MtRBD: Advancing iRBD Analysis with Multi-Task Learning for Joint Sleep Staging and RSWA Detection

— Supplementary material

S.1 Algorithmic Flowchart

The complete training procedure for our MtRBD model is outlined in Algorithm 1.

Algorithm 1 Multi-task RBD Network (MtRBD)

Require: Training data: $\mathscr{D} = (\mathbf{X}_s, \mathbf{y}_s^s, \mathbf{y}_s^r)$, Network parameters: θ, Weight factors: $\alpha, \beta, \lambda_1, \lambda_2$
Ensure: Optimized network parameters: θ^*

Initialize network parameters θ randomly
for $epoch = 1, 2, \ldots, max_epochs$ **do**
  for batch $(X_{batch}, Y_{batch}^{stage}, Y_{batch}^{RBD})$ sampled from D **do**
    // Feature extraction backbone
    $F_{ms} = \text{MSCNN}(X_{batch})$ {MSCNN with SE layer}
    $F_{MHA} = \text{SelfAttention}(F_{ms})$ {Multi-Headed Attention mechanism}
    $F_{global} = \text{GlobalAvgPool}(F_{MHA})$
    // Task 1: Sleep Staging
    $z_{task1}, L_{KL1} = \text{InformationBottleneck1}(F_{global})$
    $P_{stage} = \text{Softmax}(\text{FC}(z_{task1}))$
    $L_{task1} = \text{CrossEntropy}(P_{stage}, Y_{batch}^s) + \lambda_1 \cdot L_{KL1}$
    // Task 2: RSWA detection
    $F_{combined} = [F_{global}, P_{stage}]$ {Dynamic Feature Enhancement }
    $z_{enhanced} = F_{global} \odot \sigma(W_2 \text{ReLU}(W_1 F_{combined}))$
    $z_{task2}, L_{KL2} = \text{InformationBottleneck2}(z_{enhanced})$
    $P_{RBD} = \text{Softmax}(\text{FC}(z_{task2}))$
    $L_{task2} = \text{CrossEntropy}(P_{RBD}, Y_{batch}^r) + \lambda_2 \cdot L_{KL2}$
    // Update Parameters
    $L_{total} = \alpha \cdot L_{task1} + \beta \cdot L_{task2}$
    $\theta \leftarrow \theta - \text{Adam}(\nabla_\theta L_{total})$
  end for
  Evaluate model on validation set
end for
return θ^*

S.2 Ablation Study

In this section, we conduct an ablation study to evaluate the impact of different components of our model. The study investigates the effects of MIABNet, MtRBD, and MHA on model performance.

S.2.1 MIABNet Ablation

In order to assess the contribution of each module in MIABNet, we performed detailed ablation experiments on the SleepEDF-20 dataset (Table S.2.1).

We tested five configurations: (1) the baseline configuration with only the MSCNN backbone and no additional components; (2) the model without the SE module; (3) the model without the MHA module; (4) the MIABNet without the Information Bottleneck (IB) module; and (5) the

complete MIABNet architecture with all components. Our results show that, while the MSCNN preserves the basic structure, removing SE, MHA, or IB modules leads to significant performance degradation. Each module plays a crucial role in feature extraction, and the absence of any component compromises the model's ability to fully capture and utilize the relevant information, with the IB module being especially critical in constraining the information bottleneck feature.

Table S.2.1: Comparison of MIABNet Model Components for Ablation Study

Mode	Sleep staging		
	ACC	MF1	*k*
Only MSCNN	0.826	0.758	0.762
MIABNet w/o SE	0.850	0.792	0.795
MIABNet w/o MHA	0.841	0.776	0.783
MIABNet w/o IB	0.830	0.759	0.765
MIABNet	**0.853**	**0.792**	**0.798**

S.2.2 MtRBD Ablation

To validate the effectiveness of each component in MtRBD, we conducted comprehensive ablation experiments on the CZ-RBD dataset (Table S.2.2). We tested five configurations: (1) baseline with only MSCNN backbone without additional components, (2) standard CNN with MHA mechanism representing conventional approaches, (3) MtRBD-: MtRBD without Dynamic Feature Enhancement Module (DFEM), (4) MtRBD without Information Bottleneck (IB) modules, and (5) the complete MtRBD architecture with all components. This progressive analysis allowed us to quantify our proposed modules' individual and combined effects on both sleep staging and RSWA detection performance.

