as recognized by the European Union (Ahlström and Anund, 2024); and (2) Objective indicators, such as blinking and nodding, which are adopted in China (Bao and Xu, 2024). In addition, individual diversity and differences in collection equipment may also introduce sources of heterogeneity. These heterogeneities, which can also be described as the domain shift between different datasets, bring challenges in achieving generalizable drowsiness monitoring (Wang et al., 2025a). To verify this, we conduct a cross-dataset evaluation. As shown in Table 3, none of the methods worked well when being tested on an unseen dataset.

Recent studies have attempted to tackle this issue by jointly training the model on multiple datasets with domain generalization (DG) techniques (Kim et al., 2025; Ballas and Diou, 2024). However, existing DG methods for drowsiness detection usually took the electroencephalography (EEG) as the input. Compared to EEG, electrocardiogram

(ECG), electrodermal activity (EDA), and respiration (RESP) are practically more feasible for driver drowsiness detection, as they are less intrusive to be integrated into the vehicle cabin and at the same time, can maintain robust physiological correlation with drowsiness states (Freitas, Almeida, Gonçalves, Conceição and Freitas, 2024; Gottlieb, Whitney, Bonekat, Iber, James, Lebowitz, Nieto and Rosenberg, 1999; Saleem et al., 2023). To the best of our knowledge, no generalizable methods have been proposed with ECG, EDA, and RESP as inputs.

Motivated by the above gaps, this study seeks to answer three questions. (1) What information in physiological signals is necessary and sufficient to support generalization for driver drowsiness detection under heterogeneous inducers, sensors, and assessment criteria? (2) How can we learn state-relevant representations from raw multi-modal physiological signals (ECG, EDA, and respiration) that remain consistent across domains while preserving the discriminability between awake and drowsy states? (3) How can the learning process be made robust to the label ambiguity and noise induced by subjective self-reports, so that the model is discouraged from fitting domain-specific biases and instead captures physiology-grounded semantics that are transferable to new environments?

To answer these questions, we propose a novel driver drowsiness detection method, called DrowsyDG-Phys, incorporating the principles of domain generalization (DG). We designed a backbone network to automatically learn informative representations of input signals from both the time and frequency domains. The model was trained on multiple source domains (i.e., datasets considered known before deployment) and tested on a target domain (i.e., the unknown dataset that simulates the deployment environment). Then, inspired by prior knowledge in physiology (Sun and Li, 2024), we introduced a contrastive-learning-based regularization to enhance the model's robustness against out-of-distribution samples in unseen domains. Additionally, to address the challenges posed by domain shifts, we proposed another regularization loss that centralizes the feature space of samples with the same label but in different domains and broadens the discriminative plane between drowsiness and awakeness. To fairly evaluate the performance of our proposed approach across multi-source domains, we utilized three datasets that include three types of drowsiness inducers and two assessment criteria. Additionally, we created a self-collected dataset that focused on drowsy driving in SAE Level-3 conditions, in which 60 participants were induced to be drowsy by manipulating the cabin environment in a driving simulator. Through extensive experiments, our proposed method, DrowsyDG-Phys, outperformed the comparison baselines and enhanced the generalization performance of driver drowsiness classification.

## 2. Related Work

### 2.1. Physiological Signals Related with Driver Drowsiness

Physiological signals can be used for early detection of driver drowsiness, as they can exhibit subtle variations before behavioral changes (Javed, Arshad, Saeed and Naseer, 2021). Usually, EEG has been considered as the gold standard of drowsiness detection as it directly captures brain activity related to alertness (Hussein, Miften and George, 2023; Gao, Wang, Yang, Mu, Cai, Dang and Zuo, 2019). Nevertheless, due to its highly intrusive nature and cumbersome wearing of EEG equipment, EEG's in-vehicle applications are limited (Ma, Yan, Billington, Merat and Markkula, 2024). In contrast, ECG, EDA, and RESP-based monitoring methods are of increasing interest for their ease of use, non-invasiveness, and practical advantages, especially with the increasing prevalence of wearable technologies (Lu, Tan, Zhang, Wang, Xie, Yue and Chen, 2025; Díaz-Santos, Caballero-Gil and Caballero-Gil, 2024).

However, traditional models for drowsiness monitoring have primarily relied on manual extraction of physiological features from ECG, EDA and RESP signals, followed by ML techniques such as LightGBM and SVM (Cheon and Kang, 2017; Chowdhury et al., 2018; Zhou et al., 2022). For example, ECG-based detection that relied on time-domain and frequency-domain features of heart rate variability (HRV) metrics achieved detection accuracies exceeding 90% based on driving simulation data (Awais, Badruddin and Drieberg, 2017a; Jeppesen, Fuglsang-Frederiksen, Johansen, Christensen, Wüstenhagen, Tankisi, Qerama and Beniczky, 2020; Patel, Lal, Kavanagh and Rossiter, 2011). To further reduce the intrusiveness of the ECG sensors, recent work has integrated innovative ECG sensors in the cabin, such as steering wheel-integrated ECG electrodes (Lu, Zheng, Tang, Zhang, Sheng, Wang, Jin, Yu and Zhou, 2021) and safety belt-embedded sensors (Kundinger, Sofra and Riener, 2020b). At the same time, EDA has also gained popularity for its non-invasive nature and ease of implementation. Key EDA metrics—including tonic skin conductance level, the number of skin conductance responses, and SCR amplitude demonstrated strong correlations with subjective drowsiness measures (Villarejo, Zapirain and Zorrilla, 2012; Malathi, Jayaseeli, Madhuri and Senthilkumar, 2018; Jiao, Zhang, Chen, Fu, Jiang and Wen, 2024). As for the RESP, the respiration patterns, characterized by decreased respiration rates and increased amplitude variability, also change with levels of sleepiness (Yang, Hu, Zhai and Zhang, 2022). However, it should be noted that although single-signal approaches have shown promise, integrating multiple

physiological signals has been found to yield superior detection performance. Systematic reviews consistently indicate that poly-signal approaches can outperform mono-signal methods, thereby offering a more comprehensive assessment of driver drowsiness (Awais, Badruddin and Drieberg, 2017b).

### 2.2. Generalizability of Physiological Signal-Based Drowsiness Detection

Existing research on physiological signal-based drowsiness detection has made progress in cross-subject generalizability, i.e., ensuring detection across subjects. For example, using interpretable convolutional neural networks with EEG signals, Cui, Lan, Sourina and Müller-Wittig (2022) achieved 78.35% accuracy in detecting drowsiness when trained and tested on different individuals, establishing a benchmark for subsequent cross-subjects approaches, which can be categorized by their generalization strategies: (1) Advanced network architectures, such as single-channel EEG with residual shrinkage networks (Feng, Guo and Kwong, 2025), which reached 74.72% accuracy; (2) Transfer learning techniques, such as the GDANN framework, which reached 91.63% accuracy (Zeng, Li, Borghini, Zhao, Aricò, Di Flumeri, Sciaraffa, Zakaria, Kong and Babiloni, 2021); (3) Multi-modal fusion approaches that combine EEG, EDA, and PPG signals with attention mechanisms (Guo, Yang and Wu, 2025), which reached 82.14% accuracy; and (4) Feature optimization methods, which uses carefully selected features with random forest models for ECG and respiration data (Cos, Lambert, Soni, Jeridi, Thieulin and Jaouadi, 2023), which reached 96% accuracy. However, these cross-subject approaches were not aligned with practical application scenarios, as they typically depended on the same dataset to evaluate the generalizability of a model by dividing data from different individuals into training and testing sets. As mentioned previously, the challenges to generalizability arise not only from individual differences but also from other factors such as the device, the type of drowsiness induction, etc. These factors often cannot be replicated through cross-individual assessment.

To address these limitations, DG evaluation and methods (Wang, Lu, Wang, Yang, Chen, He and Wu, 2025c; Wang, Lu, Han, Chen, He and Wu, 2025b; Wang, Wang, Hu, Wu and He, 2024c) were leveraged to achieve robustness and generalization in real-world deployment. When compared to domain adaptation (DA) (Wang, Gu and Yao, 2024a; Ma, Zhang, Sun, Wang and Gao, 2023; Yuan, Cui, Li, Zheng, Siyal and Yi, 2024), which is commonly used to enhance generalizability, DG has the advantage of not needing additional inputs from the deployment environment. Instead, it relies solely on the source domain data for pre-training before deployment. This approach aligns better with actual product development processes, as we often cannot collect extensive data from the deployment environment in advance Wang, Lu, Wang, Chen and He (2024b); Yang, He, Wang and Wu (2025a). However, existing limited DG research in driver drowsiness detection mostly relied on EEG signals (Ballas and Diou, 2024; Kim et al., 2025), though they have shown promising results. For example, some researchers have developed multi-layer representation networks that leverage features across deep CNNs to capture invariant attributes of EEG signals, improving performance across different datasets and tasks Ballas and Diou (2024).

However, as mentioned earlier, EEG-based systems have critical drawbacks - they require cumbersome electrode caps that are impractical for everyday driving scenarios (Mu, Liao, Tao and Shen, 2024; Wang, Yang, Hu, Tang, Liu, He, Wang, Chen and Wu). In contrast, ECG, EDA, and respiration monitoring offer significant advantages in in-cabin deployment. Specifically, ECG electrodes can be integrated into steering wheels or seatbelts. EDA sensors can be placed in watches or steering wheel grips. Respiration sensors can be built into seatbelts or seat pressure systems (Lee and Chung, 2012; Leonhardt, Leicht and Teichmann, 2018; Sang-Joong, Heung-Sub and Wan-Young, 2014). Despite these practical advantages, ECG/EDA/RESP signals face unique DG challenges: (1) Inconsistent drowsiness inducers like low arousal, mental fatigue, and sleep deprivation may lead to different physiological responses; (2) Variable assessment criteria were adopted in different studies, including both subjective measures and objective indicators; (3) Individual heterogeneity and differences in collection equipment can cause signal variations; and (4) Sensor type and placement variations and environmental factors can affect signal quality. Our study aims to address these challenges using a novel DL framework.

## 3. Methodology

### 3.1. Problem Formulation

We introduce a plug-in Domain Generalization (DG) framework for estimating driver drowsiness using physiological signals from non-invasive sensors (i.e., ECG, EDA, RESP), called DrowsyDG-Phys. As shown in Fig. (2), this framework first takes a total of $N$ physiological signals, denoted as $X = \{(x_i^\alpha, x_i^\beta, x_i^\gamma)\}_{i=1}^N$. $x_i^\alpha, x_i^\beta$, and $x_i^\gamma$, representing the ECG, EDA, and RESP, respectively. Further, since the frequency characteristics of physiological

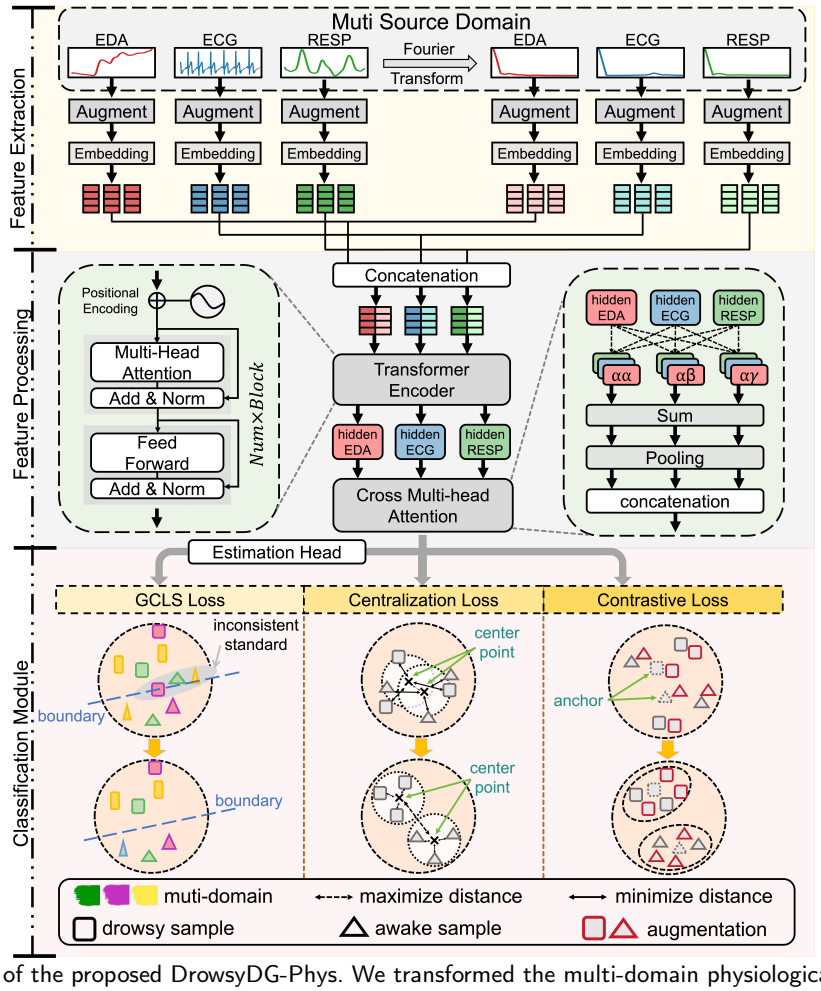


**Figure 2:** Illustration of the proposed DrowsyDG-Phys. We transformed the multi-domain physiological signals from time-domain to frequency-domain. Then, after neural network encoding, we leveraged three loss functions to mitigate the effects of heterogeneity and promote the generalizability of the model.

signals are essential for detecting drowsiness (Saleem et al., 2023), we also convert the signals into absolute frequency power series using a one-dimensional Fourier transform. We then combine the information from both the time and frequency domains to form the final $x_i$ for each signal.

