



Exploring Structure Incentive Domain Adversarial Learning for Generalizable Sleep Stage Classification

SHUO MA, Beijing Key Lab. of Mobile Computing and Pervasive Devices, Institute of Computing Technology, Chinese Academy of Sciences, University of Chinese Academy of Sciences, China

YINGWEI ZHANG and **YIQIANG CHEN***, Beijing Key Laboratory of Mobile Computing and Pervasive Devices, Institute of Computing Technology, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shandong Academy of Intelligent Computing Technology, China

TAO XIE, NIO, China

SHUCHAO SONG, Beijing Key Laboratory of Mobile Computing and Pervasive Devices, Institute of Computing Technology, Chinese Academy of Sciences, University of Chinese Academy of Sciences, China

ZIYU JIA, Brainnetome Center, Institute of Automation, Chinese Academy of Sciences, China

Sleep stage classification is crucial for sleep state monitoring and health interventions. In accordance with the standards prescribed by the American Academy of Sleep Medicine, a sleep episode follows a specific structure comprising five distinctive sleep stages that collectively form a sleep cycle. Typically, this cycle repeats about five times, providing an insightful portrayal of the subject's physiological attributes. The progress of deep learning and advanced domain generalization methods allows automatic and even adaptive sleep stage classification. However, applying models trained with visible subject data to invisible subject data remains challenging due to significant individual differences among subjects. Motivated by the periodic category-complete structure of sleep stage classification, we propose a Structure Incentive Domain Adversarial learning (SIDA) method that combines the sleep stage classification method with domain generalization to enable cross-subject sleep stage classification. SIDA includes individual domain discriminators for each sleep stage category to decouple subject dependence differences among different categories and fine-grained learning

This work was supported by the National Key Research and Development Plan of China (No. 2021YFC2501202), National Natural Science Foundation of China (No. 61972383 and No. 62302487), the Beijing Municipal Science & Technology Commission (No. Z221100002722009), and the Innovative Research Program of Shandong Academy of Intelligent Computing Technology (No. SDAICT2191010).

The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve University). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

Authors' addresses: S. Ma, Beijing Key Lab. of Mobile Computing and Pervasive Devices, Institute of Computing Technology, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing, China; e-mail: mashuo20g@ict.ac.cn; Y. Zhang and Y. Chen (Corresponding author), Beijing Key Lab. of Mobile Computing and Pervasive Devices, Institute of Computing Technology, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shandong Academy of Intelligent Computing Technology, Beijing, China; e-mails: {zhangyingwei, yqchen}@ict.ac.cn; T. Xie, NIO, Beijing, China; e-mail: xt.xie@nio.com; S. Song, Beijing Key Lab. of Mobile Computing and Pervasive Devices, Institute of Computing Technology, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing, China; e-mail: songshuchao22b@ict.ac.cn; Z. Jia, Brainnetome Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China; e-mail: jia.ziyu@outlook.com.



This work is licensed under a Creative Commons Attribution International 4.0 License.

© 2024 Copyright held by the owner/author(s).

2157-6904/2024/01-ART14

<https://doi.org/10.1145/3625238>

of domain-invariant features. Furthermore, SIDA directly connects the label classifier and domain discriminators to promote the training process. Experiments on three benchmark sleep stage classification datasets demonstrate that the proposed SIDA method outperforms other state-of-the-art sleep stage classification and domain generalization methods and achieves the best cross-subject sleep stage classification results.

CCS Concepts: • **Human-centered computing** → **Ubiquitous computing**; • **Computing methodologies** → **Transfer learning**;

Additional Key Words and Phrases: physiological signal, time-series signal, sleep stage classification, domain generalization, subject independent

1 INTRODUCTION

Sleep accounts for a significant portion of a human life, precisely one-third, and is directly related to one's physical and mental well-being. As a fundamental technique for disease monitoring [13], arrangement, and intervention, sleep stage classification has remarkable practical significance in healthcare [6]. The two principal standards governing sleep stage classification are the **Rechtschaffen & Kales (R&K)** criteria [49] and the **American Academy of Sleep Medicine (AASM)** criteria [4]. Based on these widely accepted international sleep stage classification standards, sleep monitoring is indispensable in many healthcare areas. Notably, brain disorders such as aphasia, epilepsy, and Parkinson's disease exhibit intricate and close associations with sleep disorders, prompting extensive research into the application of sleep monitoring in the intervention of brain disorders [43]. Christensen et al. [11] employed **electroencephalography (EEG)** monitoring equipment and data-driven analytical methods to reveal sleep characteristics in patients with insomnia. Coelli et al. [12] conducted benchmark research on sleep monitoring in epileptic patients, using a multiscale functional clustering approach to survey epileptic networks in various sleep stages. In Parkinson's disease, sleep disorders represent the most frequent non-motor symptoms, and monitoring sleep quality offers an effective way to anticipate Parkinson's disease onset and track disease progression [27].

The conventional method of sleep stage classification requires professional medical experts to manually analyze the **Polysomnography (PSG)** signals of subjects [51]. This approach is time-consuming, low in efficiency, and labour intensive. Moreover, this method's results are subjective and easily influenced by the expertise and experience of the analysts [26]. The development of artificial intelligence has led to the emergence of automatic sleep classification approaches that significantly improve accuracy and efficiency [26]. Typically, these methods extract time-frequency transformation features from the raw PSG signal and employ machine learning methods like Random Forest [38], **Support Vector Machine (SVM)** [1], and K-Nearest Neighbor [52] to build the final classification model. However, these methods require significant prior knowledge for feature extraction and processing. With the development of deep learning, the emergence of deep learning has brought many advancements in the accuracy and efficiency of sleep stage classification. Deep learning-based sleep stage classification methods employ end-to-end neural networks for feature extraction and model construction. **Convolutional neural networks (CNN)** have been employed to extract spatial sleep features from the PSG signal [40, 48]. Goshtasbi et al. proposed a fully convolutional neural network called SleepFCN [18], which utilizes residual dilated causal convolutions to capture temporal context information and thus enhances the accuracy and speed of recognition. **Recurrent neural networks (RNN)** have also been used to extract temporal features related to sleep from the PSG signal [8, 41, 54]. Furthermore, **Long Short-Term Memory (LSTM)** [15, 40] has been utilized to address the issue of forgetting over long-time series signals. Zhao et al. proposed SleepContextNet [57], which utilizes a CNN-LSTM model structure combined

with data augmentation techniques, significantly improving classification accuracy. Wang et al. [45] proposed a novel multi-scale attention mechanism incorporating channel and spatial attention, resulting in exceptional classification accuracy. Phan et al. proposed SeqSleepNet [33] to address the sleep stage classification problem as a sequence-to-sequence classification problem. To achieve interpretability at the epoch and sequence level and improve the accuracy of sleep stage classification, they further developed SleepTransformer [34], which is the first transformer-based sleep stage classification model and achieved state-of-the-art performance. To address the issue of heterogeneity among physiological signals, Zhu et al. proposed MaskSleepNet [59]. This model learns the joint distribution of mask and non-mask modalities by leveraging partial modalities of mask signals. It also uses multi-scale convolution and multi-head attention to extract features and make predictions at sub-scales, respectively. In addition, Researchers have utilized sparse autoencoders to categorize pre-extracted time-frequency features [44]. And some generative adversarial networks models are used for EEG and **electrocardiography (ECG)** signal generation to improve related classification tasks [17].

However, the abovementioned models are more suitable for extracting features from grid or image data. They do not utilize the functional connectivity relationship of brain structures in the PSG signal. Furthermore, the brain's cerebral cortex forms a non-Euclidean space, making it well suited for representing the feature distribution of brain space using a graph structure. Correspondingly, the **graph neural networks (GCN)** have been widely employed and worked well in graph-structured data [58]. Although existing studies have achieved acceptable sleep stage classification accuracy [21, 23, 28], these approaches have not addressed the challenge of PSG signal-based sleep stage classification, which depends on the combination of multiple physiological signals, including EEG, ECG, **electrooculography (EOG)**, and **electromyography (EMG)** signals, which vary significantly across different subjects [10]. For instance, the EEG signal can be affected by subjects' electrode drift and hair, while the EMG signal can be affected by muscle fatigue, skin resistance, and muscle strength of subjects [56]. The challenge of subject dependence limits the adaptability of sleep stage classification models, as models trained on certain subjects cannot be applied to new subjects. However, most existing methods only modify the feature extractor based on graph models without focusing on improving subject independence. Furthermore, obtaining and labeling sleep stage classification data is complex and requires professional medical expertise [39], making training a new model for each new subject with their data impractical.

Fortunately, the development of transfer learning has provided hope for achieving subject-independent sleep stage classification [30, 60]. Researchers have begun to focus on improving the generalization of the model. Jia et al. proposed the MSTGCN model [22], which integrates domain generalization [5, 46] and spatio-temporal GCN, using the **domain adversarial (DA)** method to improve the model's robustness across subjects. Tang et al. [42] employed the **Maximum Mean Discrepancy (MMD)** [19] method to reduce the distribution difference between the training set and the testing set data of the ECG signal. Most other transfer learning-based sleep stage classification methods utilize the pre-training and fine-tuning paradigm to enhance prediction accuracy [2]. However, this paradigm has many limitations due to the need for target data. Moreover, they ignored the structural characteristics of the sleep stage classification problem, resulting in unsatisfactory limited improvement results. To tackle the aforementioned challenges, we have fused the sleep stage classification problem with domain generalization [31], culminating in the proposal of a **Structure Incentive Domain Adversarial learning (SIDA)** method to augment subject generalization of the sleep stage classification model. As shown in Figure 1, the inspiration for the SIDA method came from the structure of the sleep cycle. During an entire sleep episode, there are typically five complete sleep cycles [16], each consisting of five stages

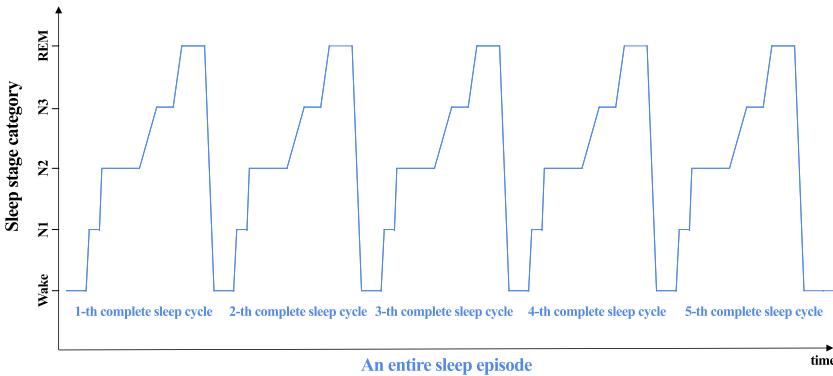


Fig. 1. The structure of the sleep cycle: Compared to other classification problems for physiological signals in time series, the sleep stage classification problem has a distinctive category-complete structure, along with a specific pattern in the category transition and cycle changes of sleep stages. Typically, during an entire sleep episode, there are five complete sleep cycles, each comprising five stages ranging from Wake to REM and back.

from the **Wakefulness (Wake)** stage to the **Rapid Eye Moment (REM)** stage and back [7]. The sleep stage categories themselves are limited and consist of five distinct stages, and each stage may exhibit unique subject dependencies. Furthermore, we generalize the problems caused by the above structure as the **Subject Dependency Differences of different sleep Categories (SDDC)** concept. More specifically, in contrast to traditional domain generalization models, SIDA establishes distinct domain (i.e., subject) discriminators for every sleep stage to dissociate the subject dependence differences amongst the various sleep stages. This strategy facilitates the model in precisely learning subject or domain invariant features. Moreover, we have bridged the sleep stage classifier and domain discriminators in SIDA with direct connections, positively influencing the training process. To our knowledge, this study marks the inaugural effort to define the SDDC notion precisely. Leveraging the PSG-based sleep stage classification's category structure, we introduce the SIDA method to attain optimal cross-subject sleep stage classification. Notably, we have utilized the leave-one-subject-out cross-validation method to rigorously validate our method. We have trained the classification model on the data from existing seen subjects and tested the efficacy of the trained model on the data of another unseen subject. Furthermore, we have validated and chosen the ultimate model on separate validation data that are randomly selected from training data. We have evaluated the effectiveness of the proposed SIDA method on three benchmark sleep stage classification datasets (i.e., ISRUC-S1 [24], ISRUC-S3 [24], and **Sleep Heart Health Study Visit 1 (SHHS1)** [36, 53]. The experimental results indicate that our proposed SIDA method outperforms other comparing methods and delivers the best cross-subject sleep stage classification results. In conclusion, the primary contributions of this study can be summarized as follows:

- We clearly define the SDDC concept and open up the idea of handling the challenge of subject dependence on the category from the perspective of transfer learning.
- We propose the SIDA method, which is a domain generalization method, to realize category-by-category subject dependency alignment and achieve direct soft weighting between the classifier and discriminators.
- Our proposed SIDA method is a plug-and-play method that can easily combine with existing methods. With experiments on the three public sleep stage classification datasets, the

extensive experiments demonstrate that the results of the existing sleep stage classification methods have been improved by combining them with our SIDA method.

2 RELATED WORK

The research of this article is mainly related to sleep stage classification and the domain generalization method. Therefore, this section will review these two parts and their intersection.

2.1 Sleep Stage Classification

Sleep stage classification is critical in monitoring and diagnosing sleep disorders. It involves collecting data during sleep and training models to classify different sleep stages. Medical experts use this information to diagnose and treat brain and neurological diseases. In 1968, Rechtschaffen and Kales proposed the R&K standard based on PSG collected during sleep, which divided sleep stages into seven stages: Wake, REM, four **Non-Rapid Eye Movement (NREM)** stages, and Movement Time stages. Four NREM stages consist of Stage 1 (S1), Stage 2 (S2), Stage 3 (S3), and Stage 4 (S4). Stages S1 and S2 are regarded as the light sleep stage, while stages S3 and S4 are regarded as the deep sleep stage, also known as slow-wave sleep. In 2007, the American Academy of Sleep Medicine merged the S3 and S4 stages in the R&K standard into stage S3 and recalled stages S1, S2, and S3 as stages N1, N2, and N3. The improved AASM standard divides the sleep stages into five categories: stages Wake, REM, N1, N3, and N3, corresponding to the five categories in sleep stage classification.