TABLE S.2.2: Comparison of MtRBD Model Components for Ablation Study

Model	Sleep Staging			RSWA Detection			
	ACC	MF1	*k*	ACC	MF1	*k*	*F1-2*
Only MSCNN	0.807	0.735	0.722	0.914	0.731	0.690	0.506
CNN+MHA	0.817	0.751	0.738	0.892	0.721	0.654	0.497
MtRBD-	0.818	0.753	0.740	0.921	0.757	0.710	0.560
MtRBD w/o IB	0.823	0.761	0.744	0.926	0.779	0.755	0.656
MtRBD	**0.830**	**0.771**	**0.756**	**0.936**	**0.824**	**0.782**	**0.705**

Results demonstrate that the complete MtRBD model significantly outperformed all partial implementations, with particularly notable improvements in RSWA detection (F1-2: 0.705 versus 0.560 without DFEM). This improvement in RSWA detection is attributable to the DFEM, which incorporates contextual sleep stage information to inform and refine RSWA detection. The integration of sleep architecture cues enables the model to capture subtle behavioral abnormalities that may be overlooked without such context. Furthermore, the Information Bottleneck approach

contributed to additional performance gains by preserving only the most task-relevant features. These results indicate that the comprehensive MtRBD architecture can more effectively detect RSWA and other abnormal activities across diverse sleep patterns.

S.2.3 Influence of MHA Multi-Head Numbers

As the Multi-Head Attention (MHA) mechanism is a fundamental part of our model, it is essential to explore how varying the number of heads impacts its performance. To do so, we keep other parameters constant and experiment with different numbers of heads in the MHA. It is important to ensure that the number of heads is divisible by the feature length d.

For the CZ-RBD dataset, where d = 64, we test configurations with 1, 2, 4, 8, 16, 32, and 64 heads. The results, shown in Fig. S.2.1, include accuracy, MF1 score, and Cohen's kappa coefficient. The model's performance remains relatively stable across different configurations. As we increase the number of heads from 1, 2, and 4 to 8 and 16, a slight performance boost is observed for both sleep staging and RSWA detection. This improvement occurs because more heads allow the model to capture additional meaningful features and feature interactions, especially for RSWA detection. However, when the number of heads increases further (to 32 and 64), the feature length per head decreases, which leads to a small decline in performance. Based on our experiments, we ultimately chose to set the number of heads to 16.

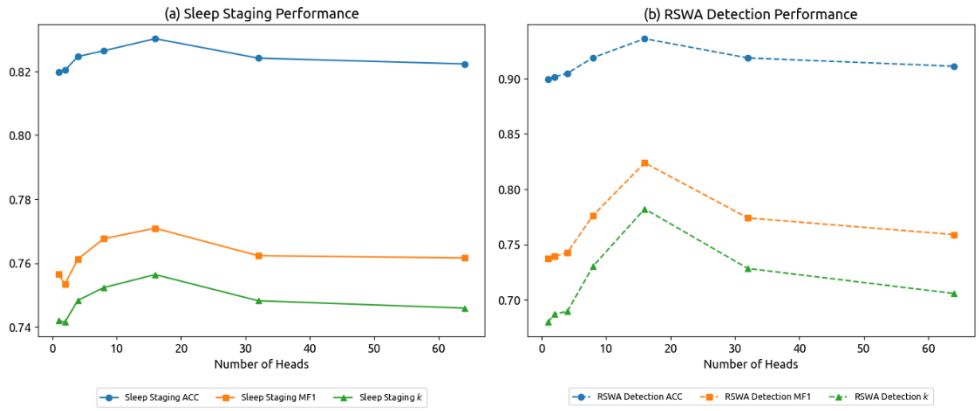

Fig. S.2.1: MtRBD performance for different numbers of heads.