Each $x_i \in \mathbb{R}^{W\times 1}$ comes from $M$ source domains. To simulate the application scenario, we directly input raw signals into DrowsyDG-Phys, which is represented as $f(X_S;\theta)$. Here, $W$ refers to the time window size plus the number of the signals in the frequency domain, and $\theta$ represents the parameters to be optimized. The DG protocol stipulates that $f(X_S;\theta)$ can only be trained using data from source domains, while testing occurs using the data $X_T$ from the target domain. Our objective is to optimize $\theta$ in order to minimize the discrepancy in the feature space between multiple source domains and the target domain, ultimately learning the best mapping $f(X_T;\theta)$ to predict the target drowsiness level $Y = \{y_i\}_{i=1}^{N}$, where $y_i \in \{0, 1\}$.

## 3.2. Architecture of Backbone Network

The overall architecture of DrowsyDG-Phys is illustrated in Figure 1. First, a backbone network $Enc(*)$ is initialized to obtain under-optimized representations $O = \{o_i\}_{i=1}^{N} \in \mathbb{R}^{N\times D}$ of the current sample in the source domains, where $D$ is the hidden layer dimension. Next, to improve precision in estimation and enhance robustness to unseen domain shifts, we simultaneously input the intermediate features $O$ into an estimation head that consists of fully connected (FC) layers, which then outputs the drowsiness label $Y$.

The initial phase involves the embedding of three physiological signals. Three single-layered linear encoders are applied to map each input from $\mathbb{R}^{W\times 1}$ to a learnable high-dimension feature space $E = \{e_i\}_{i=1}^{N} \in \mathbb{R}^{N\times W\times D}$. Then, we

**Table 1**
Summary of key symbols used in DrowsyDG-Phys.

| Symbol | Description |
|---|---|
| $X_S, X_T$ | Training data from source domains and test data from an unseen target domain. |
| $M$ | Number of source domains. |
| $N$ | Number of samples in a dataset split. |
| $y_i$ | Drowsiness label of sample $i$, $y_i \in \{0, 1\}$ (0: awake, 1: drowsy). |
| $W$ | Feature length (time-window size plus frequency-domain length). |
| $f(\cdot;\theta)$ | Drowsiness estimation model (DrowsyDG-Phys) with parameters $\theta$. |
| $Enc(\cdot)$ | Backbone encoder network. |
| $D$ | Hidden (embedding) dimension. |
| $O = \{o_i\}_{i=1}^N$ | Final sample representations, $O \in \mathbb{R}^{N\times D}$. |
| $Q, K, V$ | Query/key/value matrices in (self-/aligned-)attention. |
| $d_K$ | Key dimension used in scaled dot-product attention. |
| $B$ | Mini-batch size used in optimization. |
| $\mu_c$ | Class-$c$ centroid in the current mini-batch, $\mu_c \in \mathbb{R}^D$. |
| $\mathcal{L}_{Center}$ | Feature centralization regularization (Eq. 4). |
| $\mathcal{L}_{Contrast}$ | Contrastive regularization with temporal-consistency augmentation (Eq. 5). |
| $\mathcal{L}_{GCLS}$ | Generalizable classification loss for handling label ambiguity/noise (Eq. 8). |
| $K$ | Number of augmented views used in $\mathcal{L}_{Contrast}$. |
| $\tau$ | Temperature parameter in $\mathcal{L}_{Contrast}$. |
| $\lambda(Iter)$ | Adaptation factor (as a function of training iteration $Iter$) controlling regularization strength. |
| $k_1, k_2$ | Trade-off coefficients for $\mathcal{L}_{Contrast}$ and $\mathcal{L}_{Center}$ in $\mathcal{L}_{overall}$. |
| $\mathcal{L}_{overall}$ | Overall training objective (Eq. 10). |

employ both self-attention and aligned attention to capture the temporal dependency within and between the signals. Specifically, the attention mechanism is represented by Equation (1):

$$f_{ii} = \text{Softmax}(\frac{QK^T}{\sqrt{d_K}})V, \tag{1}$$

where, $Q$ represents the query matrix, which calculates correlation scores, or attention weights, i.e., how much attention each position pays to others. $K$ is the key matrix, which is used to determine the correlation scores (attention weights) between the query and different positions. $V$ stands for the value matrix, which contains the information that will be forwarded to the next layer. Each position retrieves information from the value matrix based on the computed attention weights. The $Q, K, V$ are obtained from the same $e_i$ by three independent linear layers. Through the multi-head attention operation (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser and Polosukhin, 2017) over the embedding of each signal, we receive the representation $e_i^{'} = \{(e_i^{'\alpha}, e_i^{'\beta}, e_i^{'\gamma})\}_{i=1}^N$ that contains the temporal dependency of the corresponding signal, where $e_i \in \mathbb{R}^{W\times D}$.

Self-attention captures the dependencies within each type of signal, while aligned attention focuses on aligning and integrating information across different signals. This approach enables the model to assess and combine the most relevant features from each signal, resulting in a cohesive and comprehensive representation. By fusing diverse data sources, this mechanism helps the model understand and predict physiological patterns through the capture of complex inter-dependencies. The aligned attention mechanism is represented by Equation (2).

$$\begin{aligned} f_{ij} &= \text{Softmax}(\frac{Q_i K_j^T}{\sqrt{d_K}})V_j, where\ i, j \in \{\alpha, \beta, \gamma\} \text{ and } i \neq j, \\ o_i &= \text{Pooling}(\sum_{i,j\in\{\alpha,\beta,\gamma\}} f_{ij} \cdot e^{'ij}). \end{aligned} \tag{2}$$

We instantiate an average pooling layer to reduce the $W$ dimension, and the final shape of $O$ is $\mathbb{R}^{N\times D}$, which stands for the final representation aggregating the self-dependency and interdependency in time-domain and frequency-domain of three physiological signals.

### 3.3. Feature Centralization Regularization

The heterogeneity of labeling criteria across multiple driver drowsiness datasets introduces domain-specific biases that undermine model generalizability. Traditional supervised learning tends to optimize feature representations that are discriminative within individual domains but may fail to align class semantics across domains. To address this challenge, inspired by (Yang, Yang, Wei, Hu and Lv, 2024), we propose a novel center loss constraint that simultaneously enforces intra-class compactness and inter-class separability across domains. This dual-objective mechanism aligns feature distributions of the same class from different labeling domains while maintaining decision boundaries between classes.

Specifically, given a mini-batch of feature representations $O = \{o_i\}_{i=1}^{B} \in \mathbb{R}^{B\times D}$ from multiple source domains with corresponding labels $Y = \{y_i\}_{i=1}^{B}$, we first define the class centroid $\boldsymbol{\mu}_c \in \mathbb{R}^D$ for each class $c$ in the current mini-batch as:

$$\boldsymbol{\mu}_c = \frac{1}{B_c} \sum_{y_i=c} o_i, \tag{3}$$

where $B$ denotes the mini-batch size, $B_c$ is the number of class-$c$ samples in the mini-batch, $o_i \in \mathbb{R}^D$ is the embedded representation of the $i$-th sample, and $y_i \in \{0, 1\}$ is its label (0: awake, 1: drowsy).

We incorporate two optimization objectives into this regularization: Intra-class Invariance and Cross-class Discriminability. To achieve intra-class invariance, we minimize the $\ell_2$ distance between each sample feature and its class-specific centroid, which helps compress features from different domains (e.g., dataset A based on subjective criteria and dataset B relying on objective criteria) into unified clusters for each semantic class. Additionally, we enhance cross-class discriminability by enlarging the distance between centroids of opposing classes, so that the decision boundary is encouraged to locate at low-density regions of the embedding space. The feature centralization regularization $\mathcal{L}_{Center}$ is formulated as:

$$\mathcal{L}_{Center} = \underbrace{\frac{1}{B}\sum_{i=1}^{B} \|o_i - \boldsymbol{\mu}_{y_i}\|_2}_{\text{Intra-class Invariance}} - \underbrace{\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_2}_{\text{Cross-class Discriminability}}. \tag{4}$$

### 3.4. Temporal Consistency Augmentation and Contrastive Regularization

In addition, to address feature space instability caused by cross-domain variations in acquisition devices, experimental conditions, and individual differences, we propose a contrastive regularization loss based on physiological temporal priors. Motivated by the biological principle that human drowsiness states exhibit short-term stability (typically $< 1$ s) (Sun and Li, 2024), we design a temporal consistency constraint through contrastive learning. This regularization enhances model robustness against distributional shifts in unseen domains by enforcing invariant feature representations under simulated signal variations.

In general, the core insight lies in treating temporal perturbations as pseudo-domain shifts. Specifically, for an input physiological signal sequence $x_i \in \mathbb{R}^{W\times 1}$, we generate $K$ augmented view $x_i^a \in \mathbb{R}^{W\times 1}$ via random backward shifting $\Delta t \sim \mathcal{U}\{0, 1s\}$ along the time axis, simulating device-specific artifacts while preserving state semantics. The contrastive loss then operates on feature representations $O$ as follows:

$$\mathcal{L}_{Contrast} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\sum_{y_k=y_i}^{K} \exp^{\text{sim}(o_i,o_k)/\tau}}{\sum_{y_j\neq y_i}^{N} \exp^{\text{sim}(o_i,o_j)/\tau}} \tag{5}$$

Where sim$(*)$ refers to the cosine similarity. In the loss function, $\tau = 0.1$ is used to modulate the concentration of similarity. This design can enhance DG through two mechanisms: 1) Temporal consistency augmentation, which increases robustness to variances in signal acquisition by maximizing feature consistency between original views and their time-shifted counterparts; 2) temporal contrastive regularization, which improves the clarity of decision boundaries by employing hard negative mining across heterogeneous domains. As illustrated in Fig. (2), the contrastive regularization works in conjunction with our centralization loss by operating at different geometrical scales.

Specifically, while centralization loss aligns the global class centroids across domains, the contrastive loss preserves the relationships among the local samples, making them invariant to physiological signal distortions. Together, these methods create a hierarchical domain-invariant feature space.

