ATTA: Adaptive Test-Time Adaptation for Multi-Modal Sleep Stage Classification

Ziyu Jia, **Xihao Yang**, **Chenyang Zhou**, **Haoyang Deng** and **Tianzi Jiang** [*]
Brainnetome Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China
{jia.ziyu, xihao.yang, chenyang.zhou1, haoyang.deng}@outlook.com, tianzi.jiang.iacas@gmail.com

Abstract

Sleep stage classification is crucial for sleep quality assessment and disease diagnosis. Although some recent studies have made great strides in sleep stage classification performance, direct application to multi-modal sleep data with cross-domain distributional variations still poses challenges: 1) How to retain the sleep knowledge acquired by the model from the source domain during cross-domain adaptation to avoid catastrophic forgetting. 2) How to evaluate the contribution of different modalities in identifying specific sleep stages to serve test-time adaptation (TTA). 3) How to dynamically adapt the sleep model to different distribution shift in data domains of different subjects. To address these challenges, we propose an Adaptive Test-Time Adaptation (ATTA) method, a multi-modal test-time adaptation method for sleep stage classification. Specifically, the intra-modal retained-adaptive module is proposed for adapting to the target domain data while retaining the sleep knowledge acquired from the source domain to avoid catastrophic forgetting. The inter-modal contribution assessment module is designed to adaptively assess the contribution of each modality to the identification of specific sleep stages. Furthermore, the adaptive learning rate strategy utilizes a memory bank to record data from different subjects during testing, and based on this, it measures the differences between the target subject and those in the memory bank. According to the difference, the model adapts to the subject samples with different learning rates. We conduct experiments on mutual migration on two sleep datasets, SleepEDF and SHHS. The results show that our ATTA method outperforms state-of-the-art baselines in sleep stage classification.

1 Introduction

Sleep stage classification is essential for the assessment of sleep quality and the diagnosis of sleep disorders [Liu and Jia, 2022; Cai *et al.*, 2021]. With the development of sensor technology, physiological time series data based on synchronized multi-sensor acquisition can effectively improve the effect of sleep stage classification [Zhou *et al.*, 2023]. Specifically, existing methods use multi-modal physiological time series such as electroencephalogram (EEG) and electrooculogram (EOG) [Liu *et al.*, 2024], which provide complementary information for accurately identifying specific sleep stages. Therefore, it is important to develop effective multi-modal fusion methods for sleep stage classification. Although existing methods have made great strides, there are still some challenges in applying them directly to multi-modal sleep data:

1) *When adapting to new subject data during testing, traditional test-time adaptation (TTA) methods may result in catastrophic forgetting in the sleep model.* Multi-modal data from different subjects are unique [Shin *et al.*, 2022]. As a result, traditional models trained based on some subjects can hardly achieve good generalization performance on other subjects. In order to solve this problem, test-time adaptation specifically has greater potential to be applied to sleep stage classification. However, when existing methods are directly applied to dynamically adapt the model to new subject data at the time of testing, the model may forget previously acquired multi-modal knowledge, which leads to catastrophic forgetting problems in the model [Niu *et al.*, 2022]. Therefore, it is challenging to adapt the model to the target data during testing while allowing the model to retain the acquired knowledge within each modality.

2) *It is a challenge to evaluate the contribution of different modalities for sleep stage classification.* Different modalities contribute differently when identifying various sleep stage. For example, in recognizing REM and N1 stages, the EOG signal contributes more than the EEG signal. In recognizing the N3 stage, the EEG signal contributes more than the EOG signal [Jia *et al.*, 2021b]. Thus, multi-modal information helps to identify sleep stages. In order to fully utilize these multi-modal data, most of the current studies focus only on the information complementarity property of multi-modality, ignoring the fact that each modality contributes differently to the identification of a specific sleep stage [Chambon *et al.*, 2018]. Therefore, it is crucial to estimate the contribution of each modality to serve TTA.

3) *Traditional TTA methods make it difficult for the model to adapt to various distribution shift in the data domain for*

[*] Corresponding Author

different subjects. The need for adaptation in sleep data vary across subjects. For example, domains with larger shift should match a larger degree of adaptation, thus facilitating feature alignment, while domains with smaller shift should be matched with a smaller degree of adaptation. Otherwise, the model might deviate from the distinctive features learned during pre-training [Yang *et al.*, 2022]. Traditional TTA methods using a fixed learning rate cannot adapt themselves to different subjects. For example, T-TIME [Li *et al.*, 2023] uses a fixed learning rate to update the model, neglecting the importance of assessing the adaptation needs of subjects. Therefore, it is a challenge to adaptively match the appropriate learning rate for different subjects.

To address the above challenges, we propose an Adaptive Test-Time Adaptation (ATTA) method for multi-modal sleep stage classification. The proposed model consists of an intra-modal retained-adaptive module, an inter-modal contribution assessment module and an adaptive learning rate strategy. The following are our main contributions:

- We propose an intra-modal retained-adaptive module that dynamically adapts to intra-modal data during testing while retaining previously acquired modality-specific knowledge and avoiding catastrophic forgetting.
- We design an inter-modal contribution assessment module, which accurately evaluates the extent to which different modalities contribute to sleep stage classification.
- We develop an adaptive learning rate strategy based on a memory bank, which can meet the adaptation needs of different subjects.
- Experimental results demonstrate that the ATTA achieves state-of-the-art performance in sleep stage classification.