S.3 Hyperparameter Validations

S.3.1 Effect of Multi-Scale MSCNN Convolution Kernels

Previous studies have demonstrated the effectiveness of multi-scale CNNs for sleep staging and related tasks [1][2][3]. Typically, for multi-scale networks, two or three branches are used, with two branches focusing on single-channel signals and three branches preferred for multi-channel, multi-modal data, as in our case. Adding more branches would increase the computational burden of the model.

In our implementation, we specifically chose kernel sizes of k=3, 5, and 7 for the three MSCNN branches [4]. These kernel sizes were selected to capture features at different temporal resolutions: k=3 to capture fine-grained local patterns, k=5 for intermediate contextual information, and k=7 for broader temporal relationships. Each branch consists of three convolutional layers that maintain the same kernel size throughout the pathway to specialize in feature extraction at each scale.

TABLE S.3.1： Multi-task Performance of the MtRBD Model with Different Multi-scale Convolution Kernels on the CZ-RBD Dataset

kernel sizes	**Sleep Staging**			**RSWA Detection**		
	ACC	MF1	*k*	ACC	MF1	*k*
3，5，7	**0.830**	**0.771**	**0.756**	**0.936**	**0.824**	**0.782**
3，50，700	0.815	0.746	0.735	0.893	0.725	0.661
30，50，70	0.824	0.763	0.749	0.903	0.743	0.688
300，500，700	0.792	0.712	0.700	0.874	0.692	0.612

Our convolution kernel ablation experiments (Table S.3.1) demonstrate that using smaller convolution kernels yields better performance in multi-task learning. We attribute this to several key factors:

1. Receptive Field Expansion: Stacking multiple convolutional layers sequentially increases the effective receptive field of the network. This allows smaller kernels to capture long-range temporal dependencies through depth rather than width.

2. Complementary Feature Extraction: The use of three different kernel sizes enables the extraction of complementary features across varying temporal resolutions. Although larger kernels can theoretically capture broader temporal dependencies, they come at the cost of higher computational complexity and increased parameter count.

3. Support from Self-Attention Mechanism: The integrated self-attention mechanism explicitly models the relationships between all time steps, regardless of their distance. This helps compensate for any limitations smaller kernels may have in modeling long-range dependencies.

4. Impact on RSWA Detection: Our experiments reveal that larger kernels do not significantly improve performance in sleep staging and, in fact, degrade accuracy in RSWA detection. Since RSWA detection relies heavily on electromyographic (EMG) signals—which contain high-frequency components—smaller kernels are more effective in capturing these fine-grained features.

5. Suitability for High-Frequency, Local Patterns: Smaller kernels are inherently better at extracting localized and fine-grained features, making them especially well-suited for high-frequency patterns. Given that the diagnosis of REM Sleep Behavior Disorder (RBD) involves analyzing EMG activity during REM sleep (which contains high-frequency components), a multi-scale design centered on smaller kernels is particularly advantageous for this task.

In summary, our choice to prioritize smaller kernels within a multi-scale framework effectively balances task-specific requirements with computational efficiency. This design leads to superior performance in both sleep staging and RSWA detection within the multi-task learning context.

S.3.2 Model Hyperparameter Selection: Effect of SE Reduction Ratios

Regarding the different reduction ratios for the SE block, we referenced highly-cited, open-source sleep staging works, such as AttnSleep, which uses a reduction ratio of 16 [1]. We also performed a parameter search within a reasonable range, and our results demonstrate the effectiveness of this choice.

TABLE S.3.2: Multi-task Performance of the MtRBD Model with Different SE Reduction Ratios on the CZ-RBD Dataset

Reduction ratios	**Sleep Staging**			**RSWA Detection**		
	ACC	MF1	k	ACC	MF1	k
4	0.826	0.765	0.750	0.908	0.748	0.699
8	0.827	0.771	0.752	0.921	0.786	0.739
16	**0.830**	**0.771**	**0.756**	**0.936**	**0.824**	**0.782**
32	0.824	0.758	0.747	0.922	0.782	0.739
64	0.817	0.748	0.736	0.901	0.741	0.685

These results demonstrate that smaller kernels and the reduction ratio of 16 for the SE block are optimal for our tasks, which is further confirmed through our ablation studies in the appendix.

