Modification Description

First of all, I would like to express my sincere thanks for your comments. I will make improvements according to your suggestions.

1. **Please correct the term “non-European space” in line 50.**

Reply: Correction: non-Euclidean space

2. **Please add a paragraph describing the manuscript’s structure at the end of the Introduction section.**

Reply: The paper structure is as follows: Section I briefly introduces the recent work on sleep staging and its shortcomings, leading to the study and discussion of the proposed model. Section II details the preliminary work, network structure of the proposed model. Section III talks about the overall performance, training duration and other related experiments of proposed model. Section IV discusses the above experiments, and section V concludes this paper.

3. **In the literature review, please briefly state the advantages of your proposed and improved approach based on the MGANet over the recent studies utilizing conventional CNN architectures with the 2D time-frequency signal representations. Please provide some examples for these alternative approaches in various applications, as, in my opinion, your proposed method is also not limited only to application to EEG signals but could be extended to other applications in future work. Please consider briefly mentioning here the following papers for illustration purposes: 10.1007/s10044-020-00921-5, 10.1109/ACCESS.2021.3139850, 10.1109/TNNLS.2020.3008938.**

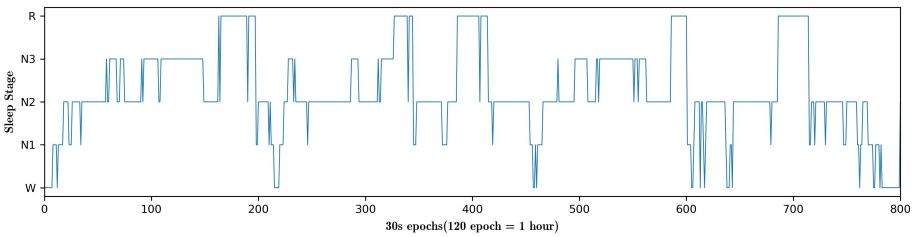
Reply: Add 2 paragraphs in the literature review as follows:

Generally speaking, there are three types of information fusion in EEG signals processing: spatial information fusion, temporal information fusion, spatial and temporal information fusion. It not only fuses the relationship dynamics of EEG signals in different brain regions such as correlation, but also calculates time-frequency domain features at different times through fusion. The traditional CNN architecture and 2D time-frequency signal processing method can well process the temporal information fusion of EEG signals. However, the fixed convolution kernel cannot maintain the translation invariance of non Euclidean distance signals, nor can it directly capture all nodes’ information, so it cannot fuse and process the rich spatial information of EEG.

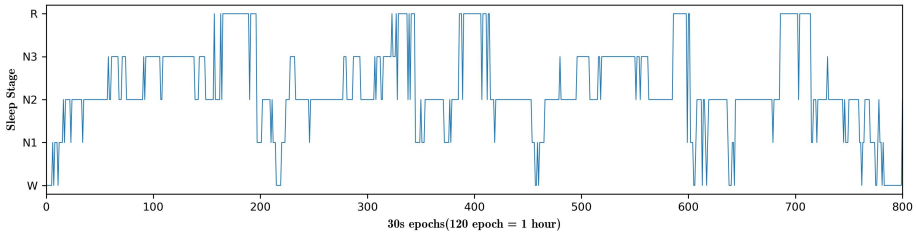
The MGANet proposed in this paper provides a general framework for representing and fusing deep semantic features from input EEG data, and could be extended in emotion recognition, speech processing, non-stationary signal analysis and other fields in future work [34, 35, 36].

4. **Please improve the resolution of Figure 6.**

Reply: The resolution of Figure 6 is increased.



(a) Hypnogram manually scored by sleep experts.



(b) Hypnogram automatically scored by the proposed MGANet.

Figure 6. The scored hypnogram by sleep experts and proposed MGANet.

5. Please provide a more detailed discussion of the results presented in the form of graphs and tables.

Reply: Each section adds discussion as follows:

4.1. Classification performance

Table 1 shows the performance comparison of several baseline results. The proposed model has the best performance on the 2 data sets The Accuracy, MF1 and Kappa of the model were 0.825, 0.814 and 0.775 on the ISRUC dataset, and 0.873, 0.801 and 0.827 on the SHHS dataset, respectively.

4.2. Number of parameters and training duration

As can be seen from Table 3, the number of parameters of the proposed model is significantly less than that of other models, and the training duration of this model is also better than the others.Although the model establishes a complex graph channel structure, the parameter number of the proposed model is 1.5×10^5, which is still on the same order of magnitude as other state-of-the-art methods. And the number of parameters affects the complexity and robustness of the model, the experiments prove that the proposed model achieves better results under the same parameter magnitude.

4.3. Hypnograms

Figure 6 shows that the MGANet proposed on 800 epochs can correctly classify most sleep stages and has a high accuracy in each transition stage. In the first 20 epochs, the model mistakenly confused the N1 and W phases, which is consistent with the human sleep pattern, because humans usually cycle repeatedly from waking to light sleep stages during this period. Around the 330 epoch, the model confuses REM with N2, a period of human dreaminess. Some wrong results in this model are understandable because of the confusing nature of the transition stage, which is characterized by multiple sleep stages. Actually, sleep experts also get it wrong even when manually scoring. Most of the results

are misjudgment of N2 as N1 or REM as N2, which are also widespread in other baseline models.

4.4. Ablation experiments results

Figure 7 and Table 4 shows that the performance of the model using baseline to extract the non-Euclidean spatial features of sleep is poor, and the model accuracy is only 80.9%. After RNN is added, the model can extract time-dependent features well, and the model performance is slightly improved to 81.5%. However, when spatial attention is added to the baseline, the model performance deteriorates compared to the baseline. May the first is the cause of the lack of time domain signal feature extraction, the second is the model of network layer number is less, attention is not obvious. Because the sleep transition stage confuses the characteristics of multiple stages, it brings difficulties to sleep staging. After adding TSE, the performance of the model was significantly improved, which proved its help in accurately classifying confusion stages. Finally, it is proved that the proposed MGANet adds variables A, B and C simultaneously, and the obtained model has the best performance, which accuracy, F1-score and kappa is 82.5, 81.4 and 77.5 respectively. RNN and transition stage estimator methods are used to achieve suboptimal results.

4.5. Heads numbers

In Figure 8, the head numbers of MGANet represents the standard for distinguishing features of center nodes. Target features are extracted from multiple angles to avoid feature redundancy or single feature. As shown in Figure 8, the model achieves best performance when h=8 is used. When the number of attention heads is set to be small, like h =2 or h =4, the acc of the model is only 80.3% and 80.7%, respectively. Due to the lack of aggregation features in each head, resulting in ineffective multi-attention mechanism.

4.6. Speed training

As shown in Figure 9, the loss of the proposed model decreases the fastest among the first 10 epochs, reaching the lowest loss compared with the other two methods.

6. **In the Conclusions section, please address some limitations of the presented study.**
7. **In the Conclusions section, please also provide some directions for future work.**

Reply: I will integrate the above two comments and added a paragraph after the conclusion as follows:

At present, most EEG research is carried out in a task driven manner, which would be largely limited by data labels to a great extent. It is difficult and expensive to obtain labeled dataset, hence the model proposed in this paper cannot be well generalized and is not suitable for clinical diagnosis at present. In the next work, we will consider to design a semi-supervised model or an unsupervised model, and will use the easily obtained unlabeled data to classify sleep stages to improve the generalization ability of the model.

Finally, I would like to express my sincere thanks again. Your advice is valuable and constructive. Wish you a wonderful day!