2 Related Work

2.1 Sleep Stage Classification

Sleep stage classification serves as the foundation role of studying sleep disorders, holding significant clinical importance [Liang *et al.*, 2023]. Early studies employ machine learning methods such as Support Vector Machines (SVM) and Random Forests (RF) for sleep stage classification. Subsequently, models combining single-modal physiological signals with Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are widely used. For instance, DeepSleepNet [Supratak *et al.*, 2017] extracts time-invariant features from EEG channels, and utilizes Bidirectional Long Short-Term Memory (BiLSTM) to learn the transition rules among sleep stages. After that, TinySleepNet [Supratak and Guo, 2020] is proposed to reduce the complexity of existing models such as DeepSleepNet. SleepEEGNet [Mousavi *et al.*, 2019] can extract time-invariant features from raw channel signals and capture long short-term dependencies from context. U-Time [Perslev *et al.*, 2019] proposes a time-domain information method, employing a temporal fully convolutional network for physiological time series segmentation, used for the analysis of sleep data. Furthermore, some research has indicated that multi-modal methods outperform single-modal methods in sleep stage classification. For example, SleepPrintNet [Jia *et al.*, 2020a] integrates spatio-temporal features of EEG and multi-modal features such as EOG and Electromyogram (EMG) into the sleep stage classification model, enhancing the classification performance. XSleepNet [Phan *et al.*, 2021] proposes a sequence-to-sequence sleep stage classification model, which is able to learn a joint representation from both raw signals and time-frequency images effectively. SalientSleepNet [Jia *et al.*, 2021b] utilizes a temporal fully convolutional network based on the U^2-Net architecture for multi-modal salient wave detection in sleep stage classification.

Inter-subject variability is a significant challenge in sleep stage classification. The MSTGCN [Jia *et al.*, 2021a] suggests a domain-generalized multi-view spatiotemporal graph convolutional network for sleep stage classification. The SEN-DAL [Jia *et al.*, 2022] puts forward a multi-modal physiological signal-based domain adversarial learning squeeze and excitation network to reduce the inter-subject variability. An Unsupervised Domain Adaptation (UDA) method [Yoo *et al.*, 2021] utilizes transferring structured knowledge in a sleep staging network, reducing the inter-subject variability by reorganizing domains and producing domain-invariant features within the same space. A transfer learning method [Phan *et al.*, 2020] applies end-to-end deep learning frameworks to sequence-to-sequence sleep stage classification models, focusing on the problem of inter-subject variability.

Although existing sleep stage classification models have achieved high classification performance, most of them are still unable to effectively solve the multi-modal data fusion and distribution shift problems.

2.2 Test-Time Adaptation

The purpose of TTA is to make existing models quickly adapt the new target data without accessing the source data used for training. As an effective technique for dealing with dynamic domain displacement problems in the real world, TTA has received increasing attention in practical applications. Wang et al. [2020] propose the first TTA method, TENT, a simple yet effective entropy minimization approach for optimizing test-time batch normalization parameters without the need for any proxy tasks during the training process. Building upon TENT, Wang et al. [2022] improve test performance through knowledge distillation and a teacher-student model for adaptation. Niu et al. [2022] study on the update efficiency of TTA, utilizing entropy to filter prediction-confident samples and weighting test samples based on both entropy and a mean vector updated by exponential moving average.

Furthermore, to enhance transfer robustness, the following researches are presented. Lim et al. [2023] adjust the weight of statistical moments interpolated between conventional batch normalization and transductive batch normalization layers based on the sensitivity to domain shift in each batch normalization (BN) layer. This adjustment improves the robustness of this model across various batch sizes and practical evaluation scenarios in the mobile domain. Chen et al. [2022] improve well-trained models through contrastive learning and pseudo-labeling strategies, optimizing the entire source model module. Finally, focusing on the stability