S.3.3 Effect of the Number of Attention Heads: Complete Results

As the Multi-Head Attention (MHA) mechanism is a critical component of our model architecture, we conducted a detailed analysis to investigate how varying the number of attention heads affects model performance. To ensure a controlled comparison, we fixed all other hyperparameters and varied only the number of heads. Since the feature dimension $d = 64$, the number of heads was selected such that it evenly divides d. We tested configurations with 1, 2, 4, 8, 16, 32, and 64 heads.

Results on SleepEDF-20 (MIABNet)

Table S.3.3 reports the performance of backbone model MIABNet on the SleepEDF-20 dataset for different head configurations. The results show that the overall performance of sleep staging remains stable across different numbers of heads. A slight improvement is observed when increasing the number of heads from 1 to 16. Specifically, the accuracy (ACC) and macro F1 score (MF1) reach a peak at 8–16 heads (ACC = 0.853, MF1 = 0.792), suggesting that moderate head sizes allow the model to better extract and aggregate multi-scale temporal features. However, further increasing the number of heads to 32 and 64 results in a minor performance drop, likely due to reduced feature dimensionality per head (i.e., each head receives fewer features, making learning less effective).

TABLE S.3.3：MIABNet performance for different numbers of heads

Number	Sleep staging		
	ACC	MF1	*k*
1	0.847	0.787	0.789
2	0.848	0.788	0.799
4	0.850	0.787	0.793
8	0.852	0.792	0.798
16	**0.853**	**0.792**	**0.798**
32	0.850	0.794	0.793
64	0.847	0.787	0.789

Results on CZ-RBD (MtRBD)

Table S.3.4 presents a similar trend on the CZ-RBD dataset using the MtRBD backbone. For the sleep staging task, performance improves steadily up to 16 heads (ACC = 0.830, MF1 = 0.771, κ = 0.756), after which it plateaus or slightly declines. The RSWA detection task, which heavily relies on fine-grained EMG features, benefits more significantly from increased attention heads. A notable improvement is observed as the number of heads increases from 1 to 16 (ACC improves from 0.899 to 0.936, and MF1 from 0.737 to 0.824). However, similar to sleep staging, performance slightly deteriorates beyond 16 heads due to diminished per-head representation capacity.

These results suggest that increasing the number of attention heads enhances the model's ability to capture diverse temporal dependencies and feature interactions—especially for tasks requiring fine-grained signal discrimination like RSWA detection. However, excessive partitioning of features among too many heads can lead to loss of discriminative power. Based on a balance between performance and computational efficiency, we selected 16 heads as the optimal setting, which is also used across other datasets. Additional sensitivity analyses can be found in the supplementary materials.

TABLE S.3.4：MtRBD performance for different numbers of heads

Number	Sleep staging			RSWA Detection		
	ACC	MF1	*k*	ACC	MF1	*k*
1	0.820	0.757	0.742	0.899	0.737	0.680
2	0.820	0.753	0.742	0.901	0.739	0.687
4	0.825	0.761	0.748	0.905	0.743	0.690
8	0.826	0.768	0.752	0.919	0.776	0.731
16	**0.830**	**0.771**	**0.756**	**0.936**	**0.824**	**0.782**
32	0.824	0.762	0.748	0.919	0.774	0.728
64	0.822	0.762	0.746	0.911	0.759	0.706

S.4 Additional Dataset Validation

S.4.1 Dataset Description

MASS-SS3 [8]: The MASS-SS3 dataset contains PSG signals from 62 healthy subjects (sampling rate 256 Hz), recorded using 20 EEG, three EMG, two EOG, and one ECG channel. Sleep stages were annotated by experts according to the AASM standard into five stages: Wake, REM, N1, N2, and N3.

The preprocessing follows the data set processing method in the main text, and selects three modal signals: EEG, EOG, and EMG. Experiments on the MASS-SS3 dataset were conducted using 21-fold cross-validation, with the first 20 folds containing data from three subjects each, and the last fold containing data from two subjects. We follow previous research and adopt a channel configuration consisting of the C4-A1 EEG, an average EOG (ROC–LOC), and an average EMG (CHIN1–CHIN2) in our experiments [6].