### 3.5. Optimization Goal

Given the existence of different subjective feelings across individuals in assessing drowsiness through self-report questionnaires (Paulhus, Vazire et al., 2007), the relationship between physiological features and drowsiness state labels may not be consistently reliable when training the model with data from different individuals. To enhance the generalizability of drowsiness estimation, we instantiate the generalizable cross-entropy loss (Zhang and Sabuncu, 2018) as our classification loss, denoted as $\mathcal{L}_{GCLS}$. Specifically, the thresholds $\pi(0 < \psi < 1)$ are used to distinguish uncertain samples common in subjective evaluations. As shown in Equation (7), for the sample $i$ with prediction $f(x)$, if $\mathcal{L}_q(f(x_i;\theta), y_i)$ is below threshold $\psi$, the $w_i$ is set to 0 to prevent this sample from affecting model parameters during backpropagation. We define our base loss function $\mathcal{L}_{\mathrm{q}}(f(x), y)$ as follows:

$$\mathcal{L}_{\mathrm{q}}(f(x), y) = \frac{1 - f(x)^q}{q} \tag{6}$$

$$w_i = \begin{cases} 1 & \text{if } \mathcal{L}_q(f(x_i;\theta), y_i) \leq \psi \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Here, $f(x;\theta)$ represents the predicted probability for the true label $y$, and $q$ is a positive number less than 1 that controls how the loss function responds to transitional drowsiness states. A smaller $q$ increases sensitivity to drowsiness. The constant $\psi$ is calculated as $\psi = \frac{1-\pi^q}{q}$. Based on these definitions, we can present our improved loss function, denoted as $\mathcal{L}_{GCLS}$, below:

$$\mathcal{L}_{GCLS} = \frac{1}{B} \sum w_i \left( \mathcal{L}_q(f(x_i;\theta^{(Iter-1)}), y_i) - \psi \right) \tag{8}$$

$w_i$(0 or 1) represents the binary weight for sample $i$, $\theta^{(Iter-1)}$ denotes the classification model parameters at last iteration. This design provides a progressive sample selection mechanism that gradually identifies and filters out ambiguous transitional states and mislabeled samples that are common in driver drowsiness datasets with self-reported criteria.

Additionally, to mitigate any unintended effects of regularization during the early iterations, we introduce adaptation factors, $\lambda$, as shown in Equation (9). Notably, $\mathcal{L}_{GCLS}$ is bounded in magnitude, while $\mathcal{L}_{Center}$ and $\mathcal{L}_{Contrast}$ can take substantially larger values depending on batch composition and feature norms. Therefore, we use small trade-off coefficients to prevent the regularizers from dominating the optimization dynamics and to avoid over-regularization. In all, we combine the main loss and the two aforementioned regularizations into one loss formula $\mathcal{L}_{overall}$, to conduct the joint training with trade-off parameters $k_1, k_2$ according to Equation (10).

$$\begin{aligned} t &= \frac{Iter_{current}}{Iter_{total}}, \\ \lambda &= \frac{2}{1 + \exp^{-10t}}. \end{aligned} \tag{9}$$

$$\mathcal{L}_{overall} = \mathcal{L}_{GCLS} + \lambda * k_1 * \mathcal{L}_{Contrast} + \lambda * k_2 * \mathcal{L}_{Center}. \tag{10}$$

In practice, the three loss terms target complementary sources of heterogeneity in multi-dataset physiological drowsiness detection. $\mathcal{L}_{GCLS}$ serves as the primary optimization objective, whereas $\mathcal{L}_{Center}$ and $\mathcal{L}_{Contrast}$ are used as regularizers to promote domain-invariant representations. $\mathcal{L}_{GCLS}$ primarily mitigates label ambiguity/noise introduced by subjective self-reports by reducing the influence of uncertain samples during optimization. $\mathcal{L}_{Center}$ mainly addresses

**Table 2**
Comparison of datasets for drowsiness estimation. In this table, Avg indicates the mean value, SD stands for standard deviation, M is male and F is female.

| Dataset | Subject | Age (Avg, Min-Max, SD) | Gender (M, F) | Device | Frequency | Inducer | Criteria |
|---|---|---|---|---|---|---|---|
| DDCE | 60 | 27.2, 22-44, 4.7 | 30, 30 | Ergoneers | 100 HZ | Cabin temperature and carbon dioxide | Subjective |
| AdVitam | 63 | 23.8, 18-64, 4.8 | 45, 18 | BioPac MP36 | 1000 HZ | Sleep deprivation | Subjective |
| MCDD | 42 | 35.3, 23-53, 9.1 | 25, 17 | Ergoneers | 100 HZ | Mental fatigue | Subjective |
| LAD | 27 | 36.7, 19-74, 14.4 | 13, 14 | Becker Meditec | 256 HZ | Low arousal | Objective |

cross-domain semantic misalignment by aligning global class centroids and enlarging inter-class margins across heterogeneous labeling rubrics. $\mathcal{L}_{Contrast}$ focuses on signal-level domain shift by enforcing temporal-consistency invariance under physiology-preserving perturbations, thereby stabilizing local neighborhood structure. Together, $\mathcal{L}_{Center}$ (global alignment) and $\mathcal{L}_{Contrast}$ (local invariance) form a hierarchical domain-invariant embedding space.

# 4. Datasets

In this work, we utilize four datasets and summarize their key characteristics in Table 2 for comparison.

## 4.1. Self-collected Dataset

Previous public drowsy driving datasets include various types of drowsiness inducers, such as sleep deprivation (Meteier, Capallera, De Salis, Angelini, Carrino, Widmer, Abou Khaled, Mugellini and Sonderegger, 2023), mental fatigue (Wang, Wang, Yan, He and Wu, 2024d), and low arousal (Ayas, He and Donmez, 2024b). However, the cabin environment was usually ignored, although it can also affect the driver's state. To address this gap, we collected a new drowsy driving dataset (**DDCE**) that investigated the influence of temperature and carbon dioxide concentrations on drowsiness in the context of conditional driving automation.

### *4.1.1. Experiment Design*

The experiment adopted a 2×3 between-subject design, examining the effects of carbon dioxide concentration (High Concentration and Low Concentration) and temperature (Slightly Cool, Neutral, and Slightly Warm) on drivers' behaviors and states. Each participant was assigned to one environmental condition, resulting in six distinct groups, each consisting of 10 participants. The assignment was randomized and balanced by gender, with five males and five females in each group.

The three temperature conditions were determined using the Predicted Mean Vote (PMV) model (Sunagawa, Shikii, Beck, Kek and Yoshioka, 2023), which is a widely accepted standardized thermal comfort model utilized in both buildings and vehicle cabins. The PMV model predicts thermal comfort on a scale from -3 to +3, where "-3" indicates "Cold," "-2" is "Cool," "-1" is "Slightly Cool," "0" is "Neutral," "+1" is "Slightly Warm," "+2" is "Warm," and "+3" is "Hot." This model assesses thermal comfort by combining environmental variables (air temperature, radiant temperature, relative humidity, and air velocity) with personal variables (metabolic rate and clothing insulation), based on the assumption of balance of human heat under steady-state conditions (Fanger, 1970). In this experiment, participants wore trousers and short-sleeve shirts, providing a clothing insulation value of 0.57 clo. The metabolic rate was estimated at 1 met, consistent with seated activities during performance tasks. The indoor air velocity was maintained at 0 m/s, and relative humidity (RH) was approximately 60%, with negligible variation. Using these standard parameters, three temperature conditions corresponding to different thermal comfort levels (Slightly Cool, Neutral, and Slightly Warm) were calculated based on the PMV model (Tartarini, Schiavon, Cheung and Hoyt, 2020).

Two carbon dioxide concentration conditions were established to reflect the two common ventilation modes used in vehicle HVAC systems. The high concentration condition simulated the recirculation ventilation (RC) mode, whereas the low concentration condition represented the outside air ventilation (OA) mode. The RC mode is often used to enhance heating or cooling efficiency, but it can lead to elevated carbon dioxide levels due to the accumulation of exhaled air in a confined cabin space. In accordance with previous studies, the low carbon dioxide concentration was maintained at approximately 1200 ± 300 ppm, while the high concentration was set at about 4200 ± 300 ppm (Angelova, Markov, Simova, Velichkova and Stankov, 2019; Zhao, Jiang and Song, 2022), representing the environmental conditions associated with two ventilation modes.

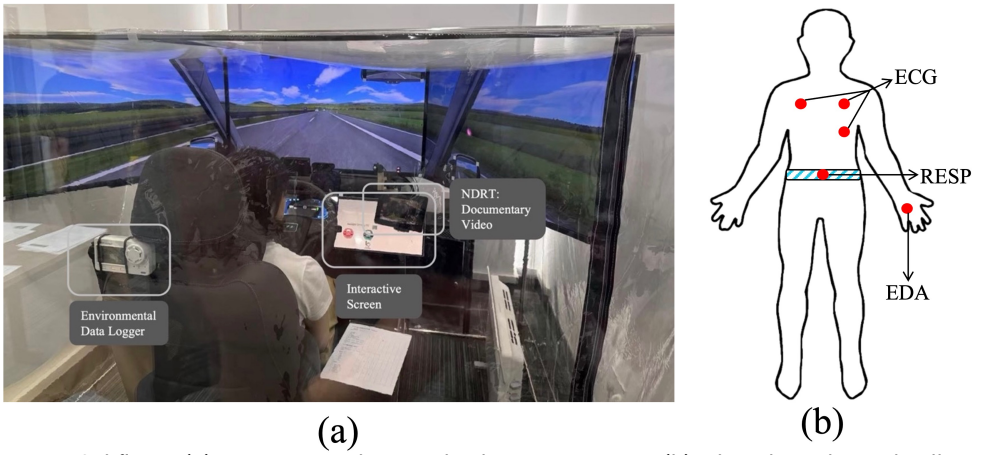


**Figure 3:** Subfigure (a): Driving simulator and cabin environment. (b) Physiological signal collection.

### *4.1.2. Participants*

Participants were recruited through advertisements on the university campus and in nearby residential areas. The experiment lasted approximately 2 hours, and participants were compensated at a rate of 80 RMB/h. All participants were required to be non-novice drivers, defined as having held a valid C1/C2 driver's license in China for at least three years and having driven more than 5,000 km in the past year. Participants were instructed to avoid alcohol and caffeine on the day of the experiment and to maintain their regular sleep schedule the night before. Their usual daily sleep duration and the previous night's sleep duration were recorded prior to the experiment.

Initially, 71 participants were recruited. Due to coffee intake, inadequate sleep the night before, overwork, or sensitivity to temperature, 60 participants met the experimental criteria and completed the study. The study was approved by the Ethics Compliance Committee at the Hong Kong University of Science and Technology (Guangzhou) under protocol HSP-2023-0015.

### *4.1.3. Experiment Procedures*

The experimental procedure was designed to ensure consistency and reliable data collection. Upon arrival, participants were briefed about the purpose of the study, reviewed a consent form, and completed a 5-minute practice session on the driving simulator to familiarize themselves with the simulator.

The formal experiment involved a 90-minute drive with driving automation on a highway, with five random takeover events: the first four at 15-minute intervals and the last, an emergency braking task, 30 minutes after the fourth one. At the end of the interval, participants responded to a takeover request and manually drove for approximately 500 meters. They then received a voice command to reactivate the driving automation and resume non-driving related tasks (watching a documentary), simulating periods of being "out-of-the-loop." Temperature and carbon dioxide levels were controlled in the simulator for consistency. Notably, in this work, we only utilized data before the takeover request. The KSS questionnaire was filled out by the participant after each takeover task to assess the drowsiness level before the takeover.

### *4.1.4. Apparatus*

The experiment used a fixed-base driving simulator developed by Info Tech, Shanghai, as shown in Figure 3(a). Driving scenarios were created using SILAB 7.1 software (also from WIVW GmbH) and displayed on three 43-inch screens, positioned approximately one meter away from the driver. Each screen had a resolution of 1920×1080, providing a horizontal viewing angle of 150° and a vertical viewing angle of 47°,. Driving data was logged at a frequency of 60 Hz using the simulation software, ensuring accurate and detailed data collection for analysis. Additionally, physiological data, including ECG, EDA, and RESP, were collected at a frequency of 100 Hz using sensors from Ergoneers, as shown in Figure 3(b). Participants filled out the KSS questionnaire at the end of each interval throughout the experiment.

### 4.2. Other Datasets

We also evaluated our method using three additional datasets. Specifically, the **AdVitam** dataset (Meteier et al., 2023) assessed the drowsiness of 63 drivers during SAE Level 3 (L3) automated driving. We utilized parts of data from its 'Experiment 4', where participants' drowsiness was induced through sleep deprivation (i.e., less than six hours of sleep the night before the experiment) and assessed using the KSS questionnaire during the one-hour driving.