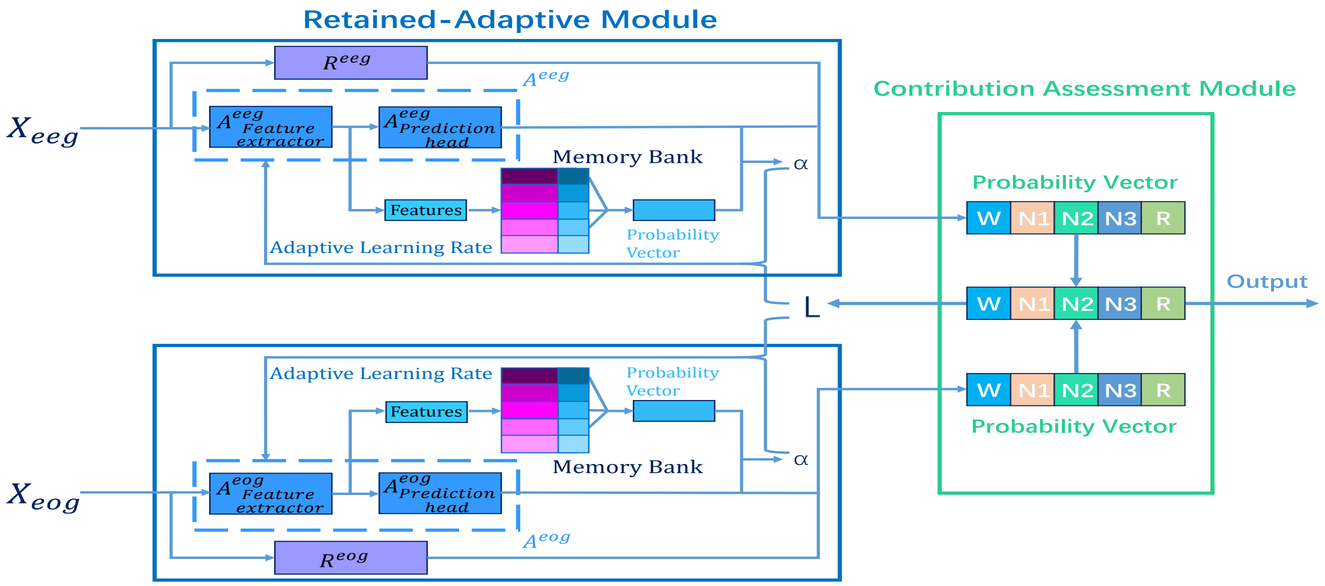

Figure 1: The overview of the proposed adaptive test-time adaptation (ATTA). The proposed ATTA consists of two modules: the intra-modal retained-adaptive module and the inter-modal contribution assessment module. For the intra-modal retained-adaptive module, the retained model effectively preserves the sleep knowledge obtained from the source domain, while the adaptive model rapidly adapts to the sleep data distribution of the target domain. Regarding the inter-modal contribution assessment module, we employ a balanced modal fusion strategy to merge pseudo-labels from each modality. To enable the adaptive model to better adapt to the sleep data distribution of the target domain, the adaptive learning rate strategy is proposed to update the adaptive model. R^M represents the retained model, and A^M represents the adaptive model. Here, $M \in \{\text{eeg}, \text{eog}\}$.

of the TTA runtime, Niu et al. [2023] propose a sharpness-aware and reliable entropy minimization method to stabilize the TTA process. Yang et al. [2022] present a discrepancy measurement to address the frequent changes in the scale of distribution shift. Zhao et al. [2023] apply a dynamic online reweighting strategy to weight each test sample when calculating conditional entropy, coping with the imbalance in testing data. However, most of the above methods ignore the significance of combining multi-modal data with adaptive TTA, for which we propose ATTA.

3 Adaptive Test-Time Adaptation

The overall pipeline of our ATTA is illustrated in Figure 1. Our ATTA consists of an intra-modal retained-adaptive module, an inter-modal contribution assessment module, and an adaptive learning rate strategy. We summarize three key ideas of our ATTA method:

1) The intra-modal retained-adaptive module can rapidly adapt to modality-specific data during testing and generate effective modality-specific pseudo-labels while retaining the knowledge acquired from the source domain to avoid catastrophic forgetting.

2) In the inter-modal contribution assessment module, we adopt a balanced modal fusion strategy, merging probability distribution of each modality based on similarity metric to fully utilize the information from each modality.

3) We adopt an adaptive learning rate strategy based on a memory bank to assess the adaptation needs of different test samples and dynamically adjust the learning rate for each test sample.

3.1 Intra-Modal Retained-Adaptive Module

During the test-time adaptation process, to mitigate the instability of the model caused by catastrophic forgetting, such as prediction fluctuations due to inter-subject variability, we design the retained-adaptive model. This model outputs probability distribution for each modality to better facilitate model updates. During testing, the initial weight parameters of both the retained model and the adaptive model are consistent with the model parameters learned from the source domain, but they are updated in different ways. The retained model effectively preserves the sleep knowledge acquired during the source domain learning process while adapting to the target domain. This ensures that our model does not lose or weaken critical information accumulated in the source domain when adapting to target domain data. Furthermore, the adaptive model demonstrates the ability to rapidly adapt to the sleep data distribution of the target domain and align with the features of the target domain, thereby enhancing the generalization ability of the model on the target domain. For each batch, we average the probability distribution output by the retained model and the adaptive model to compute the intra-modal probability distribution:

$$p(x^M) = \frac{p_A(x^M) + p_R(x^M)}{2}, \tag{1}$$

where the test sample x^M is input into the retained model and the adaptive model, respectively, generating corresponding probability distributions $p_A(x^M)$ and $p_R(x^M)$. Here,

$M \in \{\text{eeg}, \text{eog}\}$. A represents the adaptive model, R represents the retained model, and $p(x^M)$ represents the probability distribution within the modality. The probability vector for each modality is defined as:

$$\hat{y}^M = \text{argmax}(p(x^M)), \tag{2}$$

where $\hat{y}^M$ represents the intra-modal pseudo-label. We update the retained model and the adaptive model with different strategies. The adaptive model is updated using an adaptive learning rate strategy and its batch normalization statistics are updated according to the testing data. After the update of the adaptive model is completed, the retained model adopts a momentum-based update strategy [Shin *et al.*, 2022] to update the parameters of the batch normalization layer (BN):

$$\Omega^M_{R_{t_{i+1}}} = (1-\lambda)\Omega^M_{R_{t_i}} + \lambda\Omega^M_{A_{t_{i+1}}}, \tag{3}$$

where $\Omega^M_{R_{t_{i+1}}} = (\mu, \sigma, \gamma, \beta)^M_{R_{t_{i+1}}}$ represents the batch normalization parameters, including normalization statistics and transformation parameters. μ^M and σ^M are the normalization statistics, while γ^M and β^M are the transformation parameters. The initial statistics $\Omega^M_{R_{t_0}}$ are from the source pre-trained model. The update is restricted to the parameters of the BN layer for efficiency considerations. λ is the momentum used to control the speed at which the retained model moves toward the target domain. A larger λ results in a faster movement of the retained model toward the target, while a smaller λ leads to a slower movement.