The internationally accepted method for sleep stage classification relies on multimodal time-series physiological signals, known as the PSG signal, which are collected simultaneously using various sensors attached to different parts of the subjects, such as the brain, heart, or legs. However, traditional analysis of physiological signals heavily depends on extracting statistical and spatial features. Although spectrum analysis is interpretable, it requires prior solid knowledge, and its actual classification performance is unsatisfactory. To address this issue, Hassan et al. [20] proposed a tunable- Q wavelet transform to analyze the EEG signal's spectral features, followed by bootstrap aggregating for classification. Their method achieved state-of-the-art EEG-based sleep stage classification performance on the benchmark Sleep-EDF and DREAMS subjects databases. Furthermore, their approach works well and performs equally well for both R&K and AASM sleep scoring standards. Researchers have integrated machine learning techniques to improve the effectiveness of sleep stage classification. For instance, Rahman et al. [37] utilized Discrete Wavelet Transform to extract and analyze the spectral characteristics of the EOG signal and used Random Forest and SVM as the sleep stage classification model. They evaluated their approach on three publicly available databases, including the Sleep-EDF, Sleep-EDFX, and ISRUC-Sleep databases, and demonstrated that it outperforms state-of-the-art EOG-based techniques in accuracy. Similarly, Alickovic et al. [1] proposed a Rotational Support Vector Machine for sleep stage classification. Besides the traditional SVM, they integrated three components: multiscale principal component analysis, discrete wavelet transform, and rotational support vector machine, to enhance the accuracy of EEG signal sleep stage classification. Their approach achieved sensitivity and accuracy values of 84.46% and 91.1%, respectively, across all subjects on the open-source sleep-edfx dataset.

In recent years, numerous researchers have utilized various simple neural networks, including CNN, RNN, and LSTM, for sleep stage classification. Notably, the Time Distributed Multivariate Network, introduced by Chambon et al. [9], has become a standard approach for sleep stage classification problems. This network aggregates the previous d epochs, the subsequent d epochs, and the d th epoch itself to extract features that identify the sleep stage of the d th epoch. Additionally, they employed two convolution kernels of different sizes to extract dual-channel features [40]. Following the proposal of FeatureNet by Jia et al. [22], this model structure is widely employed in the

pre-extraction of features in sleep stage classification. This pre-extraction of features enhances the rapidity of neural network training while guaranteeing accuracy. Goshtasbi et al. proposed Sleep-FCN [18], which involves multi-scale feature extraction and residual diffusion causal convolution. This method yields state-of-the-art classification results in the Sleep-EDF dataset consisting of 20 subjects and a sample of 240 subjects from the SHHS1 dataset. Zhao et al. used the CNN-LSTM-based model called SleepContextNet [57] to extract long-term and short-term temporal context information and developed a data enhancement method. Excellent results were achieved on the Sleep-EDF dataset of 20 subjects and the Sleep-EDFx dataset of 78 subjects, as well as data of 329 subjects selected from the SHHS1 dataset. In addition to using a simple CNN network, more complex network models, such as the U-Net model [32] and its variants, are also employed for sleep stage classification. Phan et al. proposed a sequence-to-sequence method called SeqSleepNet [33] to interpret epoch and sequence levels. Based on this, they developed the first transformer-based sleep stage classification model called SleepTransformer [34] and achieved state-of-the-art performance on the SHHS1 dataset of 5,791 subjects and SleepEDF-78 of 78 subjects. Wang et al. [45] designed a residual attention layer that includes channel attention and spatial attention and achieved state-of-the-art results on the Sleep-EDF dataset and the Sleep-EDFx dataset with 197 PSG records. Zhu et al. proposed MaskSleepNet [59], which consists of a mask module, a multi-scale convolutional network module, a compression and excitation module, and a multi-head attention module. It enables simultaneous learning of both masking and non-masking modality information and performs multi-scale feature extraction and prediction. The proposed model achieved outstanding classification performance on Sleep-EDFx, as well as datasets from the Montreal Archive of Sleep Studies [29] and Huashan Hospital, Fudan University. Nowadays, many sleep stage classification methods rely on GCN due to the similarity between the brain's functional areas and graph structures. These methods make good use of the information about the position and function of the brain. Jia et al. [23] proposed a groundbreaking GCN-based method called GraphSleepNet for sleep stage classification. In this method, each PSG signal channel corresponds to a node in the sleep graph, with a connection between two nodes forming an edge. The features are constructed based on the brain's functional connections, and spatial-temporal graph convolution is used to classify sleep stages. GraphSleepNet is considered a pioneering work in using GCNs for sleep stage classification. Following this, Ji et al. [21] proposed JK-STGCN, a module for aggregating features from different layers. Li et al. [28] developed MVF-SleepNet by adding spectral features from **time-frequency (TF)** images [50] of the PSG time-series signal. Spectral features are extracted with models such as VGG16 and fused with features extracted by the GCN.

2.2 Domain Generalization

Some existing models incorporate transfer learning techniques to enhance cross-subject generalization in sleep stage classification, typically pre-trained on sizable datasets and fine-tuned on smaller ones. Nonetheless, this approach has its limitations, requiring labeled targets for subject data fine-tuning and thereby impeding the model's ability to generalize to previously unseen subjects. A meta-learning-based method called MetaSleepLearner [2] has been proposed, which involves pre-training on the Montreal Archive of Sleep Studies dataset and fine-tuning on new samples from the Sleep-EDF, CAP Sleep Database, ISRUC, and UCD datasets. The outcomes have been encouraging in the realm of sleep stage classification. Phan et al. [35] have also leveraged Kullback-Leibler divergence regularization to facilitate the model's generalization. According to their empirical findings on the Sleep-EDF Expanded database, which contains 75 subjects, their method can boost accuracy by 4.5 percentage points relative to the baseline, resulting in a sleep stage classification accuracy of 79.6%. However, a similar pre-training and fine-tuning paradigm is still necessary, necessitating substantial data for pre-training and labeled data for fine-tuning.

Moreover, this approach can only enhance the model's generalization to specific subjects of the fine-tuning data. Gathering data and carrying out extensive calculations are resource-intensive, making these issues unrealistic in real-world application scenarios where the model cannot access future subject data in advance. To improve the effect of classifying sleep stages exclusive to the subjects, Tang et al. [42] used the MMD method to solve the problem of inconsistency in the distribution of ECG signal data between the training set and the test set and achieved remarkable results on the four datasets including SHHS. Nevertheless, the authors did not investigate methods to enhance model generalization when the test set is unavailable, which is a more pragmatic scenario.

Domain generalization can enhance the model's ability to generalize to unseen subjects without sacrificing accuracy when resources are limited. It improves the model's cross-domain generalization using techniques like Domain-invariant representation learning and Feature disentanglement. Domain adversarial is the most widely used and effective method in domain generalization, which confuses the model's differentiation between domains by introducing a gradient flip layer, thus improving the model's cross-domain robustness. The main advantage of domain generalization is that it improves the cross-domain generalization of the model through the method itself rather than relying on other processes, such as fine-tuning. Moreover, after training, it does not require additional information about the new testing set, making it capable of achieving better results on previously unseen data. Therefore, domain generalization is highly suitable for medical scenarios involving unseen subjects. In the sleep stage classification, Jia et al. [22] proposed a novel framework called MSTCGN, which integrates domain generalization and GCN to extract subject-independent sleep features. Their approach employed the adversarial domain generalization method during training to prevent the model from discerning which source domain the data belonged to, thus enabling it to learn subject-independent information. While their approach achieved state-of-the-art performance at the time, they did not consider the subject dependence difference of the category, and the data from different subjects may be aligned indiscriminately. Therefore, the data from different categories may also be incorrectly aligned. Additionally, their use of cross-validation to only divide the training and testing sets, and the model saving the best result on the testing set during training, could be more rigorous. A better approach is to randomly select a portion of the training set as the validation set, save the best model on the validation set during training, and then test on the unseen testing set data. This ensures complete invisibility of the testing set and generalization verification of the model. It is worth noting that the emergence of different sleep stages is related to the age of the subjects, and it is crucial to take age-related differences into account. To address this issue, Baumert et al. [3] divided their subjects into pediatric, adult, and older adult groups to conduct their research, which is meaningful and provides new insights.

3 PRELIMINARIES AND MOTIVATION

3.1 Sleep Stage Classification Problem

PSG is often employed to record various human body electrical signals during sleep. It contains multi-channel EEG, ECG, EOG, and EMG signals. The PSG signal can be segmented into multi-segment multi-channel signals with 30-second epochs each for sleep stage classification. According to the AASM standard, sleep stages are divided into five stages: Wake, REM, N1, N3, and N3, corresponding to the five categories in sleep stage classification.

The sleep stage classification aims to make the model learn the mapping relationship between the input signal and the sleep stage category. The sleep stage classification problem is defined as $\hat{y}_i = G_y(G_f(x_i))$, building a sleep stage classification model based on the input sample x_i , where G_f is the feature extractor, and G_y is the label classifier. Given the input signal sequence

$S = (S_{i-d}, \dots, S_i, \dots, S_{i+d}) \in \mathbb{R}^{N \times T_n \times T_s}$, where N denotes the number of channels, T_s denotes the time series length of each epoch, $T_n = 2d + 1$ denotes the number of samples of neighbouring $2d + 1$ epochs, \mathcal{S} represents the temporal context of S_i . The classification model will jointly predict the characteristics of the i th epoch according to the transition characteristics of sleep stage rules [9]. Features of each sleep epoch are pre-extracted from the dual-channel FeatureNet [22] and an N -channel feature matrix of the i th epoch is defined as $X_i = (x_i^1, x_i^2, \dots, x_i^N)^T \in \mathbb{R}^{N \times F}$, where $x_i^n \in \mathbb{R}^F, n \in \{1, 2, \dots, N\}$ denotes features pre-extracted from channel n at epoch i . Sometimes, features are preprocessed by bandpass filters according to the frequency distribution of different signals. However, current sleep stage classification methods generally use full unfiltered features.

3.2 Domain Generalization

Suppose we have M subjects (i.e., subjects), we randomly divide M subjects into M' groups, where $M' = \lfloor \frac{M}{num} \rfloor$, num is the number of subjects in each group. Group $m' = \{m'_1, \dots, m'_{num}\}$, where $\{m'_1, \dots, m'_{num}\}$ is random group sampling without replacement from the set $\{1, \dots, M\}$. The data of the M' group constitutes M' domains (i.e., $\mathcal{D}_{m'} = \{(x_{m',k}, y_{m',k})|k \in \{1, \dots, K\}\}$) and K denotes the number of samples of $D_{M'}$, and the joint distributions between each pair of domains are different (i.e., $P_{XY}^{j_1} \neq P_{XY}^{j_2}, 1 \leq j_1 \neq j_2 \leq M'$). Cross-subject classification is the following process, suppose the sample of $\{1, \dots, M' - 1\}$ domains constitutes $\mathcal{D}_{train} = \{\mathcal{D}_1, \dots, \mathcal{D}_{M'-1}\} = \{(x_j, y_j, d_j)|j \in \{1, \dots, J\}\}$. (x_j, y_j, d_j) is the sample composed of $M' - 1$ domains, where x_j denotes the training sample (i.e. the pre-trained feature), y_j denotes the sleep stage label, $d_j \in \{1, \dots, M' - 1\}$ denotes the subject domain label. J is the sum of the numbers of $M' - 1$ domain samples. The sample of the M' th domain constitutes $\mathcal{D}_{test} = \mathcal{D}_{M'} = \{(x_j^{te}, y_j^{te}, d_j^{te})\}$. $(x_j^{te}, y_j^{te}, d_j^{te})$ is the data composed of the M' th domain, where x_j^{te} denotes the sample (i.e., the pre-trained feature), y_j^{te} denotes the sleep stage label, d_j^{te} denotes the subject domain label.

3.3 Motivation

We aim to enhance our model's cross-subject sleep stage classification robustness through domain generalization. Domain generalization eliminates differences between domains (i.e., subjects) through domain alignment. The alignment process aims to align all data of each domain without distinction. However, the biggest challenge in classification tasks is always the category difference, as different sleep stage categories have subject dependency differences. As illustrated in Figure 2, different shapes represent different sleep stage categories, and different colors represent different subject domains. In aligning subject data, if data of the same category are correctly aligned (i.e., the green box in the figure is a positive transfer), then it will enhance the model's cross-subject generalization and improve its classification accuracy. However, if data from the different categories are incorrectly aligned (i.e., the red box in the figure is a negative transfer), then it will severely impact the model classification accuracy. Inspired by the subject dependency difference of categories of sleep stage classification, we hope to align subjects in a fine-grained way by category. Fortunately, the sleep stage classification problem for subject generalization has a category-complete and recurrent structure with information from domain supervision. This structure motivates us to propose the category-specific domain adversarial method SIDA.

4 STRUCTURAL INCENTIVE DOMAIN ADVERSARIAL METHOD

To mitigate the influence of individual subject differences in physiological signals, we introduce the concept of SDDC and propose a SIDA method. The overall architecture of the SIDA method is depicted in Figure 3. In the subsequent sections, we will outline how we employ neural networks to obtain an effective representation of multimodal physiological signals in Section 4.1. We

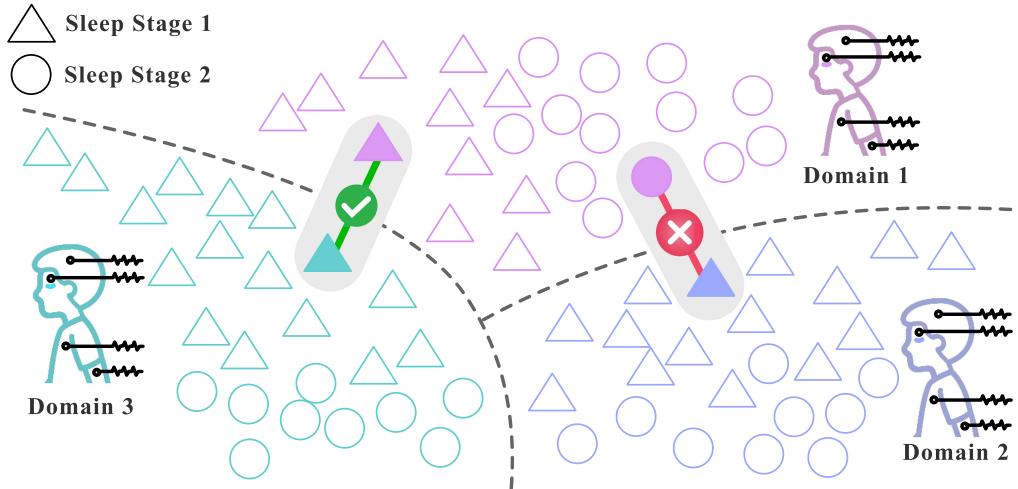


Fig. 2. The SDDC. The varying distributions among subjects are particularly evident for different sleep stage categories. In the figure, different colors represent different subject domains, while different shapes represent different sleep stage categories. The objective of domain generalization is to strengthen model robustness by aligning multiple domains. However, haphazard alignment across subjects can result in different categories' unintended alignment, which will cause considerable inaccuracies in sleep stage classification.

then describe the design of category-special domain discriminators of the sleep stage based on the structure of the sleep stage classification in Section 4.2. Section 4.3 presents how SIDA establishes direct connections between the sleep stage classifier and category-specific domain discriminators to promote the model. Finally, Section 4.4 outlines the overall implementation procedure.