The **MCDD** dataset (Wang et al., 2024d) involved 42 drivers who experienced mental fatigue from conducting three types of non-driving cognitive tasks (i.e., N-back, math calculation, spatial search) in a 2.5-hour SAE L3 automated driving. Similar to AdVitam, drowsiness in this dataset was measured using the KSS questionnaire every 5 minutes.

Finally, the **LAD** dataset (Ayas et al., 2024b) observed drowsiness in 27 drivers with monotonous drive characterized by low arousal. This refers to the participants driving an L2 automated vehicle for up to 1.5 hours in a monotonous scenario with no interaction with surrounding vehicles. The drowsiness annotations were provided by independent raters based on objective indicators derived from facial video recordings (Wierwille and Ellsworth, 1994). Note that, AdVitam and MCDD are publicly available, while LAD is currently being prepared for public release.

### 4.3. Prepocessing

Since the raw data was directly used as inputs in the model, only noise elimination was performed to improve data quality. Specifically, all signals were down-sampled to a frequency of 100 Hz to enhance computational efficiency and maintain consistency across datasets. For EDA, a low-pass filter with a cutoff frequency of 5 Hz was applied. At the same time, ECG and RESP signals were processed using band-pass filters with frequency ranges of 3–45 Hz and 0.1–0.35 Hz, respectively (Meteier, Capallera, Ruffieux, Angelini, Abou Khaled, Mugellini, Widmer and Sonderegger, 2021). Note that, for datasets annotated by the KSS questionnaire, we labeled samples with KSS$\geq$ 7 as drowsy, and KSS$<$ 6 as awake (Buendia, Forcolin, Karlsson, Arne Sjöqvist, Anund and Candefjord, 2019; Ahlström and Anund, 2024).

## 5. Experiment Setting

### 5.1. Baseline Models

We first selected four classic ML models for comparison: Logistic Regression (**LR**) (Hosmer Jr, Lemeshow and Sturdivant, 2013): A generalized linear model that estimates class probabilities by optimizing regression parameters through prediction function identification and loss minimization. Support Vector Machine (**SVM**) (Hearst, Dumais, Osuna, Platt and Scholkopf, 1998): it constructs maximum-margin hyperplanes using kernel functions (e.g., RBF) to separate classes while minimizing structural risk. Random Forest (**RF**) (Breiman, 2001): An ensemble method combining multiple decision trees (Song and Ying, 2015) to aggregate predictions for robust classification or regression. **LightGBM** (Ke, Meng, Finley, Wang, Chen, Ma, Ye and Liu, 2017): A gradient-boosting framework with innovative techniques (Gradient-based Single-Side Sampling and Exclusive Feature Bundling) to enhance classification efficiency and prevent overfitting.

Besides, we compared our model with selected DL-based approaches from previous studies: **GRU** (Dey and Salem, 2017): it simplifies LSTM by using update and reset gates to capture long-term dependencies. We added a classification head after hidden feature pooling. **MTCNN** (Xie, Murphey and Kochhar, 2019): A CNN-based architecture for drowsiness detection by integrating multivariate temporal features. In our work, vehicle signal components were removed. **CLSTM** (Hultman et al., 2021): it combines CNN-based feature extraction with LSTM recurrent modeling. In our work, it was modified to accept ECG, EDA, and RESP signals. **DSCNN** (Lyu, Akbar, Manimurugan and Jiang, 2025): It utilizes depthwise separable CNNs for drowsiness detection. In our work, its inputs were adapted to the three physiological signals as ours. **Informer** (Zhou, Zhang, Peng, Zhang, Li, Xiong and Zhang, 2021): Designed for long-sequence forecasting with ProbSparse self-attention and attention distillation. In our work, only the encoder was retained and augmented with a classifier. **TFormer** (Li, Hu, Gao, Wang, Suganthan and Sourina, 2024): it enhances time-series features through global time-frequency patterns. We modified it to process the multi-physiological signals as ours.

Furthermore, four DG methods were used for comparison: **AD** (Ganin and Lempitsky, 2015): It aligns source and target domain distributions via adversarial training with a gradient reversal layer. **VREx** (Krueger, Caballero, Jacobsen, Zhang, Binas, Zhang, Le Priol and Courville, 2021): it enhances generalization by minimizing risk variance across multiple source domains. **GroupDRO** (Sagawa, Koh, Hashimoto and Liang, 2020): it optimizes worst-case

group risk to improve robustness against domain shifts. **Mixstyle** (Zhou, Yang, Qiao and Xiang, 2024): it augments domain-invariant features by interpolating feature statistics across source domains.

### 5.2. Implementation Details

The whole work was implemented by the Pytorch framework and conducted on one NVIDIA RTX3090. For the data preparation, after filtering the physiological signals, we applied the sliding window $W = 400$ (i.e., 4s) from $[200, 400, 600, 800]$ and step size 200 to ensure a more accurate estimation. In practice, we set up a hidden size of 128, 8 attention heads, and 1 layer of the backbone network. The model was trained using the Adam optimizer with a learning rate of 0.0001, a dropout rate of 0.1, and a batch size of 32. The training was capped at 100 iterations with early stopping based on the highest accuracy in the target domain. For the choice of $k_1, k_2$, among values of $\{0.1, 0.05, 0.01, 0.005, 0.001\}$, we observed the best performance around 0.01 and stable performance in the neighborhood of this value in pilot tuning; thus, we fixed both $k_1, k_2$ to 0.01 throughout the following experiments. The maximum iterations were set to 20000.

Since the ML model relies on manually extracted features, we fixed 1 minute as the time window with a 30-second step, and extracted a total of 28 key physiological features following (Meteier, Favre, Viola, Capallera, Angelini, Mugellini and Sonderegger, 2024) for ECG, EDA, and RESP. As for the DG methods, we deployed them on our proposed backbone network, which was mentioned in Section 3.2.

### 5.3. Evaluation Protocol and Metrics

The evaluation protocol (DG protocol) is implemented in a cross-domain evaluation scenario (Wang et al., 2024c) by dividing four datasets into two groups: three datasets for training (source domains) and then testing on the remaining dataset (target domain).

Besides, the average Accuracy (ACC), F1 score (F1), Sensitivity (SEN), Specificity (SPE), and Precision (PRE) on the target domain were used for model evaluation. The evaluation metrics are formulated as:

$$\begin{aligned}
\text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \\
\text{F1 Score} &= 2 \cdot \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}, \\
\text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\
\text{Sensitivity} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\
\text{Specificity} &= \frac{\text{TN}}{\text{TN} + \text{FP}}.
\end{aligned} \tag{11}$$

where TP, FP, FN, and TN denote true positives, true negatives, false positives, and false negatives, respectively. ACC quantifies the overall proportion of correct predictions made by the model for both awake (0) and drowsy (1) states, serving as an indicator of the model's general classification performance. The F1 represents a harmonic mean of precision and sensitivity, providing a balanced measure of the model's ability to accurately identify drowsiness states (minimizing false positives) while ensuring overall detection performance (i.e., minimizing false negatives). SPE evaluates the model's capability to correctly classify the awake state, reflecting its reliability in avoiding the misclassification of awake instances as drowsy. SEN, on the other hand, assesses the model's effectiveness in accurately detecting drowsiness states, ensuring that the true drowsiness instances are captured to reduce the likelihood of missed detections. PRE evaluates the drowsiness detection correctness of the model, representing the proportion of true drowsiness instances among all segments flagged as drowsy, thereby reflecting the model's reliability in minimizing false alarms. We report the mean results of five independent evaluations initialized by five random seeds.

## 6. Results and Discussion

### 6.1. Comparison Experiment on DG Protocol

Table 3 shows the comparison results on the DG protocol of four datasets. Firstly, we notice that, traditional ML methods perform relatively poorly. Particularly, ML models generated extreme outputs and identified all states

**Table 3**
Results of drowsiness estimation tested on different datasets based on DG protocol.

| | DDCE | | | | | AdVitam | | | | | MCDD | | | | | LAD | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | ACC | F1 | SEN | SPE | PRE | ACC | F1 | SEN | SPE | PRE | ACC | F1 | SEN | SPE | PRE | ACC | F1 | SEN | SPE | PRE |
| *Machine Learning Methods* | | | | | | | | | | | | | | | | | | | | |
| LR | 50.0 | 0.0 | 0.0 | 1.0 | 0.0 | 43.6 | 58.8 | 82.7 | 6.3 | 45.6 | 35.1 | 51.9 | 98.8 | 0.0 | 35.2 | 74.3 | 85.2 | 98.2 | 0.0 | 75.2 |
| SVM | 50.0 | 0.0 | 0.0 | 1.0 | 0.0 | 51.2 | 0.0 | 0.0 | 1.0 | 0.0 | 35.5 | 52.4 | 100.0 | 0.0 | 35.5 | 75.6 | 86.1 | 100.0 | 0.0 | 75.6 |
| RF | 50.0 | 5.7 | 3.0 | 96.9 | 44.5 | 53.3 | 53.2 | 54.4 | 52.3 | 52.1 | 37.5 | 52.9 | 98.8 | 3.7 | 36.1 | 25.0 | 2.3 | 1.2 | 98.9 | 22.8 |
| LightGBM | 51.5 | 11.1 | 6.0 | 96.9 | 74.0 | 47.6 | 64.4 | 97.2 | 0.5 | 48.1 | 43.6 | 52.4 | 87.3 | 19.6 | 37.4 | 65.5 | 77.9 | 80.5 | 18.9 | 75.5 |
| *Deep Learning Methods* | | | | | | | | | | | | | | | | | | | | |
| GRU | 68.3 | 60.8 | 65.4 | 48.2 | 56.8 | 58.1 | 58.7 | 50.2 | 68.4 | 70.6 | 65.3 | 52.0 | 47.5 | 56.2 | 57.5 | 68.9 | 70.5 | 75.6 | 50.3 | 66.0 |
| MTCNN | 72.5 | 68.7 | 80.2 | 58.7 | 60.1 | 67.3 | 60.5 | 53.6 | 72.9 | 69.4 | 69.8 | 56.4 | 51.8 | 60.7 | 61.9 | 75.3 | 78.4 | 85.2 | 52.1 | 72.6 |
| CLSTM | 71.0 | 66.5 | 78.9 | 55.3 | 57.5 | 63.8 | 59.2 | 52.1 | 70.6 | 68.6 | 67.5 | 55.1 | 50.3 | 58.4 | 60.9 | 72.6 | 76.9 | 82.7 | 51.8 | 71.9 |
| DSCNN | 73.8 | 70.1 | 82.5 | 60.0 | 61.0 | 68.5 | 59.4 | 51.3 | 75.2 | 70.5 | 70.2 | 58.3 | 53.6 | 63.1 | 64.0 | 76.8 | 80.2 | 87.1 | 53.7 | 74.3 |
| Informer | 70.2 | 69.9 | 84.1 | 62.4 | 60.0 | 69.7 | 60.1 | 52.8 | 75.5 | 69.8 | 71.6 | 59.7 | 55.9 | 63.8 | 64.1 | 77.5 | 81.7 | 89.0 | 55.2 | 75.5 |
| TFormer | 72.6 | 70.4 | 86.0 | 63.8 | 59.6 | 70.4 | 63.8 | 55.5 | 77.0 | 75.0 | 72.9 | 58.5 | 54.2 | 64.3 | 63.5 | 78.2 | 81.1 | 88.1 | 56.5 | 75.1 |
| *Domain Generalization Methods* | | | | | | | | | | | | | | | | | | | | |
| AD | 72.4 | 63.6 | 70.7 | 50.1 | 57.8 | 61.2 | 62.3 | 55.1 | 74.8 | 71.7 | 69.9 | 56.1 | 51.3 | 61.9 | 61.9 | 72.6 | 76.1 | 82.3 | 55.7 | 70.8 |
| VREx | 77.8 | 72.4 | 88.7 | 64.0 | 61.1 | 72.1 | 61.9 | 56.8 | 79.7 | 68.0 | 74.2 | 60.1 | 55.1 | 67.1 | 66.1 | 79.7 | 82.8 | 90.5 | 54.0 | 76.3 |
| GroupDRO | 75.1 | 70.2 | 85.2 | 68.3 | 59.7 | 65.9 | 60.5 | 51.2 | 70.1 | 73.9 | 69.8 | 59.3 | 54.6 | 66.5 | 64.9 | 75.8 | 81.5 | 86.8 | 49.0 | 76.8 |
| Mixstyle | 64.9 | 65.5 | 73.4 | 56.6 | 59.1 | 61.7 | 59.0 | 53.5 | 66.0 | 65.8 | 70.1 | 57.4 | 49.2 | 64.3 | 68.8 | 69.3 | 77.7 | 82.0 | 58.0 | 73.8 |
| **Ours** | **79.7** | **81.6** | **92.3** | **70.0** | **73.1** | **77.6** | **69.6** | **64.2** | **86.5** | **76.0** | **76.3** | **67.3** | **62.4** | **73.7** | **73.0** | **81.8** | **88.4** | **93.7** | **57.4** | **83.7** |

in the test dataset as one state (i.e., outputs were all awake or drowsy). This suggests that these models struggle to adapt to changes in the source domain and have significant difficulty managing domain shifts when confronted with unknown domains, ultimately leading to a breakdown in model learning. This finding validates our motivation, that ML methods have worse capacity with complex variations. In contrast, most DL methods outperform ML methods across all datasets and with more balanced output. For instance, on DDCE, GRU achieves an accuracy of 68.3%, a 32.6% increase compared to the best ML method (LightGBM at 51.5%). Additionally, in AdVitam, the sensitivity of RF, which is the best model in ML methods in this domain, still performs worse than GRU. This demonstrates that DL models can achieve more robust performance in complex environments through finer-grained signal feature extraction methods, because of their broader space of learnable parameters.