3.2 Inter-Modal Contribution Assessment Module

To take full advantage of the complementary advantages of multi-modal data, we merge probability distribution from the two modalities and generate the inter-modal pseudo-label to serve TTA by comparing the similarity metric for each modality. This allows the model to better characterize the roles of different modalities during updating. Firstly, the similarity metric for each modality is defined as:

$$S^M = Sim(A^M(x^M), R^M(x^M)), \tag{4}$$

where the similarity metric S^M is used to assess whether the probability distribution outputted by the retained-adaptive model within each modality are reliable. If the similarity between the probability distribution outputted by the retained model and the adaptive model for a certain modality is low, it indicates that the output of the modality is unreliable. Conversely, if the similarity is high, it indicates that its output is reliable. $Sim(x, y)$ denotes the similarity function, which is used to calculate the similarity metric. We employ the Kullback-Leibler (KL) divergence to measure the similarity, which is defined as:

$$Sim(x, y) = \frac{\frac{1}{D_{\text{KL}}(x\|y)+\epsilon} + \frac{1}{D_{\text{KL}}(y\|x)+\epsilon}}{2}. \tag{5}$$

where D_{KL} denotes the KL divergence and ϵ is a small scalar constant to prevent division-by-zero. Then, we propose the balanced modal fusion strategy to evaluate the contributions of different modalities and fuse the probability distribution of each modality, and obtain the inter-modal pseudo-label $\hat{y}$:

$$\hat{y} = \text{argmax}(p(x)), \tag{6}$$

$$p(x) = \frac{S^{\text{eeg}}}{S^{\text{eeg}} + S^{\text{eog}}} p(x^{\text{eeg}}) + \frac{S^{\text{eog}}}{S^{\text{eeg}} + S^{\text{eog}}} p(x^{\text{eog}}), \tag{7}$$

where we calculate the contribution of each modality by normalizing the similarity metrics S^{eeg} and S^{eog} for different modalities, and then obtain the probability distribution $p(x)$ through weighted summation. Finally, we obtain the inter-modal pseudo-label.

3.3 Adaptive Learning Rate Strategy

To address the adaptation variability of sleep data from different subjects, we employ an adaptive learning rate strategy to update the adaptive model for each modality, which helps dynamically adjust the learning rate for different subjects. In fact, the proposed strategy enables the adaptive model to rapidly adapt to the diverse sleep data distribution in the target domain. For test samples with significant distribution shift from the training data, the model needs to adjust the learning rate more drastically to fully learn the features of the target samples; while for test data with smaller distribution shift, the model only needs to update the learning rate in small increments without drastically updating the learning rate, which may otherwise deviate from the discriminative sleep features learned from the source domain. Our adaptive model update process can be summarized as the following flow:

For the test samples arriving sequentially, the model outputs a feature vector after passing through the feature extraction layer. Then, based on the feature vector, the memory bank is retrieved to find the K nearest feature vectors. We use the L2 distance to measure the proximity d_i^M between them. d_i^M is defined as:

$$d_i^M = d(A_{\text{Fe}}(x^M), q_i^M) = \| A_{\text{Fe}}(x^M) - q_i^M \|_2, \tag{8}$$

where q_i^M denotes the i-th feature vector to be queried in the memory bank. $A_{\text{Fe}}(x^M)$ represents the feature vector output after the mini-batch input x^M passes through the feature extraction layer of the adaptive model. Then, the weighted sum of the probability vectors corresponding to these K feature vectors is defined as:

$$\hat{p}(x^M) = \frac{1}{K}\sum_{i=1}^{K} d_i^M \cdot v_i^M, \tag{9}$$

where v_i^M is the probability vector corresponding to q_i^M in the memory bank. Then, we compute the similarity α between $\hat{p}(x^M)$ and $p(x^M)$ by KL divergence and adjust the learning rate:

$$\eta = \alpha \cdot \eta_0, \tag{10}$$

where η_0 is the initial value of the learning rate. Then, the learning rate and the loss function are used to update the adaptive model to achieve the desired level of adaptation for the sample. The loss function L is defined as:

$$L = L^{\text{eeg}}_{\text{ent}} + L^{\text{eog}}_{\text{ent}}, \tag{11}$$

$$L^M_{\text{ent}} = L_{\text{ent}}(p(x^{\text{eeg}}), \hat{y}^{\text{eeg}}) + L_{\text{ent}}(p(x^{\text{eog}}), \hat{y}^{\text{eog}}), \tag{12}$$

where L consists of the cross-entropy loss of both modalities.

It should be noted that corresponding to each test sample, at the end of each adaptation step, the adaptive model performs forward passes again to regain the prediction results and update the element pairs in the memory bank according to the First In First Out (FIFO) principle. In the process of test-time adaptation, the parameters of the model are closely related to the test samples in a number of recent time steps. According to the memory bank, we predict the probability vector output by the adaptive model at the current time step based on the probability vectors corresponding to the samples from the most recent number of time steps. The hidden idea behind this is that the most recent test samples influence the current output to a greater extent. In contrast, updates to the model from more distant time-step samples are overwritten by updates to the model from more recent time-step samples. As new test samples arrive in sequence, the adaptive model for each modality is iteratively updated according to the above process.

4 Experiments

We evaluate the performance of our ATTA method and baseline methods on SleepEDF and SHHS datasets.