4.1 The Effective Representation of Multimodal Physiological Signal

Due to their non-linear and stationary characteristics and their multimodal heterogeneity [47], effectively representing multimodal physiological signals is a challenging problem. This difficulty arises from three factors: (1) The PSG signal acquisition adheres to medical norms, such as the international standard 10-20 system electrode placement method for the EEG [14, 25] signal, and the EMG signal placement at various muscle sites based on different measurement objectives, resulting in complex spatial structures. (2) The PSG signal is temporal, with temporal dependencies along the timeline, but integrating context information is difficult. (3) There are large differences between multimodal signals across modalities, yet modal consistency exists. Thus, researchers have been grappling with how to fully exploit the consistency between modalities and the difference between compatible modalities. Furthermore, there are still variations between different subjects within the same category or modality. Mainstream methods mostly utilize neural network-based methods to efficiently capture multimodal physiological signals' temporal and spatial features. As Equation (1), CNN-based methods [55] can extract the spatial features of these signals by implementing linear maps through convolution operations with trainable kernels,

$$x_{\beta}^{l+1}(\tau, \mu) = \sigma \left(b_{\beta}^l + \sum_{\gamma=1}^{F^l} U_{\beta\gamma}^l * x_{\beta}^l(\tau, \mu) \right) = \sigma \left(b_{\beta}^l + \sum_{\gamma=1}^{F^l} \left[\sum_{\psi=1, \varphi=1}^{\psi^l \varphi^l} U_{\beta\gamma}^l(\psi, \varphi) * x_{\beta}^l(\tau - \psi, \mu - \varphi) \right] \right), \quad (1)$$

where $x_{\beta}^{l+1}(\tau, \mu)$ denotes the feature map β in layer $(l + 1)$, σ is a non-linear function, F^l is the number of feature maps in layer l , $U_{\beta\gamma}^l$ is the kernel convolved over feature map γ in layer l to create

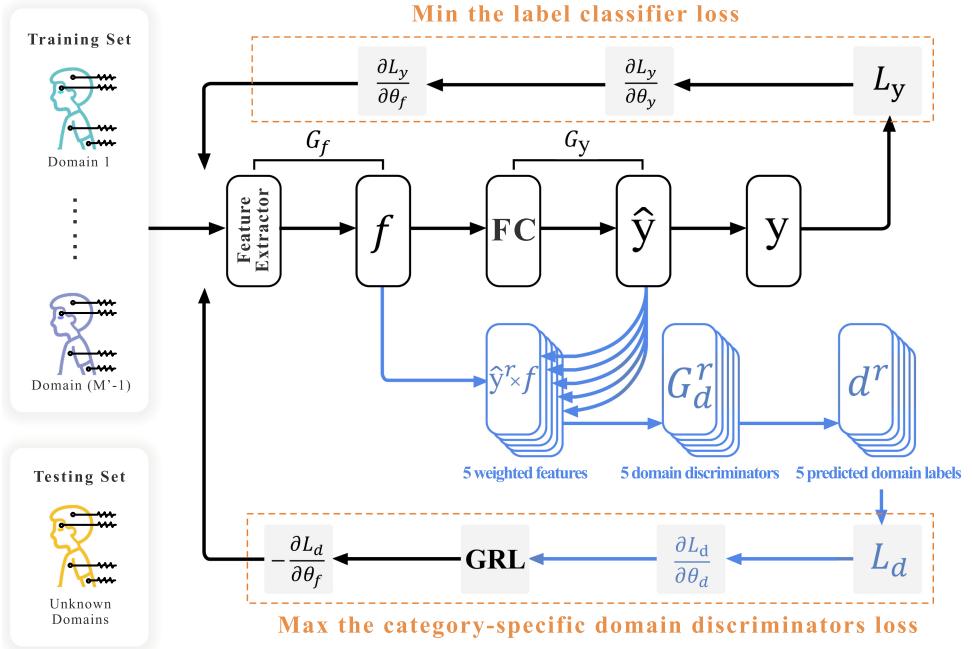


Fig. 3. The overview of the proposed SIDA method. Hidden features from multimodal physiological signals are extracted using CNN, GCN, and so on, and are divided into two streams for sleep stage classification and domain adversarial learning. Category-specific domain discriminators are employed to align subject data in a fine-grained way, promoting the positive transfer and preventing the negative transfer. The direct connection between the sleep stage classifier and domain discriminators facilitates the domain adversarial training process.

the feature map β in layer $(l+1)$, τ, μ are the horizontal and vertical coordinates of the convolution position, respectively, ψ^l, φ^l are the length and width of kernels in layer l , respectively, and b^l is a bias vector.

However, the performance of CNN in temporal extraction is limited. While RNN can extract features by combining sequence context, it lacks information filtering when processing time-series data and is susceptible to gradient disappearance and explosion in long sequences. LSTM addresses these issues through various gating mechanisms, thereby addressing long-term catastrophic forgetting. The updated representation of the LSTM layer is as follows:

$$e_t = \sigma_e (W_{ae}x_t + W_{he}h_{t-1} + W_{ce}c_{t-1} + b_e), \quad (2)$$

$$f_t = \sigma_f (W_{af}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f), \quad (3)$$

$$o_t = \sigma_o (W_{ao}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o), \quad (4)$$

$$c_t = f_t c_{t-1} + i_t \sigma_c (W_{ac}x_t + W_{hc}h_{t-1} + b_c), \quad (5)$$

$$h_t = o_t \sigma_h(c_t), \quad (6)$$

where e, f, o , and c are the input gate, forget gate, output gate, and cell activation vectors. They all define the hidden value and are the same size as vector h . The σ represents the non-linear function.

The x_t is the input to the memory cell layer at time t . W_{ae} , W_{he} , W_{ce} , W_{af} , W_{hf} , W_{cf} , W_{ac} , W_{hc} , W_{ao} , W_{ho} , and W_{co} are weight matrices, b_e , b_f , b_c , and b_o are bias vectors.

Incorporating GCN is a promising approach to constructing graph-based data structures and networks based on modalities and functional locations, allowing for the representation and fusion of multimodal physiological signals. In the graph-based framework, each signal channel is allocated to a node in the sleep graph, while the edges between the nodes represent the connections between signal channels. We have yielded exceptional results in graph-based sleep stage classification, as demonstrated by the MSTGCN [22]. MSTGCN utilizes a multi-view learning strategy that integrates the **function connections (FC)** and the **distance connections (DC)** of sleep graphs and incorporates both temporal and spatial features. We were inspired by Reference [28] to transform the sleep signal into a TF representation using **short-time Fourier transform (STFT)** and to fuse the hidden feature representations extracted by CNN and GCN.

- Spatial Feature: We extract spatial features of sleep graphs using the spatial attention mechanism and Chebyshev graph convolution. As described in Reference [22], the spatial attention is defined as follows:

$$P = V_p \cdot \sigma \left((\mathcal{X}^{l-1} Z_1) Z_2 (Z_3 \mathcal{X}^{l-1})^T + b_p \right), \quad (7)$$

$$P'_{m_1 m_2} = \text{softmax}(P_{m_1, m_2}), \quad (8)$$

where \mathcal{X}^{l-1} is the l th layer's input; V_p , b_p , Z_1 , Z_2 , and Z_3 are learnable parameters; and σ denotes the sigmoid activation function. P denotes the attention matrix, and $P'_{m_1 m_2}$ denotes the correlation between node m_1 and m_2 . The softmax operation is utilized to normalize the attention matrix P . The Chebyshev graph convolution can extract the information of neighboring 0 to $E - 1$ order neighbors centered at each node and is defined as follows:

$$g_\rho * \Omega x = g_\rho(\delta)x = \sum_{\epsilon=0}^{E-1} \rho_\epsilon T_\epsilon(\tilde{\delta})x, \quad (9)$$

$$\tilde{\delta} = \frac{2}{\lambda_{max}} \delta - I_U, \quad (10)$$

where g_ρ denotes the convolution kernel, $*\Omega$ denotes the operation of graph convolution, λ_{max} denotes the Laplacian matrix's maximum eigenvalue, and I_U denotes an identity matrix. T_ϵ denotes the Chebyshev polynomials recursively, and $\delta = D - A$ denotes the Laplacian matrix, where $D \in \mathbb{R}^{U \times U}$ denotes the degree matrix. $\rho \in \mathbb{R}^\epsilon$ denotes a vector of Chebyshev coefficients, and x denotes the input data.

- Temporal Feature: As Section 3.1, according to the primarily identical transition rules of adjacent sleep epochs, we combine temporal context information of neighboring T_n sleep epochs using the temporal attention mechanism and neural network (two-dimensional convolution [22] or a layer of GRU [28]). As described in Reference [22], the temporal attention is defined as follows:

$$Q = V_q \cdot \sigma \left(((\mathcal{X}^{l-1})^T M_1) M_2 (M_3 \mathcal{X}^{l-1}) + b_q \right), \quad (11)$$

$$Q'_{uv} = \text{softmax}(Q_{u,v}), \quad (12)$$

where \mathcal{X}^{l-1} is the l th layer's input; V_q , b_q , M_1 , M_2 , and M_3 are learnable parameters; Q denotes the attention matrix; and $Q'_{u,v}$ denotes the strength of correlation between sleep brain network G_u and G_v . The softmax operation is utilized to normalize the attention matrix. As

shown in Section 3.1, the temporal graph convolution can fuse the temporal context information of adjacent T_n sleep epochs and is defined as follows:

$$\mathcal{X}^l = \text{ReLU}(\phi * (\text{ReLU}(g_\rho * G\hat{\mathcal{X}}^{l-1}))), \quad (13)$$

where $\hat{\mathcal{X}}^{l-1}$ is the l th layer's input with temporal attention, g_ρ denotes the convolution kernel, ReLU is the non-linear activation function, ϕ denotes the parameters of the convolution kernel, and $*$ denotes the convolution operation.

- Spectral Feature: Transforming time-series data into TF image data using techniques such as STFT can effectively facilitate the model in capturing frequency-related information, allowing it to fully utilize the strengths of CNN in image classification and recognition. In a recent study [28], the ResNet and VGG models were utilized to extract features from TF images, which were then combined with GRU to integrate the temporal features of multiple sleep epochs. The resulting features were further fused with the features extracted by GCN, leading to notable improvements in performance.
- Multi-view Feature Fusion: In the MSTGCN-based [22] methods, we concatenate the graph features based on the FC and the graph features based on the DC. Each feature consists of spatial features and temporal features.

In other methods based on graph features, we only employ the graph features based on FC. Spatial-temporal features are utilized in most methods. In particular, in MVF-SleepNet [28], not only spatial-temporal features are included but also spectral-temporal features are included.

4.2 Structural Incentive Domain Adversarial Learning

To improve the cross-subject generalization of the model while ensuring classification accuracy, traditional domain generalization methods often extract subject-invariant information to ensure that the model learns a general and robust representation. We exploit an adversarial domain generalization method to enhance the generalization of various sleep stage classification models. Suppose the input signal is x_j , the feature extractor is G_f , the label classifier is G_y , and the domain classifier is G_d . Specifically, **Gradient Reversal Layer (GRL)** is implemented between the G_f and the G_d to form an adversarial relationship. During training, the model parameters of the G_f are jointly affected by the G_d and the G_y . The purpose of the G_d is to confuse the model's identification of subjects, thereby enhancing the cross-subject generalization of the model. Unlike the traditional transfer learning framework in which the pre-training and fine-tuning are separated, the domain adversarial method integrates the classification and domain generalization into a unified and end-to-end framework. Due to the existence of GRL, the model parameters θ_f of the G_f are learned by minimizing the loss \mathcal{L}_y of the label classifier and maximizing the loss \mathcal{L}_d of category-specific domain discriminators. Without loss of generality, the multi-class cross-entropy \mathcal{L}_{mc} is exploited as the basic loss function,

$$\mathcal{L}_y = \frac{1}{J} \sum_{j=1}^J \mathcal{L}_{mc}(G_y(G_f(x_j)), y_j), \quad (14)$$

$$\mathcal{L}_d = \frac{1}{J} \sum_{j=1}^J \mathcal{L}_{mc}(G_d(G_f(x_j)), d_j), \quad (15)$$

where J denotes the number of training samples and y_j and d_j denote the true label of sleep stage and subject domain, respectively. The network is optimized by minimizing the sum of two losses, and the total loss of domain generalization is defined as

$$\mathcal{L} = \mathcal{L}_y - \mathcal{L}_d. \quad (16)$$

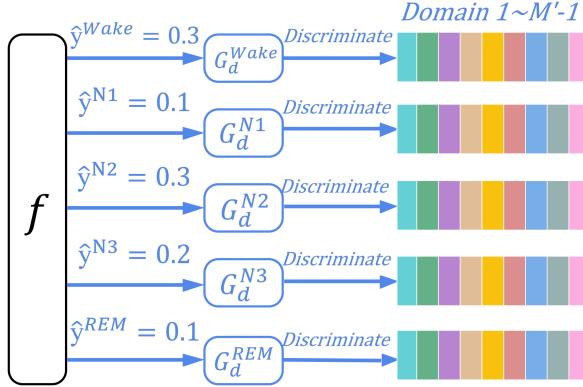


Fig. 4. This figure illustrates the Direct Dynamic Bridge between the label classifier, G_y , and the category-specific domain discriminator, G_d^r . It shows how G_y 's softmax prediction for a particular category (e.g., Wake) is used to dynamically weight the features of G_d^r corresponding to that category. Specifically, if G_y predicts a value of 0.3 for Wake, then the feature f of G_d^r corresponding to Wake will be weighted by 0.3. Unlike traditional domain adversarial methods, which lack direct interaction between G_y and G_d , our approach establishes an interpretable and dynamic connection between the two networks at the category level, resulting in improved performance.