Among DL models, performance varies depending on architecture. We notice that, compared to the recurrent neural networks (RNNs) based model (i.e., GRU), CNN-based models (i.e., MTCNN, CLSTM, DSCNN) bring advantages. For instance, compared to GRU on LAD, CLSTM achieves a 5.4% improvement, and MTCNN without any RNN-based component achieves a 9.3% improvement in accuracy. Nevertheless, transformer-based models, such as Informer and TFormer, also demonstrate strong performance across most scenarios. Notably, TFormer consistently outperforms other DL models across all target domains. For instance, despite DSCNN achieving a higher ACC on the DDCE dataset, TFormer exhibits a more balanced performance (as indicated by higher F1 score, sensitivity, and specificity) and outperformed other model DL models for most metrics on all datasets. This suggests that the transformer architecture has greater potential compared to the CNN architecture.

When comparing DL and DG methods, DG models generally exhibit superior performance across various metrics. However, it is important to note that some DG methods do not yield improvements and can lead to significant performance degradation, with improvements often being unstable. For instance, AD and Mixstyle demonstrate a performance decline of approximately 3% to 14% in ACC compared to TFormer across all target domains. We propose several possible explanations. First, the AD's principle aims to obscure the model's feature space so that it cannot be identified as originating from any specific domain. This approach facilitates domain-invariant feature learning, although adversarial training can often be unstable. Additionally, Mixstyle was designed based on the CNN architecture and not on the transformer architecture. Among DG baselines, VREx achieves the best performance and provides a more consistent improvement of around 5% on average compared to DL methods. Nonetheless, our proposed method, DrowsyDG-Phys, consistently outperforms all other methods in every case, achieving an average improvement of 8% across all metrics compared to VREx. These results indicate the importance of generalization and tailored methods for specific tasks.

**Table 4**
Ablation study of different settings. In this table, '*' indicates the significant (p-value < 0.05) best result within each column with the paired-t test.

| Variants | | | DDCE | | | AdVitam | | | MCDD | | | LAD | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{L}_{GCLS}$ | $\mathcal{L}_{Contrast}$ | $\mathcal{L}_{Center}$ | ACC | F1 | SPE | ACC | F1 | SPE | ACC | F1 | SPE | ACC | F1 | SPE |
| - | - | - | 72.0 | 72.8 | 67.2 | 70.1 | 59.5 | 79.2 | 68.9 | 58.3 | 65.4 | 74.0 | 78.2 | 50.9 |
| ✓ | - | - | 75.1 | 76.5 | 69.8 | 75.8 | 67.1 | 84.3 | 73.5 | 63.7 | 70.8 | 79.5 | 85.2 | 55.0 |
| - | ✓ | - | 73.5 | 74.2 | 68.3 | 73.2 | 63.8 | 82.4 | 72.8 | 63.5 | 69.2 | 76.1 | 80.5 | 52.4 |
| - | - | ✓ | 74.8 | 75.8 | 69.1 | 71.5 | 61.2 | 80.9 | 70.6 | 60.9 | 67.8 | 77.3 | 82.7 | 53.8 |
| ✓ | ✓ | - | 78.2 | 80.1 | 67.1 | 76.3 | 67.8 | 85.2 | 75.0 | 66.0 | 72.5 | 80.5 | 86.2 | 55.8 |
| ✓ | - | ✓ | 77.8 | 79.4 | 64.3 | 75.4 | 66.5 | 84.0 | 74.2 | 65.2 | 71.8 | 79.3 | 85.0 | 54.7 |
| - | ✓ | ✓ | 76.5 | 78.0 | 65.5 | 74.9 | 65.3 | 83.7 | 74.2 | 65.1 | 71.5 | 78.1 | 83.5 | 53.8 |
| ✓ | ✓ | ✓ | **79.7*** | **81.6*** | **70.0*** | **77.6*** | **69.6*** | **86.5*** | **76.3*** | **67.3*** | **73.7*** | **81.8*** | **88.4*** | **57.4*** |

## 6.2. Ablation Test

Table 4 presents an ablation study on the effectiveness of our proposed losses and their combinations, evaluated under DG protocol. Among them, regarding the variant related to $\mathcal{L}_{GCLS}$, we replace it with the common cross-entropy loss as the main task optimization objective. Firstly, the inclusion of the $\mathcal{L}_{GCLS}$ yields notable improvements in accuracy and other metrics on all datasets. For instance, on the DDCE dataset, the accuracy increases from 72.0% to 75.1% with the addition of $\mathcal{L}_{GCLS}$, marking a 4.3% improvement. Similarly, on the LAD dataset, accuracy rises from 74.0% to 78.1%. These enhancements highlight the effectiveness of $\mathcal{L}_{GCLS}$ in handling noisy labels from self-reported questionnaires. The $\mathcal{L}_{Contrast}$ also demonstrates a significant positive impact, particularly in enhancing the stability of the model when processing continuous signal inputs. For example, the F1 score on the AdVitam dataset improves from 59.5% to 61.2% with the application of $\mathcal{L}_{Contrast}$. Although the increase is modest, it is noteworthy given the challenge of multi-domain generalization. $\mathcal{L}_{Center}$ is designed to centralize the feature space of different classes and enlarge the inter-class distances. The introduction of $\mathcal{L}_{Center}$ results in a marked improvement in sensitivity on the LAD, rising from 84.0% to 89.2%. This indicates the effectiveness of $\mathcal{L}_{Center}$ in distinguishing between drowsy and awake states.

When all three loss functions are applied together, the model achieves the best performance across all datasets. For instance, on the DDCE dataset, both accuracy and F1 score peak at 79.7% and 81.6%, respectively; while on the MCDD dataset, specificity improves from 71.5% to 73.7%. These significant improvements (p-value<0.05) demonstrate that the combination of the three loss functions enhances generalization capability and robustness of the model. These results are consistent with the intended division of labor among $\mathcal{L}_{GCLS}$, $\mathcal{L}_{Center}$, and $\mathcal{L}_{Contrast}$, i.e., label ambiguity handling, global semantic alignment, and local invariance enhancement, respectively.

## 6.3. Effect of Different Inputs

We present a quantitative analysis of the impact of different physiological signal combinations on the performance of driver drowsiness detection models, evaluated under DG protocol (Figure 4). It is evident that different combinations of input signals affect the performance of driver drowsiness detection models. Across all domains, the full model utilizing EDA, ECG, and RESP signals demonstrates the best performance, highlighting the importance of multivariate information fusion. Specifically, on the DDCE domain, this combination achieves an accuracy of 79.7% and an F1 score of 81.6%. This combination shows similar superiority across other domains.

As for models based on a single signal, ECG appears to play a more crucial role in most cases. For instance, on the AdVitam and DDCE domains, the accuracy with ECG alone is notably higher than using EDA or RESP alone. This suggests that cardiac activities are informative of drowsiness, which is aligned with previous works (Freitas et al., 2024; Fujiwara, Abe, Kamata, Nakayama, Suzuki, Yamakawa, Hiraoka, Kano, Sumi, Masuda et al., 2018). Moreover, the combination of signals also affects the stability of model performance. In terms of variance, the combination of three signals not only provides the best performance but also exhibits smaller variance across different domains, indicating more stable performance. In contrast, single signals show relatively larger variances, which suggests multivariate signal modeling can also enhance the robustness of the model.

Among all two-signal combinations, ECG+RESP is superior. In several domains, this combination performs second best to the full model. For instance, on the AdVitam domain, this combination achieves an accuracy of 75.8% and an F1 score of 69.0%, which, although lower than the full model, is better than other combinations. Despite it, in LAD, the best two-signal combination is EDA+RESP. This suggests that the optimal combination under different

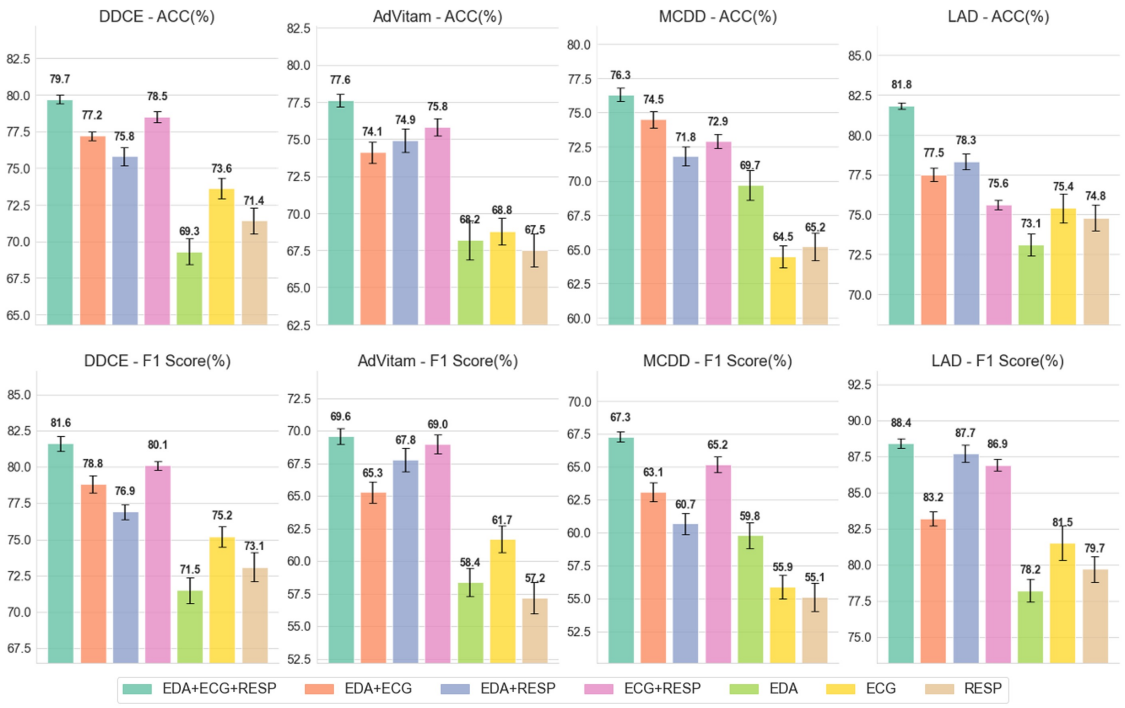


**Figure 4:** Accuracies and F1 scores of the DrowsyDG-Phys with different combinations of physiological signals

**Table 5**
Ablation study with different time windows on target domains.

| | DDCE | | AdVitam | | MCDD | | LAD | |
|---|---|---|---|---|---|---|---|---|
| W | ACC | F1 | ACC | F1 | ACC | F1 | ACC | F1 |
| W=200 | 72.3±1.0 | 73.8±1.3 | 70.5±1.3 | 62.1±1.1 | 68.9±0.9 | 59.7±1.2 | 75.2±1.1 | 81.4±1.2 |
| W=400 | **79.7**±0.3 | **81.6**±0.5 | **77.6**±0.4 | **69.6**±0.6 | **76.3**±0.5 | 67.3±0.4 | 81.8±0.2 | **88.4**±0.3 |
| W=600 | 78.9±0.5 | 81.2±0.6 | 76.1±0.6 | 68.3±0.8 | 75.1±0.6 | **68.0**±0.7 | **82.5**±0.2 | 87.1±0.4 |
| W=800 | 77.6±0.5 | 79.1±0.9 | 74.8±0.6 | 66.7±1.0 | 73.2±0.9 | 64.5±0.9 | 79.3±0.7 | 85.7±0.8 |

fatigue types or rubrics may change. In all, according to our experimental results, inputting three signals is optimal over all domains.