4.1 Datasets and Preprocessing

The SleepEDF dataset [Kemp *et al.*, 2000] is composed of 153 sleep recordings from 78 healthy subjects aged between 25 and 101 years. Except for subjects 13, 36, and 52, who lose one record each due to equipment failure, each subject has two full-day polysomnography (PSG) recordings. These PSG recordings are segmented into epochs of 30 seconds, with sleep stages labeled as {W, R, N1, N2, N3, N4, Movement, Unknown}. In this experiment, following the sleep classification standards of the American Academy of Sleep Medicine (AASM), stages N3 and N4 are merged into a single N3 stage, and the Movement and Unknown stages are removed. The final set of sleep stages includes {W, N1, N2, N3, REM}. For EEG signals within the PSG, the FPZ-CZ channel is utilized, and for EOG signals, the ROC-LOC EOG channel is employed.

The SHHS dataset [Quan *et al.*, 1997] comprises sleep data from 329 individuals selected from the SHHS1 cohort of the Sleep Heart Health Study, whose Apnea Hypopnea Indexes (AHI) is lower than 5 [Ma *et al.*, 2023]. Each subject has one PSG recording, processed in the same manner as the SleepEDF dataset. The EEG signals use the C4-A1 channel, and the EOG signals use the Right EOG channel.

In the experimental evaluation, we implement a cross-dataset transfer setup using the datasets mentioned above. Specifically, we train a source model on one dataset as the source domain and then assess the model on the other dataset as the target domain. To accommodate the different sampling frequencies of the various channels, we apply resampling techniques to standardize them to 100 Hz.

4.2 Baseline Methods

We compare our ATTA method against the following six baseline methods:

- **No Adapt**: A baseline approach that involves pre-training on source data and then testing directly on target domain data without any adaptation.
- **MSTGCN**[Jia *et al.*, 2021a]: A domain generalization method for sleep stage classification.
- **SEN-DAL**[Jia *et al.*, 2022]: A sleep stage classification method using a domain adversarial learning network based on multi-modal physiological signals.
- **TENT**[Wang *et al.*, 2020]: A method that utilizes the principle of entropy minimization to optimize test-time batch normalization parameters.
- **EATA**[Niu *et al.*, 2022]: A fully TTA method that uses entropy to filter predictably reliable samples and applies weighting to both samples and significant parameters.
- **SAR**[Niu *et al.*, 2023]: An algorithm that incorporates sharpness-awareness and reliable entropy minimization to stabilize the operation of TTA.

4.3 Experiment Settings

We implement our ATTA method using the PyTorch framework. ATTA employs a simplified version of SalientSleepNet [Jia *et al.*, 2021b] as the backbone network for pre-training, where both the MSE and MMA modules are removed. Our model is trained by Adam optimizer with initial learning rate $\eta = 10^{-3}$. The momentum parameter λ, used to control the speed at which the retained model moves toward the target domain, is set to 0.02. The batch size is 20 and the memory bank size is 16. Besides, the number of feature vectors retrieved from the memory bank is 8. More details about the model can be found in our code, which will be released on GitHub later.

4.4 Experiment Results

We conduct mutual transfer experiments on the SleepEDF and SHHS datasets. The experimental results are shown in Table 1.

Method	SleepEDF→SHHS		SHHS→SleepEDF	
	Accuracy	F1-score	Accuracy	F1-score
No Adapt	64.09	61.66	71.06	70.00
MSTGCN	68.48	62.08	71.35	71.04
SEN-DAL	67.35	55.88	71.97	65.32
TENT	70.82	66.83	73.73	**73.08**
EATA	68.18	68.98	71.08	71.63
SAR	72.75	71.81	72.91	72.66
ATTA	**78.04**	**77.81**	**74.15**	72.91

Table 1: Comparison of the results obtained by mutual transfer between SleepEDF and SHHS datasets, where the optimal results are shown in bold and the sub-optimal results are shown with a horizontal line below them. "A→B" for the transfer learning method indicates transfer from dataset A to B, and the No Adapt method indicates training on dataset A and direct testing on dataset B.

Table 1 demonstrates our proposed method ATTA, which achieves superior classification performance on most evaluation metrics compared to other baseline methods. As shown in Table 1, all TTA and domain generalization (DG) methods show improvement over the baseline method No Adapt, indicating the advantage of most transfer learning approaches in addressing domain shift issues. Our ATTA method effectively captures and utilizes the complementarity of multi-modal data, equipped with an adaptive learning rate strategy. The proposed ATTA modifies the update process of the test-time adaptive model by dynamically adjusting the learning rate according to the adaptation needs of different subject data. Therefore, our ATTA has a greater advantage in overcoming domain shift in data distribution, yielding results superior to most existing methods.

In certain cases, DG methods (MSTGCN and SEN-DAL) perform better than the poorest-performing TTA methods, such as the accuracy of MSTGCN being higher than EATA in the transfer from SleepEDF to SHHS, though their results are very close and do not exceed a difference of 1%. The results of DG methods are generally lower than most TTA methods. With a larger number of samples, SHHS datasets provide overall higher results when trained as the source domain compared to when SleepEDF serves as the source domain. This indicates that most baseline methods still rely on larger volumes of source data to achieve optimal learning effects. However, our ATTA still performs well when trained on a smaller source domain, demonstrating the robustness of the proposed method in accurately classifying with insufficient source data.