Existing domain adversarial methods in sleep stage classification improve the ability of the model to generalize to different subjects. However, there are considerable differences in the subject dependence on different sleep stage categories. As Section 3.3 shows, aligning the comprehensive data of different subjects will introduce substantial interference to the model's judgment of category. Fortunately, the sleep stage classification has a periodic category-complete structure. As Figure 3 shows, subject to the incentive structure of which we set category-specific domain discriminators and perform category-by-category fine-grained alignment on different subjects' data. This way, the domain adversarial process will not introduce additional errors to the sleep stage classification process. In addition, using the prediction results of the label classifier to weight category-specific domain discriminators dynamically also realizes an adaptive and direct correlation between them. The structurally incentive label classifier loss and total loss are consistent with Equations (14) and (16), respectively. The structural incentive category-specific domain discriminators loss is defined as

$$\mathcal{L}_d = \frac{1}{J} \sum_{r=1}^R \sum_{j=1}^J \alpha_r \mathcal{L}_d^r (G_d^r (\hat{y}_j^r G_f (x_j)), d_j), \quad (17)$$

where J and R denote the number of training samples and categories, respectively; d_j denotes the domain label of the sample x_j ; r denotes r th domain discriminator; \hat{y}_j^r denotes the predicted softmax value of the r th category by label classifier; and $\alpha_r = \frac{1}{R}$ are set as the weight of the loss of domain discriminator of the category r .

4.3 The Advantage of Direct Dynamic Bridge

In traditional domain adversarial methods, the relationship between the label classifier G_y and the domain discriminator G_d is coordinated through indirect gradient backpropagation. The nature of their relationship cannot be easily explained. Therefore, as illustrated in Figure 4, we employ the current prediction softmax value of the G_y as the weight for the features of the category-specific domain discriminator G_d^r , which aims to address the issue of subject dependency differences

category-by-category. This approach provides two benefits: (1) the fine-grained alignment at the category level avoids problems associated with the SDDC problem and (2) the original indirect association between the G_y and the G_d has been transformed into a dynamic weighting association. As a result, a soft attention weight is created that allows for interpretability between the two, which is entirely dependent on the end-to-end learning of the model. The model can adjust the proportion of features of each G_d^r adaptively based on the current prediction situation. Furthermore, the soft weight is smooth and differentiable, which ensures that each G_d^r is taken care of without becoming too absolute. Even when the G_y is inaccurate, they will still promote and adjust each other.

4.4 Method Implementation

SIDA proposed in this article is fully elucidated in Algorithm 1, SIDA aimed at achieving generalizable sleep stage classification. The framework is characterized by the loss \mathcal{L}_y of label classifier and the loss \mathcal{L}_d of category-specific domain discriminators. Initially, the features pre-trained with FeatureNet are input, which encompasses the training set \mathcal{D}_{train} and the testing set $\{x_j^{te}\}$. Subsequently, the model's parameters θ_f , θ_y , and θ_d are initialized. Features of x_j are extracted with feature extractor G_f , and then the classification result is obtained, following which the classification loss \mathcal{L}_y is computed by the fully connected classifier G_y . Based on the softmax value of the current prediction result, each category's prediction result is derived, features used by each category-specific domain discriminator are weighted, and the weighted loss sum \mathcal{L}_d of category-specific domain discriminators is calculated. The loss \mathcal{L}_d after gradient inversion is added to the loss \mathcal{L}_y to acquire the total loss \mathcal{L} . Gradient backpropagation and updates of all parameters θ_f , θ_y , and θ_d of the model continue until convergence. Finally, the best model for sleep classification is obtained.

ALGORITHM 1: The whole process of SIDA method

Input: The pre-trained feature $\mathcal{D}_{train} = \{(x_j, y_j, d_j) | j \in \{1, \dots, J\}\}$ for training, the pre-trained feature $\{x_j^{te} | j \in \{1, \dots, J'\}\}$ for testing

Output: The prediction result of the test data \tilde{y}_j^{te}

- 1 Initialize the parameter θ_f , θ_y , and θ_d of G_f , G_y , and G_d , respectively;
- 2 **repeat**
- 3 Extract the feature of x_j with G_f : $f_j = G_f(x_j)$;
- 4 Calculate the loss \mathcal{L}_y of the label classifier with Equation (14);
- 5 $\hat{y}_j = softmax(f_j)$, $\hat{y}_j = (\hat{y}_{j,1}, \dots, \hat{y}_{j,R})$;
- 6 Calculate the loss \mathcal{L}_d of category-specific domain discriminators with Equation (17);
- 7 Calculate the total loss $\mathcal{L} = \mathcal{L}_y - \mathcal{L}_d$;
- 8 Update the θ_f , θ_y , θ_d by minimizing \mathcal{L} ;
- 9 **until** *Iterate until model convergence*;
- 10 Predict the sleep stage classification results: $\tilde{y}_j^{te} = G_y(G_f(x_j^{te}))$;

5 EXPERIMENTS

All these experiments are implemented with Python 3.8.0, Nvidia-TensorFlow 1.15.0, and Keras 2.3.1. We conducted them on a computer server equipped with 960GB Memory, Ubuntu 20.04.1 operating system, and four Nvidia A100 GPUs with 80 GB GPU Memory each.

5.1 Dataset and Experiment Settings

5.1.1 Dataset. Our experiments employ three datasets: two publicly available subsets of the ISRUC-Sleep database, ISRUC-S1 and ISRUC-S3, and one large-scale dataset, **SHHS1**. The general

Table 1. Data Description

Dataset	Number of subject	Number of sleep stage					
		Wake	N1	N2	N3	REM	Total
ISRUUC-S1	100	20,098	11,062	27,511	17,251	11,265	87,187
ISRUUC-S1 with context	100	19,860	11,025	27,448	17,232	11,222	86,787
ISRUUC-S3	10	1,674	1,217	2,616	2,016	1,066	8,589
ISRUUC-S3 with context	10	1,651	1,215	2,609	2,014	1,060	8,549
SHHS1	329	46,319	10,304	142,125	60,153	65,953	324,854
SHHS1 with context	329	45,312	10,278	141,936	60,128	65,884	323,538

information is shown in Table 1. The PSG recording was segmented into 30-second-long epochs and annotated by two experts according to the AASM standards. In detail, (1) the common points of the first two datasets are as follows: Each recording contains six EEG channels (F3-A2, C3-A2, O1-A2, F4-A1, C4-A1, and O2-A1), two EOG channels (LOC-A2 and ROC-A1), three EMG channels (the chin EMG, left leg movements, and right leg movements), and one ECG channel. Considering the sleep task has little correlation with the EMG signal of the legs. We are consistent with the comparative method, such as MSTGCN, removing the EMG channels of the two legs for experiments to focus on brain signals and employing the data of 10 channels in total. In addition, signals were resampled at 100 Hz. (2) ISRUUC-S3 subgroup contains 10 healthy adults (nine male and one female, aged from 30 to 58). (3) The ISRUUC-S1 subgroup contains 100 adults with sleep disorders (55 males and 45 females, aged from 20 to 85). (4) Following References [18, 57], 329 subjects with regular sleep of SHHS1 dataset are selected according to the Apnea Hypopnea Index. The six channels (two EEG, two EOG, one ECG, and one EMG) are employed in our experiment. In addition, signals were sampled at 125 Hz. As is shown in Section 3.1, the model will jointly predict the features of the intermediate epoch according to the T_n epochs, which will better contain the context information. As is shown in Table 1, the pre-trained features of ISRUUC-S3 with context are from 10 subjects, and the pre-trained features of each subject discard a total of four epochs, so ISRUUC-S3 with context is 40 epochs less than the original features. Similarly, ISRUUC-S1 with context with 100 subjects is 400 epochs less than the original features, and SHHS1 with context with 329 subjects is 1,316 epochs less than the original features.

5.1.2 Parameter Settings. We compare SIDA with several baselines and against experiments with only integrating the traditional domain adversarial method, described in Tables 3, 4, and 5. We employ the same experimental settings for all models for a fair comparison. We reproduce each comparative method and employ 10-fold cross-validation to divide the training and testing set. In detail, the ratio of the training set to the testing set is 9:1. Then, randomly selecting 20% from the training set as the validation set, we save the best model validated on the validation set and test it on the completely invisible new subject testing set. The comparative method in their papers did not use the validation set, so our experimental results seem lower than in the comparative papers. All networks and parameters are consistent with each original method. Detailed hyper-parameters are shown in Table 2, where the parameter neighbouring epoch size means the number of neighbouring temporal epochs to aggregate (i.e., T_n), the parameter Order of Chebyshev polynomials ϵ is set to five in GraphSleepNet and MSTGCN and is set to nine among other methods to remain consistent with the original method.

5.1.3 Sleep Stage Classification Methods. Features are extracted and fused based on the dual-channel FeatureNet [22]. FeatureNet is an effective baseline method for sleep stage classification and has been commonly employed as the pre-extraction of sleep stage classification features.

Table 2. The Shared Hyper-parameters in All Methods

Hyper-parameter	Value
Neighbouring sleep epoch size (context)	5
Number of training epoch	80
Training batch size	64
Ratio of gradient reversal	0.001
Ratio of r th domain discriminator loss α_r	0.2
Optimizer	Adam
Learning rate	0.0001
Dropout ratio	0.5

Table 3. The Performance Comparison of the Mainstream Method with/without Our SIDA or Traditional Domain Adversarial Method on the ISRUC-S1 Dataset

Method	Overall results				F1-score for each class			
	Accuracy	Macro F1	Kappa	Wake	N1	N2	N3	REM
FeatureNet	0.7767	0.7592	0.7130	0.8713	0.5245	0.7588	0.8391	0.8025
MaskSleepNet	0.7507	0.7426	0.6807	0.8594	0.4974	0.7279	0.8301	0.7981
SleepContextNet	0.7422	0.7234	0.6682	0.8169	0.5035	0.7480	0.8256	0.7229
DAN	0.7436	0.7233	0.6688	0.8469	0.4553	0.7337	0.8117	0.7689
GraphSleepNet	0.7879	0.7689	0.7260	0.8842	0.5265	0.7714	0.8513	0.8112
MSTGCN	0.7900	0.7732	0.7289	0.8802	0.5286	0.7761	0.8507	<u>0.8303</u>
MSTGCN+SIDA	0.8004	0.7792	0.7411	0.8899	0.5348	0.7864	0.8563	0.8285
JK-STGCN	0.7887	0.7694	0.7271	0.8824	0.5332	0.7719	0.8519	0.8074
JK-STGCN+DA	0.7926	0.7725	0.7321	0.8832	0.5318	0.7797	<u>0.8583</u>	0.8093
JK-STGCN+SIDA	0.7963	0.7741	0.7360	0.8891	0.5222	0.7808	0.8602	0.8184
MVF-SleepNet	0.7903	0.7741	0.7296	0.8855	0.5390	0.7754	0.8470	0.8237
MVF-SleepNet+DA	0.7910	0.7747	0.7306	0.8759	0.5389	0.7788	0.8548	0.8251
MVF-SleepNet+SIDA	0.7941	<u>0.7774</u>	0.7343	0.8784	0.5391	<u>0.7823</u>	0.8566	0.8308

FeatureNet aims to extract neural network features from the raw input feature matrix, which means that original data points from each channel at each epoch will be transferred into pre-extracted feature vectors. FeatureNet, in our experimental results Tables 3, 4, and 5, is utilized to classify the pre-extracted features through the fully connected classifier directly. All comparative methods employ the features pre-extracted by FeatureNet as the original feature input. MaskSleepNet [59] consists of a masking module, a multi-scale convolutional neural network, a squeezing and excitation block, and a multi-headed attention module. SleepContextNet [57] extracts the signal's long-term and short-term context information based on the CNN-RNN backbone and designs a data augmentation method. DAN [42] is based on a CNN-GRU backbone, using the MMD [19] to align different domains. GraphSleepNet [23] is a GCN-based method. In GraphSleepNet, each PSG signals channel corresponds to a node in the sleep graph, and a connection between two nodes forms an edge in the sleep graph. This method constructed graph features based on the brain's FC and employs spatial-temporal graph convolution to classify sleep stages. Based on GraphSleepNet, MSTGCN [22] adds DC based on the electrode distance of functional brain areas, and domain generalization is incorporated into the model to extract subject-independent information. JK-STGCN [21] is proposed to aggregate features from different layers by a jumping knowledge module. In MVF-SleepNet [28], spectral features from TF images of the PSG time-series signal were added and were utilized by models such as VGG16. We also compare the results of the traditional domain adversarial method and our SIDA combined with each method. The single discriminator

Table 4. Performance Comparison of the Mainstream Method with/without Our SIDA or Traditional Domain Adversarial Method on the ISRUC-S3 Dataset

Method	Overall results				F1-score for each class			
	Accuracy	Macro F1	Kappa	Wake	N1	N2	N3	REM
FeatureNet	0.7538	0.7456	0.6855	0.8742	0.5637	0.6885	0.8274	0.7740
MaskSleepNet	0.7427	0.7256	0.6690	0.8576	0.4819	0.7209	0.8404	0.7270
SleepContextNet	0.7709	0.7618	0.7040	0.8717	0.5464	0.7477	0.8397	0.8032
DAN	0.7350	0.7133	0.6574	0.8311	0.4496	0.7277	0.8300	0.7278
GraphSleepNet	0.7763	0.7654	0.7116	0.8674	0.5409	0.7506	0.8482	0.8201
MSTGCN	0.7830	0.7725	0.7202	0.8741	0.5585	0.7587	0.8612	0.8099
MSTGCN+SIDA	0.7972	0.7802	0.7383	0.8801	0.5264	0.7832	0.8703	0.8408
JK-STGCN	0.7870	0.7762	0.7254	0.8770	0.5601	0.7652	0.8580	0.8208
JK-STGCN+DA	0.7913	0.7805	0.7309	0.8784	0.5603	0.7713	0.8562	<u>0.8364</u>
JK-STGCN+SIDA	<u>0.7952</u>	0.7798	<u>0.7351</u>	0.8786	0.5477	0.7821	<u>0.8658</u>	0.8249
MVF-SleepNet	0.7899	<u>0.7827</u>	0.7292	<u>0.8931</u>	<u>0.5842</u>	0.7720	0.8312	0.8330
MVF-SleepNet+DA	0.7917	0.7824	0.7316	<u>0.8931</u>	0.5770	0.7782	0.8396	0.8240
MVF-SleepNet+SIDA	0.7972	0.7882	0.7383	0.8959	0.5852	0.7850	0.8439	0.8307

The **bold** and underline items denote the best and second-best results, respectively.