## 6.4. Effect of Different Time Window Sizes

We assess the performance of models with different time window sizes ($W$) across four target domains. As shown in Table 5, as the time window increases, the generally reach optimal performance at around $W$=400, in terms of model accuracy and F1 score. For instance, on the DDCE domain, the accuracy at $W$=500 is 79.7%±0.4, while at $W$=600 or 800, the accuracy and F1 score drops. This suggests that although longer time windows can capture more temporal information, excessively long windows might lead to information redundancy or increased computational complexity, potentially affecting model performance.

Among all domains, the performance is most favorable at a time window of 400, indicating that this window size effectively balances information richness and computational efficiency. Additionally, the variance is relatively small across different domains at $W$=400, indicating stable model performance. Notably, even though $W$=600 provides slight performance improvements on certain domains, such as an accuracy of 82.5%±0.2 on the LAD domain, considering the overall performance and variance across all domains, $W$=400 is still the optimal choice.

## 6.5. Case Study: Robustness to Noisy Input

Considering the deployment environment, many factors (e.g. sensor failure, data transmission, motion, etc.) can lead to noisy and partially missing physiological signals, which can affect the model performance. Consequently, we

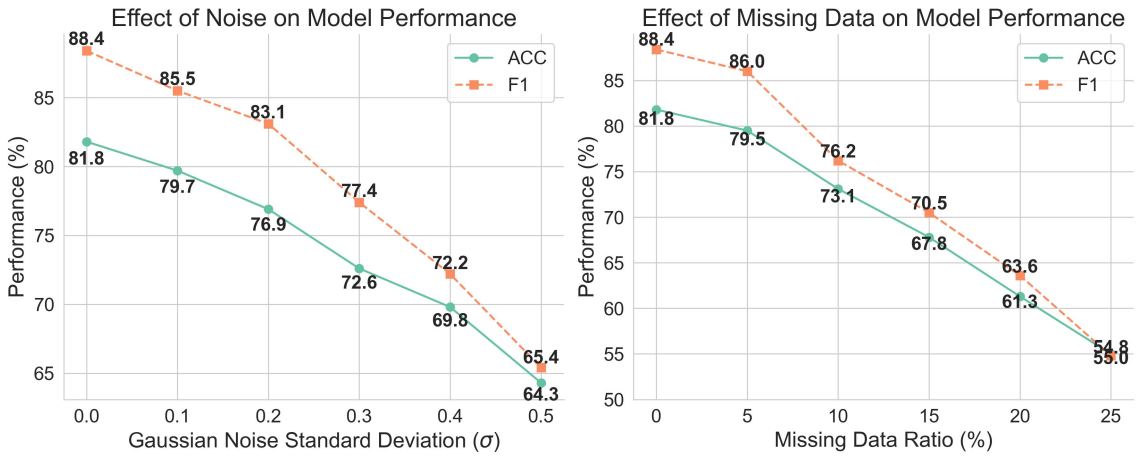


**Figure 5:** Comparison of estimation performance of DrowsyDG-Phys with varied information distortions.

**Table 6**
Comparison results of drowsiness estimation based on cross-subject protocol.

| | | Machine Learning | | | | Deep Learning | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Target** | **Metric(%)** | **LR** | **SVM** | **RF** | **LightGBM** | **GRU** | **MTCNN** | **CLSTM** | **DSCNN** | **Informer** | **TFormer** | **Ours** |
| DDCE | ACC | 77.7 | 40.7 | 66.6 | 40.7 | 65.2 | 78.5 | 73.8 | 80.1 | 82.3 | 83.9 | **85.3** |
| | F1 | 78.5 | 0.0 | 64.0 | 0.0 | 63.5 | 76.8 | 71.2 | 78.9 | 80.7 | 83.5 | **87.2** |
| | SEN | 68.7 | 0.0 | 50.0 | 0.0 | 58.9 | 72.4 | 67.3 | 76.5 | 79.1 | 81.7 | **94.5** |
| | SPE | 90.9 | 100.0 | 90.9 | 100.0 | 88.2 | 85.7 | 89.4 | 87.6 | 89.8 | 90.5 | **89.3** |
| | PRE | 91.5 | 0.0 | 88.9 | 0.0 | 68.9 | 81.8 | 75.6 | 81.5 | 82.4 | 85.4 | **80.9** |
| AdVitam | ACC | 66.8 | 52.7 | 76.4 | 75.1 | 75.0 | 80.2 | 77.6 | 82.7 | 85.1 | 87.6 | **90.0** |
| | F1 | 65.2 | 22.2 | 76.3 | 74.5 | 73.8 | 77.5 | 75.1 | 79.3 | 82.4 | 84.9 | **79.6** |
| | SEN | 58.9 | 12.8 | 70.7 | 66.9 | 68.5 | 73.9 | 69.8 | 75.2 | 77.6 | 80.3 | **71.3** |
| | SPE | 75.7 | 97.1 | 88.8 | 90.3 | 86.4 | 83.2 | 87.1 | 85.0 | 88.7 | 89.5 | **91.2** |
| | PRE | 73.0 | 84.2 | 82.9 | 84.0 | 80.0 | 81.5 | 81.3 | 83.9 | 87.8 | 90.1 | **90.1** |
| MCDD | ACC | 71.4 | 67.3 | 78.5 | 66.3 | 76.8 | 80.7 | 78.2 | 82.4 | 84.0 | 85.2 | **85.7** |
| | F1 | 54.8 | 15.8 | 61.8 | 45.9 | 58.3 | 65.7 | 62.4 | 68.9 | 71.5 | 73.8 | **76.2** |
| | SEN | 50.0 | 8.8 | 50.0 | 41.1 | 47.2 | 55.6 | 52.1 | 60.8 | 63.4 | 66.7 | **70.4** |
| | SPE | 82.8 | 98.3 | 93.7 | 79.7 | 90.5 | 88.2 | 91.0 | 89.3 | 90.8 | 91.7 | **80.0** |
| | PRE | 60.6 | 76.4 | 81.0 | 52.0 | 76.2 | 80.3 | 77.8 | 79.5 | 82.0 | 82.6 | **83.1** |
| LAD | ACC | 88.8 | 80.9 | 78.7 | 80.2 | 85.2 | 87.6 | 86.1 | 89.3 | 90.5 | 91.8 | **92.7** |
| | F1 | 92.7 | 87.6 | 75.0 | 79.4 | 89.3 | 90.8 | 89.5 | 91.2 | 92.4 | 94.0 | **95.1** |
| | SEN | 95.6 | 89.9 | 60.3 | 64.4 | 92.7 | 93.4 | 91.8 | 94.1 | 94.8 | 94.9 | **95.6** |
| | SPE | 68.4 | 53.9 | 91.6 | 90.2 | 82.1 | 79.8 | 84.3 | 83.6 | 85.0 | 86.7 | **84.2** |
| | PRE | 89.9 | 85.4 | 99.1 | 100.0 | 86.1 | 88.3 | 87.3 | 88.5 | 90.1 | 93.1 | **94.6** |

test the robustness of the model to information distortion in the operational environment by applying different degrees of Gaussian white noise (i.e., adding Gaussian white noise with different standard deviations $\sigma$ to original physiological signals) and missing data (i.e., randomly replacing some segments of the input signals with zeros) to each of the three physiological signals in the target domain.

Figure 5 illustrates the impact of information distortions, specifically Gaussian white noise with $\sigma$ from 0.0 to 0.5 and missing data with ratio from 0% to 25%, on our proposed model. The left graph demonstrates a clear decline in model performance as the standard deviation of Gaussian noise increases, with accuracy dropping from 81.8% at $\sigma = 0.0$ to 64.3% at $\sigma = 0.5$, and the F1 score decreasing from 88.4% to 65.4%. At the same time, the right graph reveals a similar downward trajectory in performance with increasing missing data ratios, where accuracy reduces from 81.8% at a 0% missing data ratio to 54.8% at a 25% ratio, and the F1 score falls from 88.4% to 55.0%. We note that the performance of the model falls to the level of the ML learning model at $\sigma$ of 0.5, or a missing ratio of 20%. These findings underscore the critical need for strategies to handle incomplete data, as even moderate levels of missing information can significantly degrade model performance. Furthermore, we observe that the model exhibits greater robustness to Gaussian noise than to missing data. We attribute this to the explicit incorporation of frequency-domain features in our modeling approach. Specifically, Gaussian noise introduced in the time domain can be effectively

disentangled from meaningful semantic information when processed in the frequency domain. In contrast, missing data leads to information loss in both the time and frequency domains, resulting in a more substantial decline in performance.

## 6.6. Case Study: Cross-Subject Protocol

Furthermore, we also evaluate our model based on the cross-subject evaluation to illustrate its generalizability to different individuals. The cross-subject evaluation protocol trains the model with 60% subjects in one dataset, and assesses the model under the rest 40% subjects. It is worth noting that since it does not fall within the scope of DG, we did not compare our model with other DG approaches.

According to Table 6, our proposed model outperforms across all domains. Overall, DL models generally outperform traditional ML models, due to their ability to capture complex temporal features and nonlinear relationships more effectively. However, it is noteworthy that in certain instances, individual ML models (such as LightGBM) perform comparably or even better than some DL models (like GRU) on specific domains, indicating that ML models still hold competitive advantages in scenarios that might with fewer variations. Comparing ML and DL models overall, DL models generally exhibit superior performance, particularly in processing complex physiological signals. However, the GRU, a specific DL model, underperforms compared to some ML models (such as LightGBM) on certain domains, possibly due to its sensitivity to features and the complexity of the domain.

On the LAD domain, which used objective labeling standards, all models generally perform well (i.e., most model accuracies and F1 scores are higher than 85%) compared to the other three domains with subjective assessment. It shows that domains labeled by objective assessment are easier to achieve generalization compared to subjective assessment. In general, our proposal shows superior performance in all domains, which indicates the robustness of our model and its adaptability to diverse labeling standards.

## 6.7. Implications

Although this work is evaluated in a research setting, the results provide several implications for building practical driver monitoring systems based on physiological signals. First, the comparisons of input features in Figure 4 suggest that multi-signal fusion is beneficial not only for improving performance but also for stabilizing the model across domains. In practice, this implies that a multi-channel configuration (ECG, EDA, and RESP) can be adopted when possible, while ECG-centric configurations may serve as a feasible alternative when only limited sensors are available. The variability observed across domains for different signal combinations also indicates that the optimal sensing configuration may depend on the dominant fatigue type and the assessment rubric in the target deployment environment.

Second, the time-window analysis in Table 5 indicates that the model performance does not monotonically increase with longer windows, and the best overall results are achieved around $W = 400$, with relatively small variance across domains. This observation is relevant to deployment because longer windows typically introduce additional latency and computational cost. Therefore, selecting an appropriate window size that balances information richness and responsiveness is necessary when the system is expected to generate timely warnings in real driving scenarios.