S.4.2 Experimental Results on the Dataset

TABLE S.4.1: Sleep Stage Classification Performance of the Proposed Model on the MASS-S3 Dataset

Sleep Stage	Predicted					Per-class metrics		
	W	N1	N2	N3	REM	Precision	Recall	F1
W	5836	342	106	10	120	0.88	0.91	0.90
N1	512	2326	1204	4	790	0.67	0.48	0.56
N2	149	521	27360	1072	693	0.89	0.92	0.90
N3	19	0	1565	6069	0	0.85	0.79	0.82
REM	106	289	453	2	9727	0.86	0.92	0.89

Table S.4.1 presents the detailed confusion matrix and per-class performance metrics of the proposed model on the MASS-S3 dataset. The model achieves high precision and recall across most stages, particularly for stages W, N2, and REM, where the F1-scores are 0.90, 0.90, and 0.89, respectively. Stage N1, as is commonly observed in previous studies, remains the most challenging due to its transitional nature between wake and non-REM stages, resulting in a lower F1-score of 0.56. Nonetheless, the overall performance across all stages demonstrates the model's robustness in handling the imbalanced class distribution and complex stage transitions in MASS-S3.

S.4.3 Comparison with Baseline Network

TABLE S.4.2: Comparison of MIABNet with State-of-the-Art Models on the MASS-S3 Dataset

for Sleep Stage Classification

Model	Sleep Staging							
	ACC	MF1	*k*	F1-W	F1-N1	F1-N2	F1-N3	F1-REM
DeepSleepNet [5]	0.851	0.779	0.800	0.823	0.482	0.896	0.820	0.879
AttnSleep [1]	0.852	0.780	0.801	0.860	0.556	0.872	0.900	0.795
SeqSleepNet [6]	**0.871**	**0.833**	**0.815**	-	-	-	-	-
MIABNet	0.866	0.813	0.801	0.895	0.560	0.905	0.820	0.888

Table S.4.2 compares MIABNet with several state-of-the-art (SOTA) models on the MASS-S3 dataset, including DeepSleepNet, AttnSleep, and SeqSleepNet. MIABNet achieves a strong overall performance with an accuracy of 0.866, a macro F1-score (MF1) of 0.813, and a Cohen's kappa of 0.801. It shows particularly high F1-scores for stages W, N2, and REM, indicating robust classification of both wakefulness and deep sleep stages.

Although SeqSleepNet achieves the highest overall accuracy (0.871) and macro F1, it requires a more complex preprocessing pipeline, including time–frequency transformation of physiological signals. This additional preprocessing increases the computational burden and implementation complexity, potentially limiting its practicality in real-time or resource-constrained scenarios.

In contrast, MIABNet operates directly on raw time-series data, avoiding the need for costly spectral transformations while still delivering competitive performance. This highlights the practicality and efficiency of MIABNet for real-world sleep staging applications.

References:

[1] Eldele, E., Chen, Z., Liu, C., Wu, M., Kwoh, C. K., Li, X., & Guan, C. (2021). An attention-based deep learning approach for sleep stage classification with single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *29*, 809-818.

[2] Zhou, W., Shen, N., Zhou, L., Liu, M., Zhang, Y., Fu, C., ... & Chen, C. (2024). PSEENet: A pseudo-siamese neural network incorporating electroencephalography and electrooculography characteristics for heterogeneous sleep staging. IEEE Journal of Biomedical and Health Informatics, 28(9), 5189-5200.

[3] Zhu, H., Zhou, W., Fu, C., Wu, Y., Shen, N., Shu, F., ... & Chen, C. (2023). MaskSleepNet: A cross-modality adaptation neural network for heterogeneous signals processing in sleep staging. *IEEE Journal of Biomedical and Health Informatics*, *27*(5), 2353-2364.

[4] Chen, Y., Lv, Y., Sun, X., Poluektov, M., Zhang, Y., & Penzel, T. (2024). ESSN: An Efficient Sleep Sequence Network for Automatic Sleep Staging. *IEEE Journal of Biomedical and Health Informatics*.

[5] Supratak, A., Dong, H., Wu, C., & Guo, Y. (2017). 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.

[6] Phan, H., Andreotti, F., Cooray, N., Chén, O. Y., & De Vos, M. (2019). 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.