Table 5. The Performance Comparison of the Mainstream Method with/without Our SIDA or Traditional Domain Adversarial Method on the SHHS1 Dataset

Method	Overall results				F1-score for each class			
	Accuracy	Macro F1	Kappa	Wake	N1	N2	N3	REM
FeatureNet	0.7959	0.6778	0.7103	0.8158	0.1561	0.8230	0.8388	0.7556
MaskSleepNet	0.7689	0.7083	0.6880	0.8412	0.2504	0.8296	0.8608	0.7596
SleepContextNet	0.8237	0.7404	0.7541	0.8220	0.3597	0.8388	0.8375	0.8442
DAN	0.8470	0.7382	0.7821	0.8820	0.2462	0.8636	0.8645	0.8348
GraphSleepNet	0.8664	0.7899	0.8117	0.8893	0.4295	0.8800	0.8695	0.8814
MSTGCN	<u>0.8763</u>	<u>0.8009</u>	0.8254	0.8987	0.4454	0.8873	0.8762	0.8968
MSTGCN+SIDA	<u>0.8802</u>	0.7660	<u>0.8298</u>	0.9019	0.2546	0.8912	<u>0.8788</u>	<u>0.9034</u>
JK-STGCN	0.8777	0.7899	0.8266	0.8935	0.3911	0.8910	0.8750	0.8987
JK-STGCN+DA	0.8782	0.7929	0.8274	0.8986	0.4009	0.8906	0.8770	0.8976
JK-STGCN+SIDA	0.8843	0.8048	0.8366	0.8995	<u>0.4387</u>	0.8958	0.8807	0.9092
MVF-SleepNet	0.8681	0.7654	0.8126	0.8882	0.3013	0.8796	0.8632	0.8947
MVF-SleepNet+DA	0.8694	0.7801	0.8149	0.8915	0.3692	0.8812	0.8645	0.8942
MVF-SleepNet+SIDA	0.8736	0.7975	0.8211	0.9101	0.4340	0.8839	0.8684	0.8912

The **bold** and underline items denote the best and second-best results, respectively.

structure is consistent with the label classifier of each original method. In summary, our experimental methods and comparative methods are as follows:

- FeatureNet [22]: The FeatureNet model is utilized for feature extraction and sleep stage classification. It is the only method in this article that does not employ the data with context, as is shown in Table 1. The results reported in the experiment are the results of all training samples, and other methods employ pre-extracted features with context of training samples.
- MaskSleepNet [59]: The FeatureNet model is utilized for features pre-extraction, and the MaskSleepNet model is exploited to study neural network-based features and classify sleep stages.

- SleepContextNet [57]: The FeatureNet model is utilized for features pre-extraction, and the SleepContextNet model is exploited to study neural network-based features and classify sleep stages.
- DAN [42]: The FeatureNet model is utilized for features pre-extraction, and the DAN model is exploited to study neural network-based features and classify sleep stages.
- GraphSleepNet [23]: The FeatureNet model is utilized for features pre-extraction, and the GraphSleepNet model is exploited to study graph-based features and classify sleep stages.
- MSTGCN [22]: The FeatureNet model is utilized for features pre-extraction, and the MST-GCN model is exploited to study graph-based features and classify sleep stages. The MSTGCN model combines the traditional domain adversarial method to extract subject-independent information.
- MSTGCN+SIDA: The FeatureNet model is utilized for features pre-extraction, and the MST-GCN model is exploited to study graph-based features and classify sleep stages. The MST-GCN+SIDA method is integrated with our SIDA to extract subject-independent information.
- JK-STGCN [21]: The FeatureNet model is utilized for features pre-extraction, and the JK-STGCN method is exploited to study graph-based features and classify sleep stages.
- JK-STGCN+DA: The FeatureNet model is utilized for features pre-extraction, and the JK-STGCN model is exploited to study graph-based features and classify sleep stages. The JK-STGCN+DA method is integrated with the traditional domain adversarial method to extract subject-independent information.
- JK-STGCN+SIDA: The FeatureNet model is utilized for features pre-extraction, and the JK-STGCN model is exploited to study graph-based features and classify sleep stages. The JK-STGCN+SIDA method is integrated with our SIDA method to extract subject-independent information.
- MVF-SleepNet [28]: The FeatureNet model is utilized for features pre-extraction, and the MVF-SleepNet model is exploited to study graph-based features and classify sleep stages.
- MVF-SleepNet+DA: The FeatureNet model is utilized for features pre-extraction, and the MVF-SleepNet model is exploited to study graph-based features and classify sleep stages. The MVF-SleepNet+DA method is integrated with the traditional domain adversarial method to extract subject-independent information.
- MVF-SleepNet+SIDA: The FeatureNet model is utilized for features pre-extraction, and the MVF-SleepNet method is exploited to study graph-based features and classify sleep stages. The MVF-SleepNet+SIDA method is integrated with our SIDA method to extract subject-independent information.

Besides FeatureNet, the same pre-extracted features extracted by the FeatureNet are employed in all methods. The word DA denotes the traditional domain adversarial method. The word SIDA means adding our SIDA method to the origin method. All DA-related and SIDA-related parameters are kept the same in the original methods. Compared with GraphSleepNet, MSTGCN mainly adds a subject discriminator for domain adversarial operation. Moreover, according to the distance between different positions of the electrodes in the brain area, MSTGCN incorporates distance connections to enrich the spatial proximity structural features of the brain. The backbone of the two methods is basically the same, so we mainly compare our SIDA method based on the upgraded version of GraphSleepNet, MSTGCN.

5.1.4 Performance Metrics. The evaluation measures, including *Accuracy* (*Acc*), *F1 score* (*F1*), *Macro F1*, and *Kappa* are defined as follows:

$$Acc = \frac{TP + TN}{TP + FP + FN + TN}, \quad (18)$$

Table 6. The Across-age Performance Comparison of the GraphSleepNet method with/without Our SIDA or Traditional Domain Adversarial Method on the ISRUC-S1 Dataset

Method	Accuracy			
	Pediatric	Adult	Older adult	Total
GraphSleepNet	<u>0.8173</u>	0.7776	0.7335	0.7761
MSTGCN	0.8143	<u>0.7870</u>	<u>0.7378</u>	<u>0.7795</u>
MSTGCN+SIDA	0.8225	0.7956	0.7386	0.7853

The **bold** and underline items denote the best and second-best results, respectively.

$$F1 = \frac{2 \times (Precision \times Recall)}{Precision + Recall}, \quad (19)$$

$$Macro\ F1 = \frac{\sum_r F1_r}{R}, \quad (20)$$

$$Kappa = \frac{p_o - p_e}{1 - p_e}, \quad (21)$$

where TP refers to the number of samples of the current sleep stage classified correctly, FP refers to the number of samples of other sleep stages classified wrongly to be the current sleep stage, FN refers to the number of samples of the current sleep stage classified wrongly to be other stages, and TN refers to the number of samples of other sleep stages classified correctly. $F1$ score is the harmonic mean of $Recall$ and $Precision$, r is the sleep stage category, R is the number of sleep stage categories, and $Macro\ F1$ is for each sleep stage category of the $F1$ score calculation arithmetic mean. Moreover, p_o is the relative observation consistency between raters, and p_e is the hypothetical probability of probability consistency.

5.2 Comparative Experiment Results

Our SIDA method provides a fine-grained distribution alignment to reduce subject-dependence variability in sleep stage classification compared to traditional domain adversarial techniques. We evaluate the classification performance using FeatureNet on raw data from ISRUC-S1, ISRUC-S3, and SHHS1 datasets (Table 1) and compare other methods on pre-extracted features with context. The experimental results presented in Table 3, 4, and 5 show that we achieved further improvement on three datasets by combining SIDA with the original methods, which also shows superiority compared to several state-of-the-art methods. Results on the ISRUC-S3 dataset demonstrate that MSTGCN+SIDA and MVF-SleepNet+SIDA attained the highest Acc result of 0.7972. Notably, the Acc result of MSTGCN+SIDA increased by over one percentage point compared to the MSTGCN method and about two percentage points compared to GraphSleepNet. The $Macro\ F1$ result of MVF-SleepNet+SIDA also reached a high of 0.7882, indicating a significant improvement. Compared to the best-performing method SleepContextNet in the comparison method, the Acc , $Macro\ F1$, and $Kappa$ of MVF-SleepNet+SIDA increased by approximately three percentage points each. On the ISRUC-S1 dataset, MSTGCN+SIDA achieved the highest Acc result of 0.8004 due to the larger number of samples. Notably, the $Macro\ F1$ and $Kappa$ results of MSTGCN+SIDA also reached a high value of 0.7792 and 0.7411 and increased by about one percentage point each compared to the MSTGCN method. In comparison to the best-performing method MaskSleepNet, the Acc , $Macro\ F1$, and $Kappa$ of MSTGCN+SIDA increased by approximately five, four, and six

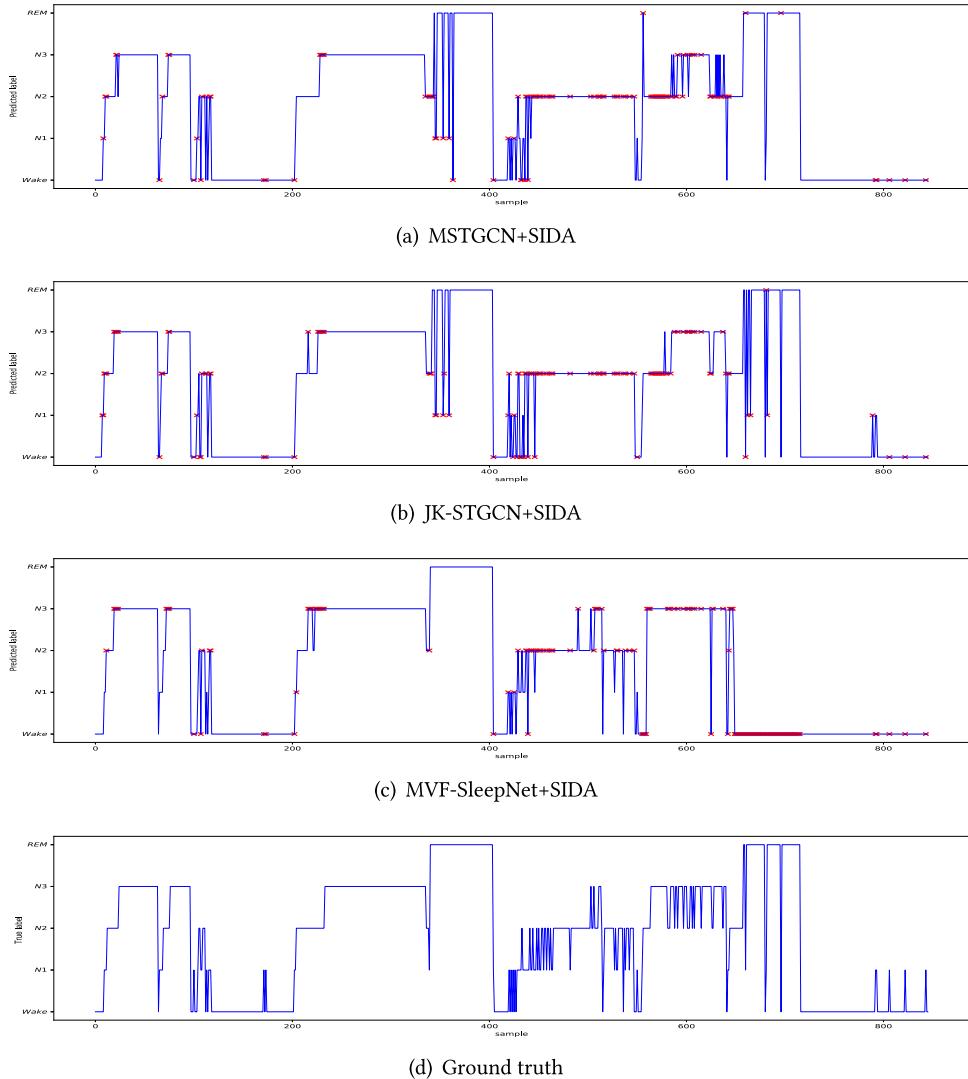


Fig. 5. The comparative between ground truth and predicted sleep stage of the subject one on the ISRUC-S1 dataset with (a) MSTGCN+SIDA, (b) JK-STGCN+SIDA, (c) MVF-SleepNet+SIDA methods, and (d) Ground truth.

percentage points, respectively, which are significant improvements. On the SHHS1 dataset, JK-STGCN+SIDA attained the highest *Acc* of 0.8843, *Macro F1* of 0.8048, and *Kappa* of 0.8366. Compared to the JK-STGCN method without SIDA, the results increased by approximately one percentage point. The *Acc*, *Macro F1*, and *Kappa* of JK-STGCN+SIDA also increased by about four, seven, and five percentage points, respectively, compared to the best-performing method DAN. Our experiments visualized the changes in sleep stage categories throughout an entire sleep episode when fusing SIDA with/without other methods. The classification performance is excellent, with most misclassifications occurring during sleep stage transitions, which are typically difficult to detect in medical practice, as demonstrated in Figures 5, 6, and 7 for the ISRUC-S1, ISRUC-S3, and SHHS1 datasets, respectively.