Third, the comparison under the DG protocol shows that the model performance can degrade substantially when tested on an unseen dataset, and that domain generalization brings benefits as compared to directly applying a model trained on a single dataset. This highlights a practical benefit of multi-source training for deployment: it can reduce the need for frequent re-collection and re-training when the operational domain changes due to different sensor configurations, vehicle platforms, populations, and drowsiness inducers. In addition, the performance gap between domains labeled by subjective questionnaires and those labeled by objective indicators in Table 6 suggests that label heterogeneity is not only a modeling challenge but also a deployment challenge, because industrial datasets are often collected under different rubrics over time. In this context, losses that explicitly handle noisy and uncertain labels can make the reuse of heterogeneous datasets possible and mitigate performance instability caused by rubric changes.

Finally, the robustness study in Figure 5 shows that the model is more sensitive to missing data than to additive Gaussian noise. This result has direct functional implications: deployment should prioritize the integrity of signals and avoid incomplete segments to avoid sharp performance degradation. For example, a practical system can include channel-level signal quality checks and have conservative decision logic that down-weights unreliable channels or delays decisions when the input integrity is compromised. These mechanisms are particularly relevant for automated driving scenarios (e.g., SAE Level-3), where the system is expected to provide stable driver state estimation under diverse operational conditions even with intermittent sensing failures.

## 7. Limitations

There are a few limitations to this study. First, due to the limited availability of public datasets, the dataset used in this study comprises three datasets labeled based on subjective questionnaires and only one labeled with objective indicators. Future research should incorporate additional datasets into the DG evaluation protocol, particularly those labeled using objective criteria.

Second, while the proposed DrowsyDG-Phys demonstrates superior performance in supervised pre-training scenarios with multi-source datasets, its performance in real-world deployment remains unknown. Specifically, all datasets were collected in simulator-based or controlled experimental settings, and therefore may not fully reflect challenges in long-term on-road monitoring, such as sensor drift, uncontrolled motion artifacts, and evolving lighting conditions. Thus, although our findings can be viewed as a step toward real-world deployment, future work should validate the proposed framework on naturalistic driving data. Additionally, alternative approaches other than DG methodologies should be explored. For example, unsupervised learning or few-shot learning (Yang, Wang, Fan, Liu, Su, Guo, Yu, He and Wu, 2025b; Wang, Yang, Lu, He and Wu, 2026) may also enhance generalizability by reducing reliance on biased input data.

Finally, we adopted the widely used KSS dichotomization to ensure comparability with prior studies; however, due to the inherent ambiguity between self-reported sleepiness states, different works have explored alternative cutoffs or multi-level/continuous modeling of KSS (Hultman et al., 2021; Ahlstrom, Fors, Anund and Hallvig, 2015). Thus, although a systematic evaluation of threshold sensitivity on the newly collected dataset is beyond the scope of this study, it constitutes an important direction for future research.

## 8. Conclusion

In this paper, we propose a novel generalizable framework DrowsyDG-Phys, incorporating three wearable physiological sensor data for driver drowsiness estimation. The proposed model explicitly extracts time and frequency-domain information to increase the robustness of noisy input. Besides, we introduce a feature centralization regularization and a contrastive regularization with temporal consistency prior knowledge to promote the discrimination ability of the model to different states and maintain the stability of model output. Lastly, corresponding to the shift in labeling criteria, we incorporate a generalizable classification loss as the main optimization goal. According to our extensive experiments on four datasets, we find that:

- ML methods with manual feature extraction perform significantly worse than DL models on unseen domains. Among DL architectures, the transformer model outperforms RNNs and CNNs for the task addressed in this paper. Future research should continue exploring generalizable DL architectures.

- DG methods do not always lead to performance improvements. Among the DG methods adopted in this study, two plug-and-play approaches that are independent from explicit domain labeling, VREx and our proposal yield superior results. Future work should explore tailored DG methods designed for specific tasks.

- Different combinations of physiological signal inputs influence model performance, with varying effects across domains. Notably, cardiac activity, represented by ECG signals, plays a crucial role in enhancing model efficacy. Future research should focus on maintaining high performance while relying on fewer inputs, thereby reducing model complexity.

- By explicitly incorporating frequency-domain information, our model demonstrates strong robustness against varying levels of Gaussian noise in the input. However, it suffers significant performance degradation in the presence of missing data. Future work should further investigate methods to enhance robustness under missing data scenarios.

- Both ML and DL models perform better in intra-domain cross-subject protocols. However, in real-world applications, test samples encountered by the model may not share the same data distribution as the training set. Therefore, we encourage future research to further evaluate model performance under the DG protocol to improve real-world applicability.

## Acknowledgments

This work was supported by the Natural Science Foundation of Guangdong Province of China (2024A1515010392).

## References

Ahlström, C., Anund, A., 2024. Development of sleepiness in professional truck drivers: Real-road testing for driver drowsiness and attention warning (ddaw) system evaluation. Journal of sleep research , e14259.

Ahlstrom, C., Fors, C., Anund, A., Hallvig, D., 2015. Video-based observer rated sleepiness versus self-reported subjective sleepiness in real road driving. European Transport Research Review 7, 38.

Alguindigue, J., Singh, A., Narayan, A., Samuel, S., 2024. Biosignals monitoring for driver drowsiness detection using deep neural networks. IEEE Access .

Angelova, R., Markov, D., Simova, I., Velichkova, R., Stankov, P., 2019. Accumulation of metabolic carbon dioxide (co2) in a vehicle cabin, in: IOP Conference Series: Materials Science and Engineering, IOP Publishing. p. 012010.

Ashraf, I., Hur, S., Shafiq, M., Park, Y., 2019. Catastrophic factors involved in road accidents: Underlying causes and descriptive analysis. PLoS one 14, e0223473.

Awais, M., Badruddin, N., Drieberg, M., 2017a. A hybrid approach to detect driver drowsiness utilizing physiological signals to improve system performance and wearability. Sensors 17, 1991.

Awais, M., Badruddin, N., Drieberg, M., 2017b. A hybrid approach to detect driver drowsiness utilizing physiological signals to improve system performance and wearability. Sensors 17, 1991.

Ayas, S., Donmez, B., Tang, X., 2024a. Drowsiness mitigation through driver state monitoring systems: a scoping review. Human factors 66, 2218–2243.

Ayas, S., He, D., Donmez, B., 2024b. Differentiating high and low cognitive load using driving performance and driver physiological data, in: [Conference Presentation] ASPIRE International Annual Meeting, Human Factors and Ergonomics Society.

Ballas, A., Diou, C., 2024. Towards domain generalization for ecg and eeg classification: Algorithms and benchmarks. IEEE Transactions on Emerging Topics in Computational Intelligence 8, 44–54. doi:10.1109/TETCI.2023.3306253.

Bao, Y., Xu, W., 2024. Design and implementation of a fatigue detection system based on dlib for driver facial features, in: Electronics, Communications and Networks. IOS Press, pp. 792–799.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Buendia, R., Forcolin, F., Karlsson, J., Arne Sjöqvist, B., Anund, A., Candefjord, S., 2019. Deriving heart rate variability indices from cardiac monitoring—an indicator of driver sleepiness. Traffic injury prevention 20, 249–254.

Cheon, S.P., Kang, S.J., 2017. Sensor-based driver condition recognition using support vector machine for the detection of driver drowsiness, in: 2017 IEEE Intelligent Vehicles Symposium (IV), IEEE. pp. 1517–1522.

Chowdhury, A., Shankaran, R., Kavakli, M., Haque, M.M., 2018. Sensor applications and physiological features in drivers' drowsiness detection: A review. IEEE sensors Journal 18, 3055–3067.

Cos, C.A., Lambert, A., Soni, A., Jeridi, H., Thieulin, C., Jaouadi, A., 2023. Enhancing mental fatigue detection through physiological signals and machine learning using contextual insights and efficient modelling. Journal of Sensor and Actuator Networks 12, 77.

Cui, J., Lan, Z., Sourina, O., Müller-Wittig, W., 2022. Eeg-based cross-subject driver drowsiness recognition with an interpretable convolutional neural network. IEEE Transactions on Neural Networks and Learning Systems 34, 7921–7933.

Dey, R., Salem, F.M., 2017. Gate-variants of gated recurrent unit (gru) neural networks, in: 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS), IEEE. pp. 1597–1600.

Díaz-Santos, S., Caballero-Gil, P., Caballero-Gil, C., 2024. Ai-driven wearables for driver health and safety, in: International Conference on Ubiquitous Computing and Ambient Intelligence, Springer. pp. 375–380.

Fan, C., Peng, Y., Peng, S., Zhang, H., Wu, Y., Kwong, S., 2021. Detection of train driver fatigue and distraction based on forehead eeg: a time-series ensemble learning method. IEEE transactions on intelligent transportation systems 23, 13559–13569.

Fanger, P.O., 1970. Thermal comfort. analysis and applications in environmental engineering. .

Feng, X., Guo, Z., Kwong, S., 2025. Id3rsnet: cross-subject driver drowsiness detection from raw single-channel eeg with an interpretable residual shrinkage network. Frontiers in Neuroscience 18, 1508747.

Freitas, A., Almeida, R., Gonçalves, H., Conceição, G., Freitas, A., 2024. Monitoring fatigue and drowsiness in motor vehicle occupants using electrocardiogram and heart rate- a systematic review. Transportation research part F: traffic psychology and behaviour 103, 586–607.

Fujiwara, K., Abe, E., Kamata, K., Nakayama, C., Suzuki, Y., Yamakawa, T., Hiraoka, T., Kano, M., Sumi, Y., Masuda, F., et al., 2018. Heart rate variability-based driver drowsiness detection and its validation with eeg. IEEE transactions on biomedical engineering 66, 1769–1778.

Ganin, Y., Lempitsky, V., 2015. Unsupervised domain adaptation by backpropagation, in: International conference on machine learning, PMLR. pp. 1180–1189.

Gao, Z., Wang, X., Yang, Y., Mu, C., Cai, Q., Dang, W., Zuo, S., 2019. Eeg-based spatio–temporal convolutional neural network for driver fatigue evaluation. IEEE Transactions on Neural Networks and Learning Systems 30, 2755–2763. doi:10.1109/TNNLS.2018.2886414.

Gottlieb, D.J., Whitney, C.W., Bonekat, W.H., Iber, C., James, G.D., Lebowitz, M., Nieto, F.J., Rosenberg, C.E., 1999. Relation of sleepiness to respiratory disturbance index: the sleep heart health study. American journal of respiratory and critical care medicine 159, 502–507.

Guo, Y., Yang, K., Wu, Y., 2025. A multi-modality attention network for driver fatigue detection based on frontal eeg, eda and ppg signals. IEEE Journal of Biomedical and Health Informatics .

Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B., 1998. Support vector machines. IEEE Intelligent Systems and their applications 13, 18–28.

Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied logistic regression. John Wiley & Sons.

Hultman, M., Johansson, I., Lindqvist, F., Ahlström, C., 2021. Driver sleepiness detection with deep neural networks using electrophysiological data. Physiological measurement 42, 034001.

Hussein, R.M., Miften, F.S., George, L.E., 2023. Driver drowsiness detection methods using eeg signals: a systematic review. Computer methods in biomechanics and biomedical engineering 26, 1237–1249.

Javed, A., Arshad, M.U., Saeed, E., Naseer, N., 2021. Real-time drowsiness detection and emergency parking using eeg .

Jeppesen, J., Fuglsang-Frederiksen, A., Johansen, P., Christensen, J., Wüstenhagen, S., Tankisi, H., Qerama, E., Beniczky, S., 2020. Seizure detection using heart rate variability: a prospective validation study. Epilepsia 61, S41–S46.

Jiao, Y., Zhang, C., Chen, X., Fu, L., Jiang, C., Wen, C., 2024. Driver fatigue detection using measures of heart rate variability and electrodermal activity. IEEE Transactions on Intelligent Transportation Systems 25, 5510–5524. doi:10.1109/TITS.2023.3333252.

Kaida, K., Takahashi, M., Åkerstedt, T., Nakata, A., Otsuka, Y., Haratani, T., Fukasawa, K., 2006. Validation of the karolinska sleepiness scale against performance and eeg variables. Clinical neurophysiology 117, 1574–1581.