The traditional No Adapt method, without additional adaptation, experiences a decline in model performance when faced with new data distribution. DG models improve performance across datasets by extracting domain-invariant features between different domains, thus enhancing the model's effectiveness compared to No Adapt method. However, this improvement is relatively minor, and the models have limited generalizability. TTA methods can adapt during testing, effectively addressing the continuous changes in real-world data distribution and further improving the accuracy of sleep stage classification. However, traditional TTA methods risk catastrophic forgetting, losing substantial knowledge from the source domain during the test adaptation process. Additionally, traditional TTA methods neglect effective modeling of multi-modal data. To address this, the proposed ATTA method employs an intra-modal retained-adaptive module to dynamically adapt to sample data while preserving source domain knowledge during testing, uses an inter-modal contribution assessment module to evaluate the contribution of each modality to sleep stage classification, and adopts an adaptive learning rate strategy to meet the adaptation needs of different subjects. Therefore, the proposed model achieves competitive results compared to other baseline methods.

Additionally, we test the performance of our ATTA method under different test sample arrival orders. Table 2 shows that regardless of the order of test samples, ATTA maintains stable performance, thus demonstrating that our method is unaffected by the order of test samples and can flexibly handle and adapt to the real-world scenario of random testing across different subjects.

4.5 Ablation Experiment

To investigate the effectiveness of the various components within ATTA, we design several different variants:

Variant 1: ATTA-Only with Adaptive (Basic). To validate the effectiveness of using the adaptive model in conjunction with the retained model, this variant removes the retained model from ATTA. The probability distribution outputs from each modality, after processing through the adaptive model, are combined via arithmetic mean to produce the final output.

Order Number	Accuracy(%)	F1-score(%)
1	78.04	77.81
2	77.61	77.41
3	77.56	77.22
4	77.40	77.43
5	77.76	77.50

Table 2: The impact of different sample arrival sequences on model performance.

Variant 2: ATTA-Equal Average Modal Fusion (EAMF). This variant keeps the retained model within ATTA and obtains the final output by taking the arithmetic average of the outputs from both modalities:

$$\hat{y}_2 = \text{argmax}(p(x)), \tag{13}$$

$$p(x) = \frac{p(x^{\text{eeg}}) + p(x^{\text{eog}})}{2}, \tag{14}$$

where $\hat{y}_2$ represents the the inter-modal pseudo-label.

Variant 3: ATTA-Preferential Modal Fusion (PMF). This variant determines the final output by selecting the output from the modality with the higher value of the similarity metric:

$$\hat{y}_3 = \begin{cases} \hat{y}^{\text{eeg}}, & \text{if } S^{\text{eeg}} > S^{\text{eeg}}, \\ \hat{y}^{\text{eog}}, & \text{otherwise}, \end{cases} \tag{15}$$

$$\hat{y}^M = \text{argmax}(p(x^M)), \tag{16}$$

where $\hat{y}_3$ represents the inter-modal pseudo-label.

Variant 4: ATTA-Random Weighted Modal Selection (RWMS). In this variant, the final output is decided based on the outputs from each modality weighted by their respective similarity metric, normalized as probabilities:

$$\Phi_{\text{eeg}} = \frac{S^{\text{eeg}}}{S^{\text{eeg}} + S^{\text{eog}}}, \Phi_{\text{eog}} = 1 - \Phi_{\text{eeg}}, \tag{17}$$

$$\hat{y}_4 = \begin{cases} \hat{y}^{\text{eeg}}, & \text{if } r < \Phi_{\text{eeg}}, \\ \hat{y}^{\text{eog}}, & \text{otherwise}, \end{cases} \tag{18}$$

$$\hat{y}^M = \text{argmax}(p(x^M)), \tag{19}$$

where Φ_{eeg} and Φ_{eog} measure the contributions of the EEG and EOG modalities respectively, r is a random number and $r \in [0, 1)$.

Our Method: ATTA-Balanced Modal Fusion (BMF). Our approach in ATTA involves multiplying the outputs from each modality by their respective similarity metric, then taking the arithmetic mean and normalizing it to produce the final output.

Table 3 displays the classification results for different variants. A comparison between Variant 1 and Variant 2 shows that the retained model significantly enhances performance. This improvement is due to the ability of the retained model to effectively preserve essential knowledge from the source

Variant	Retained-Adaptive Module		Contribution Assessment Module	Evaluation Metrics	
	Retained	Adaptive	Fusion Strategy	Accuracy(%)	F1-score(%)
Variant 1	✗	✓	-	71.32	71.15
Variant 2	✓	✓	EAMF	73.87	73.33
Variant 3	✓	✓	PMF	75.13	75.48
Variant 4	✓	✓	RWMS	75.45	75.53
Our Method	✓	✓	BMF	78.04	77.81

Table 3: Experimental results validating the effectiveness of the retained model and the fusion strategies of the contribution assessment module.

model, which prevents catastrophic forgetting and enhances robustness. Additionally, comparisons among Variants 2, 3, 4, and our method indicate that the proposed inter-modal contribution assessment module is effective, with the complete ATTA model achieving optimal performance. Our method, unlike the basic approach in Variant 2, considers the simultaneous predictions from two models within the same modality. Compared to Variants 3 and 4, the proposed inter-modal contribution assessment module accurately delineates the contributions of different modalities to sleep stage classification.