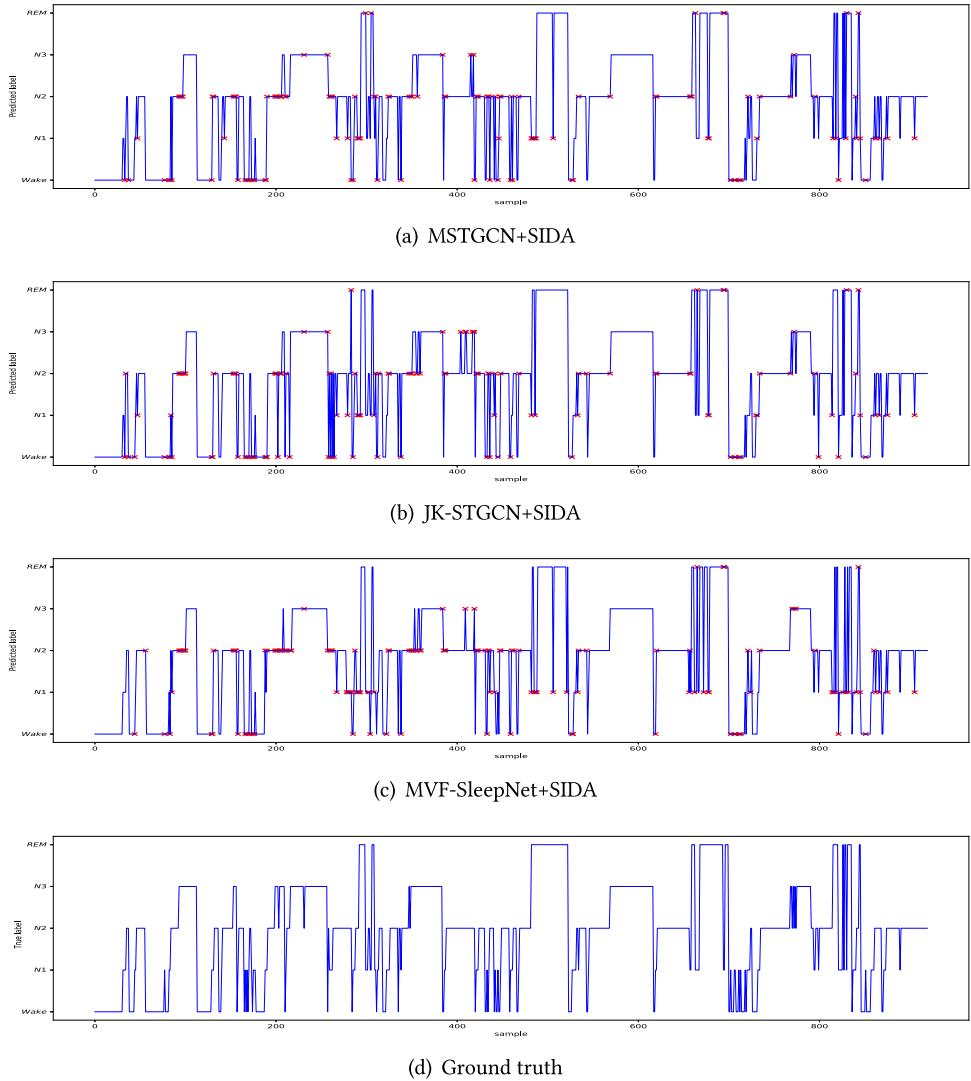


Fig. 6. The comparative between ground truth and predicted sleep stage of the subject one on the ISRUC-S3 dataset with (a) MSTGCN+SIDA, (b) JK-STGCN+SIDA, (c) MVF-SleepNet+SIDA methods, and (d) Ground truth.

5.3 Across-age Experiment Results

Following Reference [3], we divided the ISRUC-S1 dataset into three age groups: 34 pediatrics, 32 adults, and 34 older adults, treating each group as a domain. Our experiments utilized GraphSleepNet, MSTGCN, and MSTGCN+SIDA methods to evaluate performance in different age groups. Table 6 displays that our MSTGCN+SIDA method outperforms both MSTGCN and GraphSleepNet methods. Notably, we discovered a noteworthy phenomenon: the model consistently outperforms older groups in identifying younger groups, which warrants further exploration in future research.

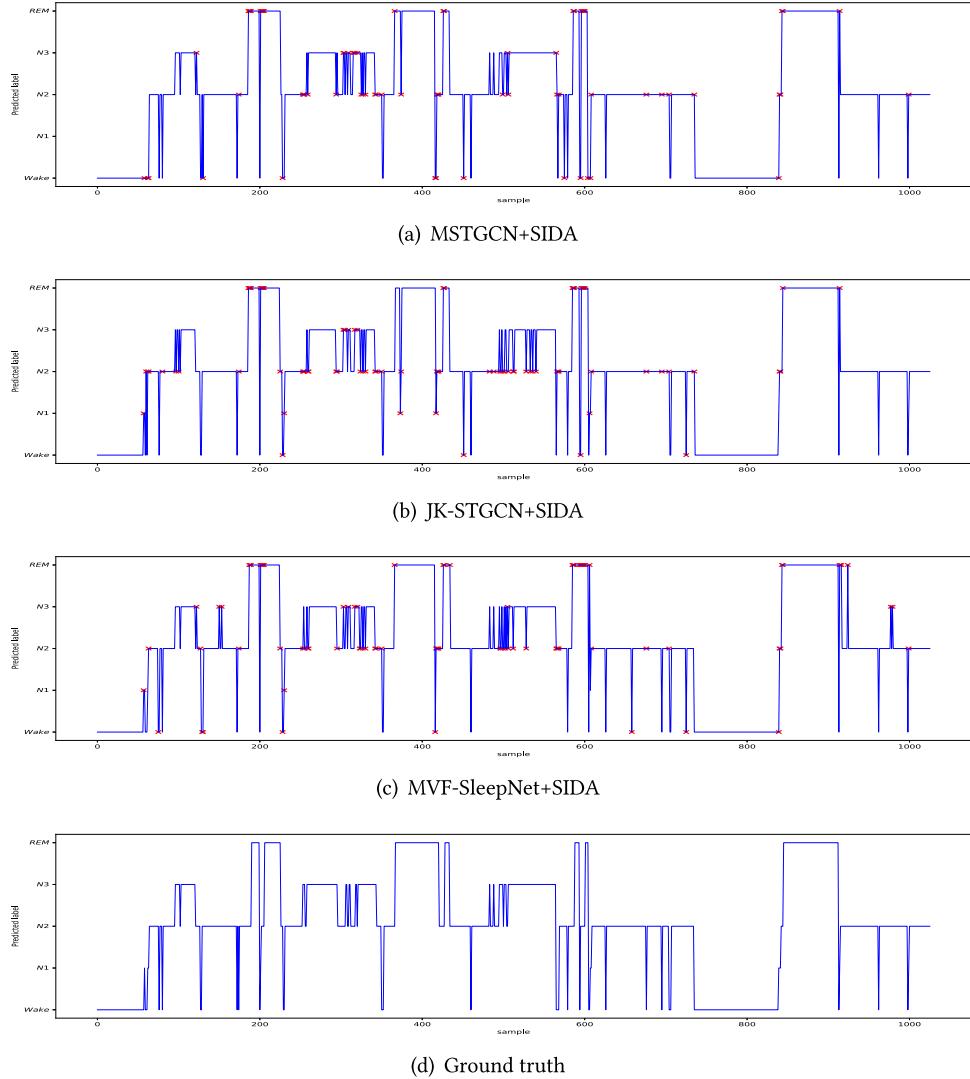


Fig. 7. The comparative between ground truth and predicted sleep stage of the subject one on the SHHS1 dataset with (a) MSTGCN+SIDA, (b) JK-STGCN+SIDA, (c) MVF-SleepNet+SIDA methods, and (d) Ground truth.

5.4 Feature Visualization Analysis

To investigate the impact of the traditional domain adversarial method and our SIDA, we performed feature visualization for some methods by selecting hidden features before the fully connected classifier and utilizing the tSNE tool to reduce their dimensions to a two-dimensional plane. As illustrated in Figure 8 (d), (e) and (f) for the ISRUC-S3 dataset, the classification boundary of the GraphSleepNet method is ambiguous, with the features of the same subjects being primarily concentrated together, showing significant differences in subject personalization. Moreover, the category boundary of sleep stage N1 is the most unclear and challenging to identify, consistent with our experimental findings. Nevertheless, after using the traditional domain adversarial

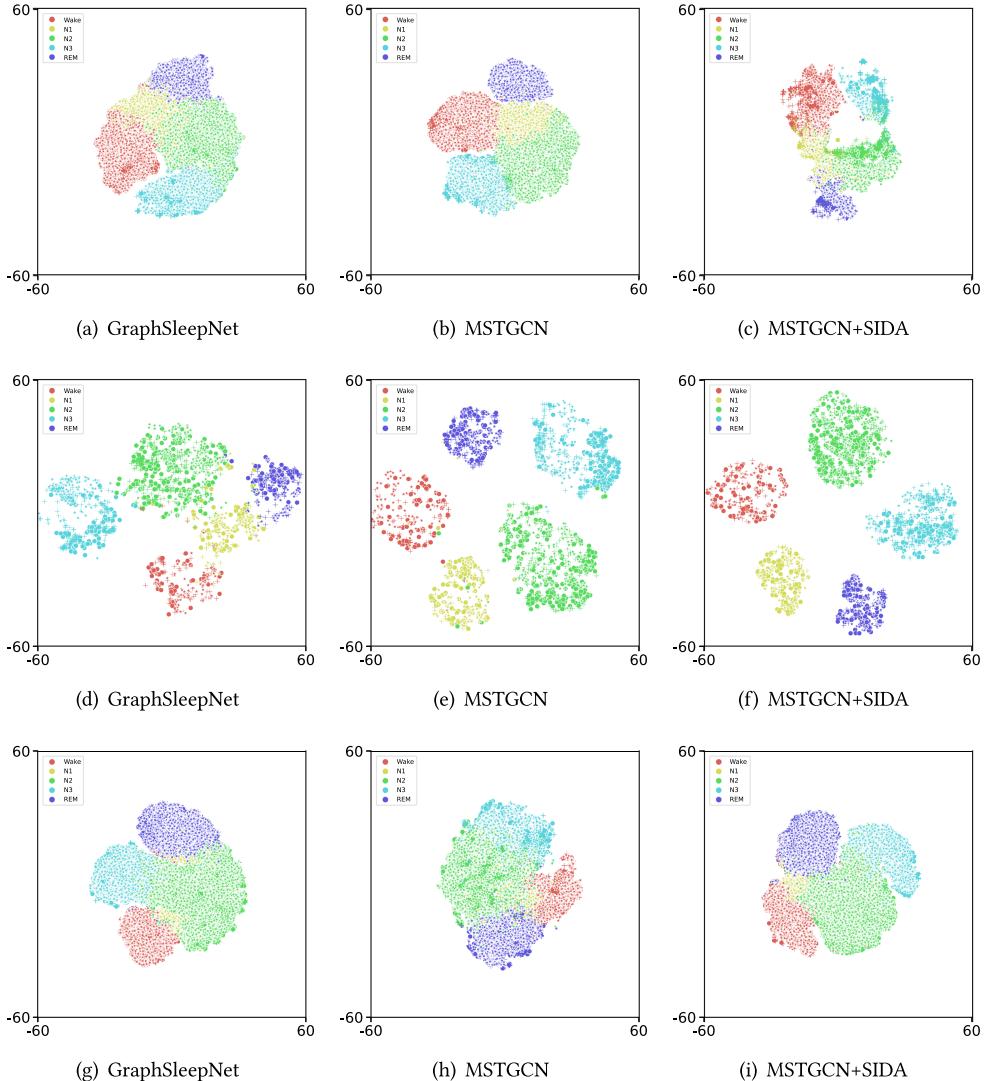


Fig. 8. Feature visualization of different methods on the ISRUC-S1 dataset with (a) GraphSleepNet, (b) MSTGCN, and (c) MSTGCN+SIDA methods, the ISRUC-S3 dataset with (d) GraphSleepNet, (e) MSTGCN, and (f) MSTGCN+SIDA methods, and the SHHS1 dataset with (g) GraphSleepNet, (h) MSTGCN, and (i) MSTGCN+SIDA methods. Compared with GrapSleepNet, MSTGCN mainly combines traditional domain adversarial methods. Three different shapes denote the data of three randomly selected subjects, and five colors denote different sleep stage categories.

method, the category boundary becomes distinct, and the model's dependency on the subject is weakened, enhancing cross-subject generalization, although misclassifications still occur. While our SIDA method results in each category's feature shape being close to circular, with a more apparent boundary and a significantly improved classification effect. Notably, the features of different subjects are evenly distributed, resulting in low subject dependency on the model. As shown in the results of the ISRUC-S1 dataset in Figure 8 (a), (b), and (c), The size of ISRUC-S1 is 10 times that of the ISRUC-S3 dataset, so it is tough to classify the GraphSleepNet method. The high-dimensional

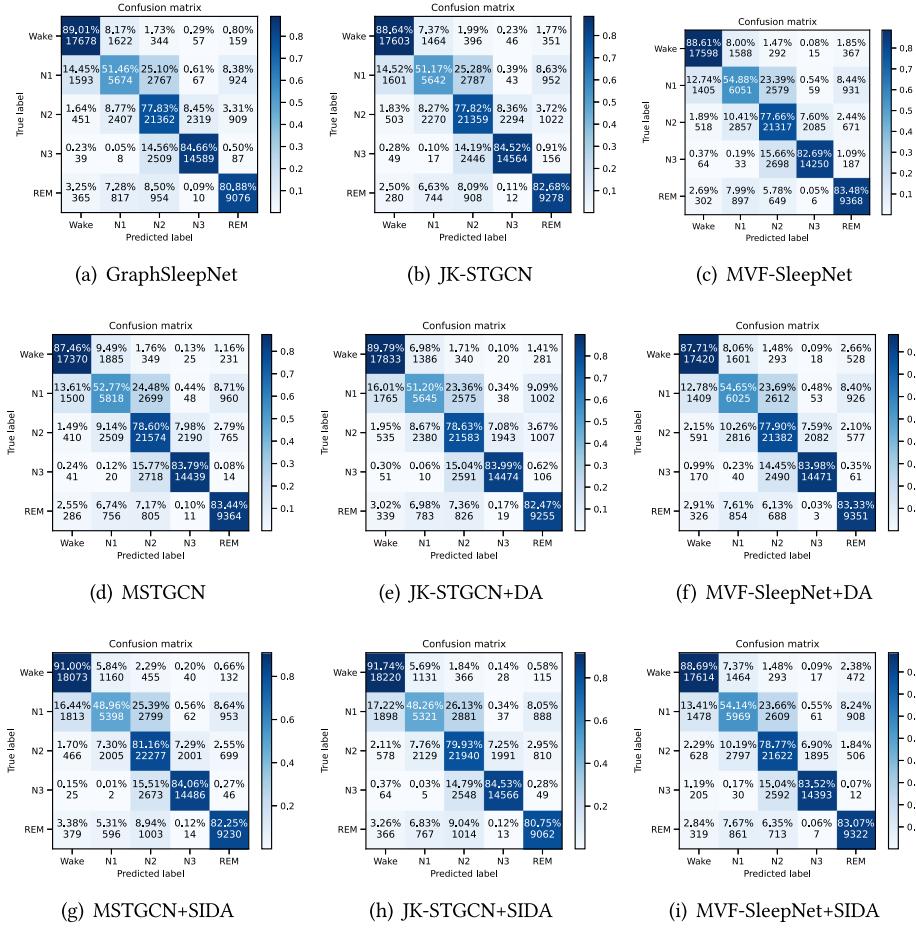


Fig. 9. The confusion matrix of different methods on the ISRUC-S1 dataset with (a) GraphSleepNet, (b) JK-STGCN, (c) MVF-SleepNet, (d) MSTGCN, (e) JK-STGCN+DA, (f) MVF-SleepNet+DA, (g) MSTGCN+SIDA, (h) JK-STGCN+SIDA, and (i) MVF-SleepNet+SIDA methods.

feature representation is difficult to separate. After adding the traditional domain adversarial method, the category boundary of the feature is vaguely visible, and each category is separated into different clusters. After using our SIDA method, the characteristics of different categories tend to be further separated. As shown in the SHHS1 dataset results. In Figure 8 (g), (h), and (i), the size of the SHHS1 dataset is about 40 times that of the ISRUC-S3 dataset, so it is especially difficult for the sleep stage N1 with a relatively small number of samples. Hidden features are difficult to assemble into a cluster, and many misclassifications occur. After adding the traditional domain adversarial method, wrong clustering is greatly improved. However, the distance between different categories is relatively close, and the interface is unclear. After using our SIDA method, the features of different categories tend to be further apart, and each category has a clear interface, the distribution of features across subjects also becomes uniform.