Kakhi, K., Jagatheesaperumal, S.K., Khosravi, A., Alizadehsani, R., Acharya, U.R., 2024. Fatigue monitoring using wearables and ai: Trends, challenges, and future opportunities. arXiv preprint arXiv:2412.16847 .

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 30.

Khushaba, R.N., Kodagoda, S., Lal, S., Dissanayake, G., 2010. Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm. IEEE transactions on biomedical engineering 58, 121–131.

Kim, D.Y., Han, D.K., Jeong, J.H., Lee, S.W., 2025. Calibration-free driver drowsiness classification with prototype-based multi-domain mixup. IEEE Transactions on Intelligent Transportation Systems .

Krueger, D., Caballero, E., Jacobsen, J.H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., Courville, A., 2021. Out-of-distribution generalization via risk extrapolation (rex), in: International Conference on Machine Learning, PMLR. pp. 5815–5826.

Kundinger, T., Mayr, C., Riener, A., 2020a. Towards a reliable ground truth for drowsiness: A complexity analysis on the example of driver fatigue 4. URL: https://doi.org/10.1145/3394980, doi:10.1145/3394980.

Kundinger, T., Sofra, N., Riener, A., 2020b. Assessment of the potential of wrist-worn wearable sensors for driver drowsiness detection. Sensors 20, 1029.

Lee, B.G., Chung, W.Y., 2012. Driver alertness monitoring using fusion of facial features and bio-signals. IEEE Sensors Journal 12, 2416–2422. doi:10.1109/JSEN.2012.2190505.

Leonhardt, S., Leicht, L., Teichmann, D., 2018. Unobtrusive vital sign monitoring in automotive environments—a review. Sensors 18, 3080.

Li, R., Hu, M., Gao, R., Wang, L., Suganthan, P.N., Sourina, O., 2024. Tformer: A time–frequency transformer with batch normalization for driver fatigue recognition. Advanced Engineering Informatics 62, 102575.

Lu, J., Zheng, X., Tang, L., Zhang, T., Sheng, Q.Z., Wang, C., Jin, J., Yu, S., Zhou, W., 2021. Can steering wheel detect your driving fatigue? IEEE Transactions on Vehicular Technology 70, 5537–5550. doi:10.1109/TVT.2021.3072936.

Lu, X., Tan, H., Zhang, H., Wang, W., Xie, S., Yue, T., Chen, F., 2025. Triboelectric sensor gloves for real-time behavior identification and takeover time adjustment in conditionally automated vehicles. Nature Communications 16, 1080.

Lyu, X., Akbar, M.A., Manimurugan, S., Jiang, H., 2025. Driver fatigue warning based on medical physiological signal monitoring for transportation cyber-physical systems. IEEE Transactions on Intelligent Transportation Systems , 1–13doi:10.1109/TITS.2025.3540895.

Ma, C., Zhang, M., Sun, X., Wang, H., Gao, Z., 2023. Dynamic threshold distribution domain adaptation network: A cross-subject fatigue recognition method based on eeg signals. IEEE Transactions on Cognitive and Developmental Systems 16, 190–201.

Ma, S., Yan, X., Billington, J., Merat, N., Markkula, G., 2024. Cognitive load during driving: Eeg microstate metrics are sensitive to task difficulty and predict safety outcomes. Accident Analysis & Prevention 207, 107769.

Malathi, D., Jayaseeli, J.D., Madhuri, S., Senthilkumar, K., 2018. Electrodermal activity based wearable device for drowsy drivers, in: Journal of Physics: Conference Series, IOP Publishing. p. 012048.

Meteier, Q., Capallera, M., De Salis, E., Angelini, L., Carrino, S., Widmer, M., Abou Khaled, O., Mugellini, E., Sonderegger, A., 2023. A dataset on the physiological state and behavior of drivers in conditionally automated driving. Data in brief 47, 109027.

Meteier, Q., Capallera, M., Ruffieux, S., Angelini, L., Abou Khaled, O., Mugellini, E., Widmer, M., Sonderegger, A., 2021. Classification of drivers' workload using physiological signals in conditional automation. Frontiers in psychology 12, 596038.

Meteier, Q., Favre, R., Viola, S., Capallera, M., Angelini, L., Mugellini, E., Sonderegger, A., 2024. Classification of driver fatigue in conditionally automated driving using physiological signals and machine learning. Transportation Research Interdisciplinary Perspectives 26, 101148.

Mu, S., Liao, S., Tao, K., Shen, Y., 2024. Intelligent fatigue detection based on hierarchical multi-scale ecg representations and hrv measures. Biomedical Signal Processing and Control 92, 106127.

Patel, M., Lal, S.K., Kavanagh, D., Rossiter, P., 2011. Applying neural network analysis on heart rate variability data to assess driver fatigue. Expert systems with Applications 38, 7235–7242.

Paulhus, D.L., Vazire, S., et al., 2007. The self-report method. Handbook of research methods in personality psychology 1, 224–239.

Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P., 2020. Distributionally robust neural networks, in: International Conference on Learning Representations.

Saleem, A.A., Siddiqui, H.U.R., Raza, M.A., Rustam, F., Dudley, S., Ashraf, I., 2023. A systematic review of physiological signals based driver drowsiness detection systems. Cognitive neurodynamics 17, 1229–1259.

Sang-Joong, J., Heung-Sub, S., Wan-Young, C., 2014. Driver fatigue and drowsiness monitoring system with embedded electrocardiogram sensor on steering wheel. IET Intelligent Transport Systems 8.

Slater, J.D., 2008. A definition of drowsiness: One purpose for sleep? Medical Hypotheses 71, 641–644.

Song, Y.Y., Ying, L., 2015. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry 27, 130.

Sun, Z., Li, X., 2024. Contrast-phys+: Unsupervised and weakly-supervised video-based remote physiological measurement via spatiotemporal contrast. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Sunagawa, M., Shikii, S.i., Beck, A., Kek, K.J., Yoshioka, M., 2023. Analysis of the effect of thermal comfort on driver drowsiness progress with predicted mean vote: An experiment using real highway driving conditions. Transportation research part F: traffic psychology and behaviour 94, 517–527.

Tartarini, F., Schiavon, S., Cheung, T., Hoyt, T., 2020. Cbe thermal comfort tool: Online tool for thermal comfort calculations and visualizations. SoftwareX 12, 100563.

Tefft, B.C., 2012. Prevalence of motor vehicle crashes involving drowsy drivers, united states, 1999–2008. Accident Analysis & Prevention 45, 180–186.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

Villarejo, M.V., Zapirain, B.G., Zorrilla, A.M., 2012. A stress sensor based on galvanic skin response (gsr) controlled by zigbee. Sensors 12, 6075–6101.

Wang, F., Gu, T., Yao, W., 2024a. Research on the application of the sleep eeg net model based on domain adaptation transfer in the detection of driving fatigue. Biomedical Signal Processing and Control 90, 105832.

Wang, J., Ayas, S., Zhang, J., Wen, X., He, D., Donmez, B., 2025a. Towards generalizable drowsiness monitoring with physiological sensors: A preliminary study, in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, SAGE Publications Sage CA: Los Angeles, CA. p. 10711813251376962.

Wang, J., Lu, H., Han, H., Chen, Y., He, D., Wu, K., 2025b. Generalizable remote physiological measurement via semantic-sheltered alignment and plausible style randomization. IEEE Transactions on Instrumentation and Measurement 74, 1–14. doi:10.1109/TIM.2024.3497058.

Wang, J., Lu, H., Wang, A., Chen, Y., He, D., 2024b. Hierarchical style-aware domain generalization for remote physiological measurement. IEEE Journal of Biomedical and Health Informatics 28, 1635–1643. doi:10.1109/JBHI.2023.3346057.

Wang, J., Lu, H., Wang, A., Yang, X., Chen, Y., He, D., Wu, K., 2025c. Physmle: Generalizable and priors-inclusive multi-task remote physiological measurement. IEEE Transactions on Pattern Analysis and Machine Intelligence , 1–19doi:10.1109/TPAMI.2025.3545598.

Wang, J., Wang, A., Hu, H., Wu, K., He, D., 2024c. Multi-source domain generalization for ecg-based cognitive load estimation: Adversarial invariant and plausible uncertainty learning, in: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1631–1635. doi:10.1109/ICASSP48485.2024.10447676.

Wang, J., Wang, A., Yan, S., He, D., Wu, K., 2024d. Revisiting interactions of multiple driver states in heterogenous population and cognitive tasks. arXiv preprint arXiv:2412.13574 .

Wang, J., Yang, X., Hu, Q., Tang, J., Liu, C., He, D., Wang, Y., Chen, Y.C., Wu, K., . Physdrive: A multimodal remote physiological measurement dataset for in-vehicle driver monitoring, in: The Thirty-ninth Annual Conference on Neural Information Processing Systems.

Wang, J., Yang, X., Lu, H., He, D., Wu, K., 2026. Align the gap: Prior-based unified multi-task remote physiological measurement framework for domain generalization and personalization. International Journal of Computer Vision .

Wang, J., Yang, X., Wang, Z., Wei, X., Wang, A., He, D., Wu, K., 2024e. Efficient mixture-of-expert for video-based driver state and physiological multi-task estimation in conditional autonomous driving. arXiv preprint arXiv:2410.21086 .

Wierwille, W.W., Ellsworth, L.A., 1994. Evaluation of driver drowsiness by trained raters. Accident Analysis & Prevention 26, 571–581.

Xie, Y., Murphey, Y.L., Kochhar, D.S., 2019. Personalized driver workload estimation using deep neural network learning from physiological and vehicle signals. IEEE Transactions on Intelligent Vehicles 5, 439–448.

Yang, C., Hu, M., Zhai, G., Zhang, X.P., 2022. Graph-based denoising for respiration and heart rate estimation during sleep in thermal video. IEEE Internet of Things Journal 9, 15697–15713. doi:10.1109/JIOT.2022.3150147.

Yang, L., Yang, H., Wei, H., Hu, Z., Lv, C., 2024. Video-based driver drowsiness detection with optimised utilization of key facial features. IEEE Transactions on Intelligent Transportation Systems 25, 6938–6950.

Yang, X., He, D., Wang, J., Wu, K., 2025a. Fedhug: Federated heterogeneous unsupervised generalization for remote physiological measurements. arXiv preprint arXiv:2510.12132 .

Yang, X., Wang, J., Fan, Y., Liu, C., Su, H., Guo, W., Yu, Z., He, D., Wu, K., 2025b. Not only consistency: Enhance test-time adaptation with spatio-temporal inconsistency for remote physiological measurement. arXiv preprint arXiv:2507.07908 .

Yuan, L., Cui, J., Li, R., Zheng, Z., Siyal, M.Y., Yi, Z., 2024. Entropy-guided robust feature domain adaptation for electroencephalogram-based cross-dataset drowsiness recognition. Engineering Applications of Artificial Intelligence 137, 109153.

Zeng, H., Li, X., Borghini, G., Zhao, Y., Aricò, P., Di Flumeri, G., Sciaraffa, N., Zakaria, W., Kong, W., Babiloni, F., 2021. An eeg-based transfer learning method for cross-subject fatigue mental state prediction. Sensors 21, 2369.

Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2022. Understanding deep learning requires rethinking generalization, in: International Conference on Learning Representations.

Zhang, Z., Sabuncu, M., 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems 31.

Zhao, Y., Jiang, C., Song, X., 2022. Seasonal patterns and semi-empirical modeling of in-vehicle exposure to carbon dioxide and airborne particulates in dalian, china. Atmospheric Environment 274, 118968.

Zhou, F., Alsaid, A., Blommer, M., Curry, R., Swaminathan, R., Kochhar, D., Talamonti, W., Tijerina, L., 2022. Predicting driver fatigue in monotonous automated driving with explanation using gpboost and shap. International Journal of Human–Computer Interaction 38, 719–729.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., Zhang, W., 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting, in: Proceedings of the AAAI conference on artificial intelligence, pp. 11106–11115.

Zhou, K., Yang, Y., Qiao, Y., Xiang, T., 2024. Mixstyle neural networks for domain generalization and adaptation. International Journal of Computer Vision 132, 822–836.