Figure 2 displays the results of fixed learning rate and adaptive learning rate under different initial learning rate settings. Clearly, the adaptive learning rate generally outperforms the fixed learning rate. For different initial learning rates, the accuracy and F1-score using a fixed learning rate are lower than those using an adaptive learning rate. Moreover, the fixed learning rate shows significant performance variations with changes in the initial learning rate. In contrast, the adaptive learning rate demonstrates exceptionally stable performance, suggesting that our proposed adaptive learning rate strategy not only enhances performance but also improves robustness, facilitating updates across different individuals.

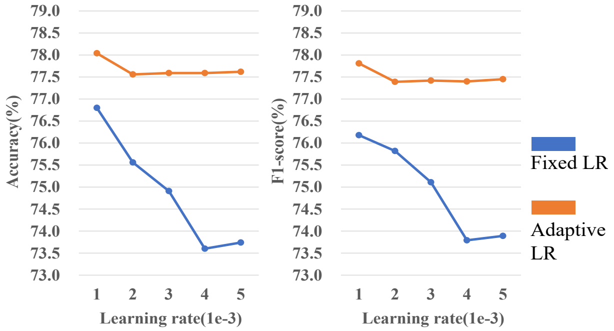

Figure 2: The performance of the model under various learning rate strategies and initial learning rates.

4.6 Hyperparameters Effection

We investigate the impact of two primary hyperparameters, buffer size and retrieval size, on the outcomes. Table 4 illustrates that the proposed ATTA method effectively performs the sleep stage classification task when buffer size $B_s \in \{4, 8, 16, 32\}$ and retrieval size $R_s \in \{2, 4, 6, 8\}$. This indicates that our ATTA method is not sensitive to variations in hyperparameters, demonstrating robust performance.

Parameter	Size	Accuracy(%)	F1-score(%)
Buffer size	4	76.83	76.62
	8	77.21	77.18
	16	78.04	77.81
	32	77.34	77.42
Retrieval size	2	77.53	77.34
	4	77.82	77.56
	8	78.04	77.81
	16	77.76	77.61

Table 4: The impact of different buffer sizes and retrieval sizes on model performance.

5 Conclusion

In this paper, we present a multi-modal test-time adaptation approach equipped with an adaptive learning rate strategy for sleep stage classification. To the best of our knowledge, this work is the first attempt to apply the multi-modal TTA approach to the classification task. Specifically, the intra-modal retained-adaptive module can dynamically adapt to the target domain data during testing, while retaining the knowledge acquired in the source domain to avoid catastrophic forgetting. The inter-modal contribution assessment module is capable of evaluating to what extent different modalities contribute in the process of classification. Meanwhile, the adaptive learning rate strategy can reduce the degradation of classification performance caused by domain shift, while adaptively adjusting the learning rate for different subjects to adapt different domain shift. The experimental result shows that our ATTA method achieves competitive classification performance in the mutual transfer between SleepEDF and SHHS datasets compared to the baseline methods.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Grant No. 62306317) and STI2030-Major Projects (Grant No. 2021ZD0200200) and Postdoctoral Fellowship Program of CPSF (Grant No.

GZC20232992) and China Postdoctoral Science Foundation (Grant No. 2023M733738).

References

[Berry *et al.*, 2012] 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. Rules for scoring respiratory events in sleep: update of the 2007 aasm manual for the scoring of sleep and associated events. *Journal of clinical sleep medicine*, 8(05):597–619, 2012.

[Cai *et al.*, 2021] Xiyang Cai, Ziyu Jia, and Zehui Jiao. Two-stream squeeze-and-excitation network for multi-modal sleep staging. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1262–1265. IEEE, 2021.

[Chambon *et al.*, 2018] Stanislas Chambon, Mathieu N Galtier, Pierrick J Arnal, Gilles Wainrib, and Alexandre Gramfort. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(4):758–769, 2018.

[Chen *et al.*, 2022] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022.

[Jia *et al.*, 2020a] Ziyu Jia, Xiyang Cai, Gaoxing Zheng, Jing Wang, and Youfang Lin. Sleepprintnet: A multivariate multimodal neural network based on physiological time-series for automatic sleep staging. *IEEE Transactions on Artificial Intelligence*, 1(3):248–257, 2020.

[Jia *et al.*, 2020b] Ziyu Jia, Youfang Lin, Jing Wang, Ronghao Zhou, Xiaojun Ning, Yuanlai He, and Yaoshuai Zhao. Graphsleepnet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification. In *IJCAI*, volume 2021, pages 1324–1330, 2020.

[Jia *et al.*, 2021a] Ziyu Jia, Youfang Lin, Jing Wang, Xiaojun Ning, Yuanlai He, Ronghao Zhou, Yuhan Zhou, and H Lehman Li-wei. Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1977–1986, 2021.

[Jia *et al.*, 2021b] Ziyu Jia, Youfang Lin, Jing Wang, Xuehui Wang, Peiyi Xie, and Yingbin Zhang. Salientsleepnet: Multimodal salient wave detection network for sleep staging. *arXiv preprint arXiv:2105.13864*, 2021.

[Jia *et al.*, 2022] Ziyu Jia, Xiyang Cai, and Zehui Jiao. Multi-modal physiological signals based squeeze-and-excitation network with domain adversarial learning for sleep staging. *IEEE Sensors Journal*, 22(4):3464–3471, 2022.

[Kemp *et al.*, 2000] Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

[Khalighi *et al.*, 2016] Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. Isruc-sleep: A comprehensive public dataset for sleep researchers. *Computer Methods and Programs in Biomedicine*, 124:180–192, 2016.