5.5 Confusion Matrix Analysis

As shown in Figures 9, 10, and 11, we analyzed the confusion matrix by combining the experimental results of our SIDA method and the comparative method on the three classical datasets

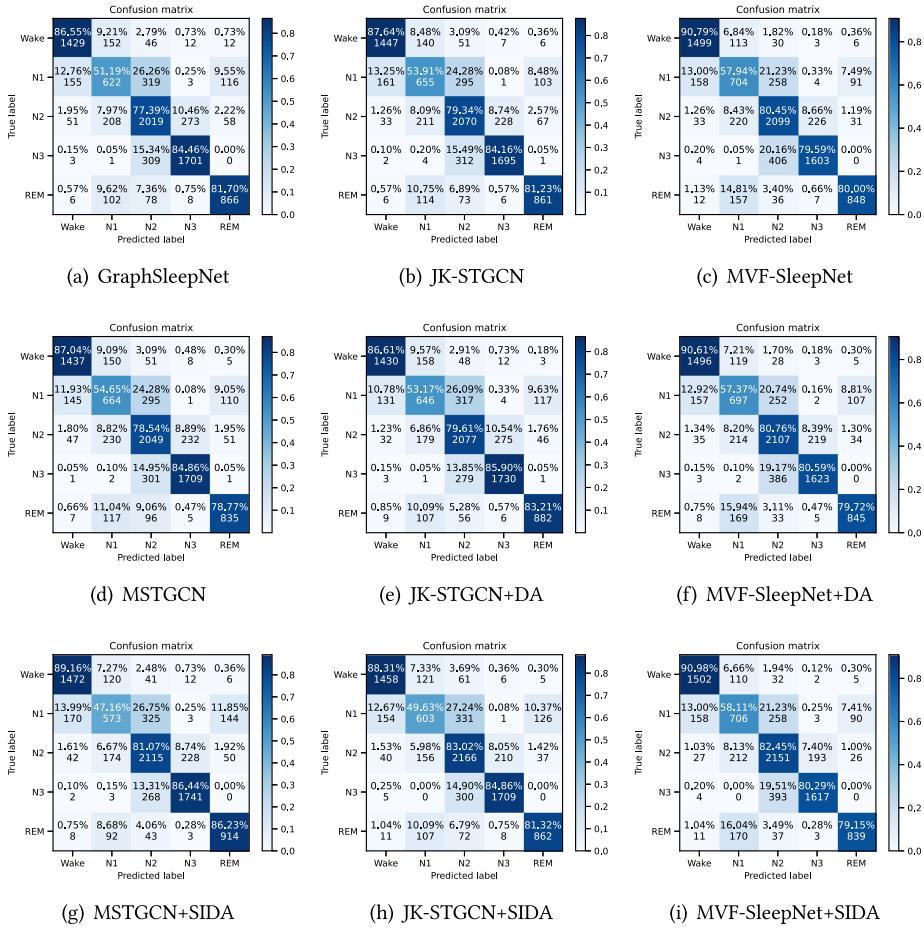


Fig. 10. The confusion matrix of different methods on the ISRUC-S3 dataset with (a) GraphSleepNet, (b) JK-STGCN, (c) MVF-SleepNet, (d) MSTGCN, (e) JK-STGCN+DA, (f) MVF-SleepNet+DA, (g) MSTGCN+SIDA, (h) JK-STGCN+SIDA, and (i) MVF-SleepNet+SIDA methods.

of the ISRUC-S1, ISRUC-S3, and SHHS1 datasets. From the confusion matrix, we can see that the classification results are generally good. However, the number of samples of the sleep stage N1 is the least, and the classification effect is the worst. This phenomenon is reflected in the previous classification methods of sleep stages. On the one hand, the number of samples of the sleep stage N1 is small. On the other hand, sleep stage N1 is a light sleep between the sleep stage Wake and N2. Physiologically, the brain is between the lightly active state and the light sleep state, and the signal fluctuation is slight, and the category changes are varied.

5.6 Loss and Accuracy Change Analysis

The loss and accuracy change of the training and validation sets of the ISRUC-S1, ISRUC-S3, and SHHS1 datasets during training are shown in Figure 12. It can be seen from the curve in the figure that the loss and accuracy of the training and validation set converge well during the training process, and the performance of the three datasets is basically the same. The loss and accuracy of the validation set converge earlier than the training set, so there is a particular gap between the effect of the training and validation set.

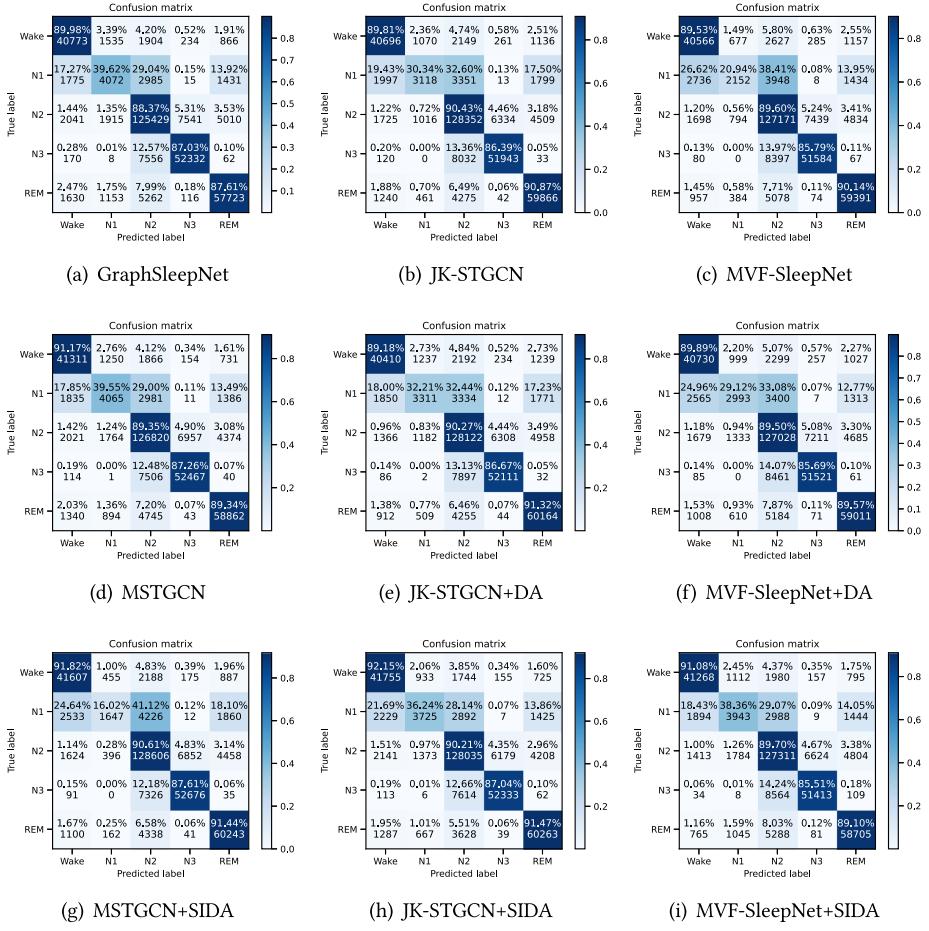


Fig. 11. The confusion matrix of different methods on the SHHS1 dataset with (a) GraphSleepNet, (b) JK-STGCN, (c) MVF-SleepNet, (d) MSTGCN, (e) JK-STGCN+DA, (f) MVF-SleepNet+DA, (g) MSTGCN+SIDA, (h) JK-STGCN+SIDA, and (i) MVF-SleepNet+SIDA methods.

5.7 Inference Time Analysis

Figure 13 depicts the average inference time per sample for each method during testing on the ISRUC-S1, ISRUC-S3, and SHHS1 datasets, with time measured in milliseconds. To maintain uniform representation, GraphSleepNet+DA in Figure 13 refers to the MSTGCN model. Since our SIDA is an improvement over the conventional method, it has more network parameters than the methods with the traditional domain generalization method and the baseline method that does not employ domain generalization. As a result, our model's inference time is longer. As seen in Figure 13(b), our method takes longer to process a single sample than other methods, with the inference time for a single sample with SIDA increasing by less than 25% compared to the methods with the traditional domain generalization method. The time increase is still acceptable. As seen in Figure 13(a) and (c), when incorporating some methods with the SIDA method on the ISRUC-S1 and SHHS1 datasets, the increase of inference time may be insignificant, and the time cost is also within an acceptable range.

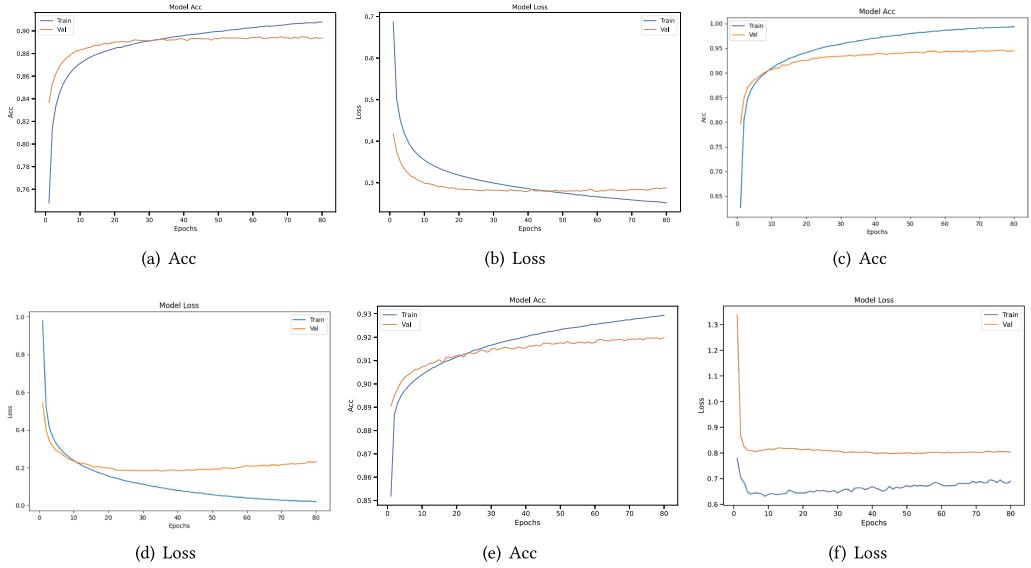


Fig. 12. Accuracy and loss change of training and validation set during MSTGCN with SIDA training on the ISRUC-S1 dataset with (a) and (b), the ISRUC-S3 dataset with (c) and (d), the SHHS1 dataset with (e) and (f).

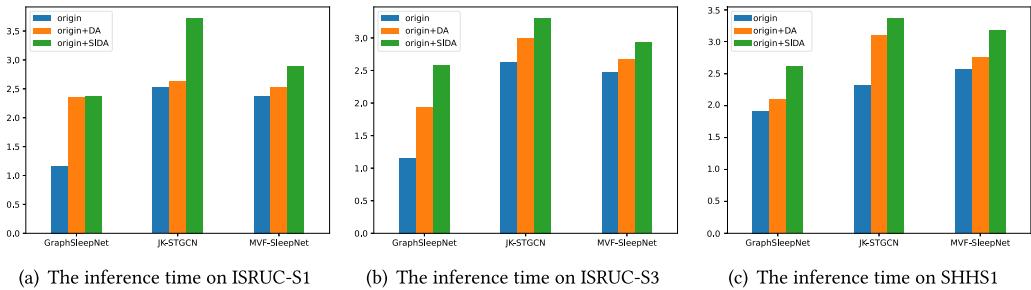


Fig. 13. The inference time of average per sample during methods testing on ISRUC-S1, ISRUC-S3, and SHHS1 datasets, respectively.

6 CONCLUSION AND FUTURE WORK

Inspired by the structure of sleep stage classification, we propose a plug-and-play method called the SIDA method. It considers the subject dependency differences between different sleep stage categories and aligns them category-by-category to improve the model's classification accuracy and subject generalization robustly. We integrate mainstream sleep stage classification methods and compare our method against the traditional domain adversarial method and the three latest state-of-the-art methods. Experimental results on three classic sleep stage classification datasets show that our SIDA method significantly improves the model's subject generalization and classification accuracy.

Looking to the future, we acknowledge that some challenges remain to be addressed. Our current method of soft weighting between classifiers and category-specific domain discriminators relies on neural network learning in an end-to-end manner, which is both convenient and efficient. However,

we observe that classifier performance is unsatisfactory in the early stages of training, leading to domain discriminator features being given inappropriate weights by the classifier, leading to training bias. Therefore, our goal is to improve the training process of the domain discriminator by using the training labels to adjust the weight distribution based on the classifier's current prediction accuracy. This approach aims to improve the accuracy and generalization of the model in the early stages of training by correcting the weight distribution based on the classifier's performance.

REFERENCES

- [1] Emina Alickovic and Abdulhamit Subasi. 2018. Ensemble SVM method for automatic sleep stage classification. *IEEE Trans. Instrum. Meas.* 67, 6 (2018), 1258–1265.
- [2] Nannapas Banluesombatkul, Pichayoot Ouppaphan, Pitshaporn Leelaarporn, Payongkit Lakan, Busarakum Chaitusaney, Nattapong Jaimchariyatam, Ekapol Chuangsuvanich, Wei Chen, Huy Phan, Nat Dilokthanakul, et al. 2020. MetaSleepLearner: A pilot study on fast adaptation of bio-signals-based sleep stage classifier to new individual subject using meta-learning. *IEEE J. Biomed. Health Inf.* 25, 6 (2020), 1949–1963.
- [3] Mathias Baumert, Simon Hartmann, and Huy Phan. 2023. Automatic sleep staging for the young and the old—Evaluating age bias in deep learning. *Sleep Med.* 107 (2023), 18–25.
- [4] Richard B. Berry, Rohit Budhiraja, Daniel J. Gottlieb, David Gozal, Conrad Iber, Vishesh K. Kapur, Carole L. Marcus, Reena Mehra, Sairam Parthasarathy, Stuart F. Quan, et al. 2012. Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events: Deliberations of the sleep apnea definitions task force of the American Academy of Sleep Medicine. *J. Clin. Sleep Med.* 8, 5 (2012), 597–619.
- [5] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. 2021. Domain generalization by marginal transfer learning. *J. Mach. Learn. Res.* 22, 1 (2021), 46–100.
- [6] Reza Boostani, Foroozan Karimzadeh, and Mohammad Nami. 2017. A comparative review on sleep stage classification methods in patients and healthy individuals. *Comput. Methods Progr. Biomed.* 140 (2017), 77–91.
- [7] Alexander A. Borbély, Fritz Baumann, Daniel Brandeis, Inge Strauch, and Dietrich Lehmann. 1981. Sleep deprivation: Effect on sleep stages and EEG power density in man. *Electroencephalogr. Clin. Neurophysiol.* 51, 5 (1981), 483–493.
- [8] Erik Bresch, Ulf Großkathöfer, and Gary Garcia-Molina. 2018. Recurrent deep neural networks for real-time sleep stage classification from single channel EEG. *Front. Comput. Neurosci.* 12 (2018), 1–12.
- [9] Stanislas Chambon, Mathieu N. Galtier, Pierrick J. Arnal, Gilles Wainrib, and Alexandre Gramfort. 2018. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Trans. Neural Syst. Rehabil. Eng.* 26, 4 (2018), 758–769.
- [10] Lan-lan Chen, Ao Zhang, and Xiao-guang Lou. 2019. Cross-subject driver status detection from physiological signals based on hybrid feature selection and transfer learning. *Expert Syst. Appl.* 137 (2019), 266–280.
- [11] Julie Anja Engelhard Christensen, Rick Wassing, Yishul Wei, Jennifer R. Ramautar, Oti Lakbila-Kamal, Poul Jørgen Jennum, and Eus J. W. Van Someren. 2019. Data-driven analysis of EEG reveals concomitant superficial sleep during deep sleep in insomnia disorder. *Front. Neurosci.* 13 (2019), 1–12.
- [12] Stefania Coelli, Eleonora Maggioni, Annalisa Rubino, Chiara Campana, Lino Nobili, and Anna M. Bianchi. 2019. Multiscale functional clustering reveals frequency dependent brain organization in type II focal cortical dysplasia with sleep hypermotor epilepsy. *IEEE Trans. Biomed. Eng.* 66, 10 (2019), 2831–2839.
- [13] Jessamyn Dahmen and Diane J. Cook. 2021. Indirectly supervised anomaly detection of clinically meaningful health events from smart home data. *ACM Trans. Intell. Syst. Technol.* 12, 2 (2021), 1–18.
- [14] Chenglong Dai, Dechang Pi, and Stefanie I. Becker. 2020. Shapelet-transformed multi-channel EEG channel selection. *ACM Trans. Intell. Syst. Technol.* 11, 5 (2020), 1–27.
- [15] Hao Dong, Akara Supratak, Wei Pan, Chao Wu, Paul M Matthews, and Yike Guo. 2017. Mixed neural network approach for temporal sleep stage classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* 26, 2 (2017), 324–333.
- [16] Irwin Feinberg. 1974. Changes in sleep cycle patterns with age. *J. Psychiatr. Res.* 10, 3-4 (1974), 283–306.
- [17] Nan Gao, Hao Xue, Wei Shao, Sichen Zhao, Kyle Kai Qin, Arian Prabowo, Mohammad Saiedur Rahaman, and Flora D. Salim. 2022. Generative adversarial networks for spatio-temporal data: A survey. *ACM Trans. Intell. Syst. Technol.* 13, 2 (2022), 1–25.
- [18] Narjes Goshtasbi, Reza Boostani, and Saeid Sanei. 2022. Sleepfcn: A fully convolutional deep learning framework for sleep stage classification using single-channel electroencephalograms. *IEEE Trans. Neural Syst. Rehabil. Eng.* 30 (2022), 2088–2096.
- [19] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. 2006. A kernel method for the two-sample-problem. *Adv. Neural Inf. Process. Syst.* 19 (2006), 1–8.
- [20] Ahnaf Rashik Hassan and Abdulhamit Subasi. 2017. A decision support system for automated identification of sleep stages from single-channel EEG signals. *Knowl.-Bas. Syst.* 128 (2017), 115–124.

- [21] Xiaopeng Ji, Yan Li, and Peng Wen. 2022. Jumping knowledge based spatial-temporal graph convolutional networks for automatic sleep stage classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* 30 (2022), 1464–1472.
- [22] Ziyu Jia, Youfang Lin, Jing Wang, Xiaojun Ning, Yuanlai He, Ronghao Zhou, Yuhan Zhou, and H Lehman Li-wei. 2021. Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* 29 (2021), 1977–1986.
- [23] Ziyu Jia, Youfang Lin, Jing Wang, Ronghao Zhou, Xiaojun Ning, Yuanlai He, and Yaoshuai Zhao. 2020. GraphSleepNet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI'20)*. 1324–1330.
- [24] Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. 2016. ISRUC-sleep: A comprehensive public dataset for sleep researchers. *Comput. Methods Progr. Biomed.* 124 (2016), 180–192.
- [25] Georgios Koutroulis, Leo Botler, Belgin Mutlu, Konrad Diwold, Kay Römer, and Roman Kern. 2021. KOMPOS: Connecting causal knots in large nonlinear time series with non-parametric regression splines. *ACM Trans. Intell. Syst. Technol.* 12, 5 (2021), 1–27.
- [26] Miroslav Kubat, Gert Pfurtscheller, and Doris Flotzinger. 1994. AI-based approach to automatic sleep classification. *Biol. Cybernet.* 70, 5 (1994), 443–448.
- [27] Annie C. Lajoie, Anne-Louise Lafontaine, and Marta Kaminska. 2021. The spectrum of sleep disorders in Parkinson disease: A review. *Chest* 159, 2 (2021), 818–827.
- [28] Yujie Li, Jingrui Chen, Wenjun Ma, Gansen Zhao, and Xiaomao Fan. 2022. MVF-SleepNet: Multi-view fusion network for sleep stage classification. *IEEE J. Biomed. Health Inf.* (2022), 1–11.
- [29] Christian O’reilly, Nadia Gosselin, Julie Carrier, and Tore Nielsen. 2014. Montreal archive of sleep studies: An open-access resource for instrument benchmarking and exploratory research. *J. Sleep Res.* 23, 6 (2014), 628–635.
- [30] Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22, 10 (2010), 1345–1359.
- [31] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. 2018. Multi-adversarial domain adaptation. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. Vol. 32, 3934–3941.
- [32] Mathias Perslev, Sune Darkner, Lykke Kempfner, Miki Nikolic, Poul Jørgen Jennum, and Christian Igel. 2021. U-sleep: Resilient high-frequency sleep staging. *NPJ Digit. Med.* 4, 1 (2021), 1–12.
- [33] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. 2019. SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Trans. Neural Syst. Rehabil. Eng.* 27, 3 (2019), 400–410.
- [34] Huy Phan, Kaare Mikkelsen, Oliver Y. Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos. 2022. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Trans. Biomed. Eng.* 69, 8 (2022), 2456–2467.
- [35] Huy Phan, Kaare Mikkelsen, Oliver Y. Chén, Philipp Koch, Alfred Mertins, Preben Kidmose, and Maarten De Vos. 2020. Personalized automatic sleep staging with single-night data: A pilot study with Kullback–Leibler divergence regularization. *Physiol. Meas.* 41, 6 (2020), 1546–1555.
- [36] Stuart F. Quan, Barbara V Howard, Conrad Iber, James P. Kiley, F Javier Nieto, George T. O’Connor, David M. Rapoport, Susan Redline, John Robbins, Jonathan M. Samet et al. 1997. The sleep heart health study: Design, rationale, and methods. *Sleep* 20, 12 (1997), 1077–1085.
- [37] Md Mosheyur Rahman, Mohammed Imamul Hassan Bhuiyan, and Ahnaf Rashik Hassan. 2018. Sleep stage classification using single-channel EOG. *Comput. Biol. Med.* 102 (2018), 211–220.
- [38] Rajeev Sharma, Ram Bilas Pachori, and Abhay Upadhyay. 2017. Automatic sleep stages classification based on iterative filtering of electroencephalogram signals. *Neural Comput. Appl.* 28, 10 (2017), 2959–2978.
- [39] Sudeep Sharma, Ashok Chhetry, Md Sharifuzzaman, Hyosang Yoon, and Jae Yeong Park. 2020. Wearable capacitive pressure sensor based on MXene composite nanofibrous scaffolds for reliable human physiological signal acquisition. *ACS Appl. Mater. Interf.* 12, 19 (2020), 22212–22224.
- [40] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. 2017. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* 25, 11 (2017), 1998–2008.
- [41] Akara Supratak and Yike Guo. 2020. TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG. In *Proceedings of the 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC’20)*. IEEE, 641–644.
- [42] Minfang Tang, Zhiwei Zhang, Zhengling He, Weisong Li, Xiuying Mou, Lidong Du, Peng Wang, Zhan Zhao, Xianxiang Chen, Xiaoran Li, et al. 2022. Deep adaptation network for subject-specific sleep stage classification based on a single-lead ECG. *Biomed. Sign. Process. Contr.* 75 (2022), 1–13.
- [43] Mohamed A. Tork, Hebatallah R. Rashed, Lobna Elnabil, Nahed Salah-Eldin, Naglaa Elkhayat, Ayman A. Abdelhady, M. Ossama Abdulghani, and Khaled O. Abdulghani. 2020. Sleep pattern in epilepsy patients: A polysomnographic study. *Egypt. J. Neurol. Psychiatr. Neurosurg.* 56, 1 (2020), 1–5.

- [44] Orestis Tsinalis, Paul M Matthews, and Yike Guo. 2016. Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders. *Ann. Biomed. Eng.* 44, 5 (2016), 1587–1597.
- [45] Huafeng Wang, Chonggang Lu, Qi Zhang, Zhimin Hu, Xiaodong Yuan, Pingshu Zhang, and Wanquan Liu. 2022. A novel sleep staging network based on multi-scale dual attention. *Biomed. Sign. Process. Contr.* 74 (2022), 1–10.
- [46] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. 2022. Generalizing to unseen domains: A survey on domain generalization. *IEEE Trans. Knowl. Data Eng.* 14, 8 (2022), 1–20.
- [47] Zaijian Wang, Shiwen Mao, Lingyun Yang, and Pingping Tang. 2018. A survey of multimedia big data. *Chin. Commun.* 15, 1 (2018), 155–176.
- [48] Liangjie Wei, Youfang Lin, Jing Wang, and Yan Ma. 2017. Time-frequency convolutional neural network for automatic sleep stage classification based on single-channel EEG. In *Proceedings of the IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI'17)*. IEEE, 88–95.
- [49] Edward A. Wolpert. 1969. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. *Arch. Gen. Psychiatr.* 20, 2 (1969), 246–247.
- [50] Ziliang Xu, Xuejuan Yang, Jinbo Sun, Peng Liu, and Wei Qin. 2020. Sleep stage classification using time-frequency spectra from consecutive multi-time points. *Front. Neurosci.* 14 (2020), 1–10.
- [51] Runze Yan, Xinwen Liu, Janine Dutcher, Michael Tumminia, Daniella Villalba, Sheldon Cohen, David Creswell, Kasey Creswell, Jennifer Mankoff, Anind Dey, et al. 2022. A computational framework for modeling biobehavioral rhythms from mobile and wearable data streams. *ACM Trans. Intell. Syst. Technol.* 13, 3 (2022), 1–27.
- [52] Rui Yan, Chi Zhang, Karen Spruyt, Lai Wei, Zhiqiang Wang, Lili Tian, Xueqiao Li, Tapani Ristaniemi, Jihui Zhang, and Fengyu Cong. 2019. Multi-modality of polysomnography signals fusion for automatic sleep scoring. *Biomed. Sign. Process. Contr.* 49 (2019), 14–23.
- [53] Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. 2018. The national sleep research resource: Towards a sleep data commons. *J. Am. Med. Inf. Assoc.* 25, 10 (2018), 1351–1358.
- [54] Xiang Zhang, Lina Yao, Chaoran Huang, Tao Gu, Zheng Yang, and Yunhao Liu. 2020. DeepKey: A multimodal biometric authentication system via deep decoding gaits and brainwaves. *ACM Trans. Intell. Syst. Technol.* 11, 4 (2020), 1–24.
- [55] Yingwei Zhang, Yiqiang Chen, Hanchao Yu, Zepign Lv, Pan Shang, Yiyi Ouyang, Xiaodong Yang, and Wang Lu. 2019. Wearable sensors based automatic box and block test system. In *Proceedings of the IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI'19)*. IEEE, 952–959.
- [56] Yingwei Zhang, Yiqiang Chen, Hanchao Yu, Xiaodong Yang, and Wang Lu. 2020. Dual layer transfer learning for sEMG-based user-independent gesture recognition. *Pers. Ubiq. Comput.* (2020), 1–12.
- [57] Caihong Zhao, Jinbao Li, and Yahong Guo. 2022. SleepContextNet: A temporal context network for automatic sleep staging based single-channel EEG. *Comput. Methods Progr. Biomed.* 220 (2022), 1–12.
- [58] Guohun Zhu, Yan Li, Peng Paul Wen, and Shuaifang Wang. 2014. Analysis of alcoholic EEG signals based on horizontal visibility graph entropy. *Brain Inf.* 1, 1 (2014), 19–25.
- [59] Hangyu Zhu, Wei Zhou, Cong Fu, Yonglin Wu, Ning Shen, Feng Shu, Huan Yu, Chen Chen, and Wei Chen. 2023. MasksleepNet: A cross-modality adaptation neural network for heterogeneous signals processing in sleep staging. *IEEE J. Biomed. Health Inf.* 27 (2023), 2353–2364.
- [60] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2020. A comprehensive survey on transfer learning. *Proc. IEEE* 109, 1 (2020), 43–76.

Received 22 February 2023; revised 29 June 2023; accepted 5 September 2023