[Li *et al.*, 2023] Siyang Li, Ziwei Wang, Hanbin Luo, Lieyun Ding, and Dongrui Wu. T-time: Test-time information maximization ensemble for plug-and-play bcis. *IEEE Transactions on Biomedical Engineering*, 2023.

[Liang *et al.*, 2023] Heng Liang, Yucheng Liu, Haichao Wang, and Ziyu Jia. Teacher assistant-based knowledge distillation extracting multi-level features on single channel sleep eeg. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI*, pages 3948–3956, 2023.

[Lim *et al.*, 2023] Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. Ttn: A domain-shift aware batch normalization in test-time adaptation. *arXiv preprint arXiv:2302.05155*, 2023.

[Liu and Jia, 2022] Yuchen Liu and Ziyu Jia. Bstt: A bayesian spatial-temporal transformer for sleep staging. In *The Eleventh International Conference on Learning Representations*, 2022.

[Liu *et al.*, 2024] Chenyu Liu, Xinliang Zhou, Zhengri Zhu, Liming Zhai, Ziyu Jia, and Yang Liu. Vbh-gnn: Variational bayesian heterogeneous graph neural networks for cross-subject emotion recognition. In *The Twelfth International Conference on Learning Representations*, 2024.

[Ma *et al.*, 2023] Shuo Ma, Yingwei Zhang, Yiqiang Chen, Tao Xie, Shuchao Song, and Ziyu Jia. Exploring structure incentive domain adversarial learning for generalizable sleep stage classification. *ACM Transactions on Intelligent Systems and Technology*, 2023.

[Mousavi *et al.*, 2019] Sajad Mousavi, Fatemeh Afghah, and U Rajendra Acharya. Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach. *PloS one*, 14(5):e0216456, 2019.

[Niu *et al.*, 2022] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Guanghui Xu, Haokun Li, Peilin Zhao, Junzhou Huang, Yaowei Wang, and Mingkui Tan. Boost test-time performance with closed-loop inference. *arXiv preprint arXiv:2203.10853*, 2022.

[Niu *et al.*, 2023] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*, 2023.

[Perslev *et al.*, 2019] Mathias Perslev, Michael Jensen, Sune Darkner, Poul Jørgen Jennum, and Christian Igel. U-time: A fully convolutional network for time series segmentation applied to sleep staging. *Advances in Neural Information Processing Systems*, 32, 2019.

[Phan *et al.*, 2019] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. Seqsleepnet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging.

IEEE Transactions on Neural Systems and Rehabilitation Engineering, 27(3):400–410, 2019.

[Phan *et al.*, 2020] Huy Phan, Oliver Y Chén, Philipp Koch, Zongqing Lu, Ian McLoughlin, Alfred Mertins, and Maarten De Vos. Towards more accurate automatic sleep staging via deep transfer learning. *IEEE Transactions on Biomedical Engineering*, 68(6):1787–1798, 2020.

[Phan *et al.*, 2021] Huy Phan, Oliver Y Chén, Minh C Tran, Philipp Koch, Alfred Mertins, and Maarten De Vos. Xsleepnet: Multi-view sequential model for automatic sleep staging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5903–5915, 2021.

[Quan *et al.*, 1997] 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. The sleep heart health study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997.

[Shin *et al.*, 2022] Inkyu Shin, Yi-Hsuan Tsai, Bingbing Zhuang, Samuel Schulter, Buyu Liu, Sparsh Garg, In So Kweon, and Kuk-Jin Yoon. Mm-tta: multi-modal test-time adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16928–16937, 2022.

[Sokolovsky *et al.*, 2019] Michael Sokolovsky, Francisco Guerrero, Sarun Paisarnsrisomsuk, Carolina Ruiz, and Sergio A Alvarez. Deep learning for automated feature discovery and classification of sleep stages. *IEEE/ACM transactions on computational biology and bioinformatics*, 17(6):1835–1845, 2019.

[Supratak and Guo, 2020] Akara Supratak and Yike Guo. Tinysleepnet: An efficient deep learning model for sleep stage scoring based on raw single-channel eeg. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 641–644. IEEE, 2020.

[Supratak *et al.*, 2017] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017.

[Wang *et al.*, 2020] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

[Wang *et al.*, 2022] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022.

[Yang *et al.*, 2022] Hongzheng Yang, Cheng Chen, Meirui Jiang, Quande Liu, Jianfeng Cao, Pheng Ann Heng, and Qi Dou. Dltta: Dynamic learning rate for test-time adaptation on cross-domain medical images. *IEEE Transactions on Medical Imaging*, 41(12):3575–3586, 2022.

[Yoo *et al.*, 2021] Chaehwa Yoo, Hyang Woon Lee, and Je-Won Kang. Transferring structured knowledge in unsupervised domain adaptation of a sleep staging network. *IEEE journal of biomedical and health informatics*, 26(3):1273–1284, 2021.

[Zhao *et al.*, 2023] Bowen Zhao, Chen Chen, and Shu-Tao Xia. Delta: degradation-free fully test-time adaptation. *arXiv preprint arXiv:2301.13018*, 2023.

[Zhou *et al.*, 2023] Xinliang Zhou, Chenyu Liu, Liming Zhai, Ziyu Jia, Cuntai Guan, and Yang Liu. Interpretable and robust ai in eeg systems: A survey. *arXiv preprint arXiv:2304.10755*, 2023.