Multi-Layer Graph Attention Network for Sleep Stage Classification Based on EEG

Qi Wang, Yecai Guo *, Yuhui Shen, Shuang Tong and Hongcan Guo

School of Electronics and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China
* Correspondence: guoyecai@nuist.edu.cn; Tel.: +86-151-8980-2968

Abstract: Graph neural networks have been successfully applied to sleep stage classification, but there are still challenges: (1) How to effectively utilize epoch information of EEG-adjacent channels owing to their different interaction effects. (2) How to extract the most representative features according to confused transitional information in confused stages. (3) How to improve classification accuracy of sleep stages compared with existing models. To address these shortcomings, we propose a multi-layer graph attention network (MGANet). Node-level attention prompts the graph attention convolution and GRU to focus on and differentiate the interaction between channels in the time-frequency domain and the spatial domain, respectively. The multi-head spatial-temporal mechanism balances the channel weights and dynamically adjusts channel features, and a multi-layer graph attention network accurately expresses the spatial sleep information. Moreover, stage-level attention is applied to easily confused sleep stages, which effectively improves the limitations of a graph convolutional network in large-scale graph sleep stages. The experimental results demonstrated classification accuracy; MF1 and Kappa reached 0.825, 0.814, and 0.775 and 0.873, 0.801, and 0.827 for the ISRUC and SHHS datasets, respectively, which showed that MGANet outperformed the state-of-the-art baselines.

Keywords: graph attention network; gated recurrent unit; node-level and stage-level attention; sleep staging; transitional stage estimator

1. Introduction

Sleeping is a crucial condition to maintain normal cognitive function. Long-term sleep deprivation causes normal cognitive diseases, such as learning and memory disorders [1,2]. Different sleep stages are complex and dynamic biological processes that periodically repeat. Polysomnography (PSG) contains physiological signals such as electroencephalogram (EEG) and electroophthalmogram (EOG) [3]. In a clinical setting, the transition rules between sleep stages assist experts to accurately classify sleep stages within frameworks determined by the American Academy of Sleep Medicine (AASM) and the Rechtschaffen and Kales (R&K) standards [4,5], which divide human sleep stages into five classes: wake (W), rapid eye movement (REM), non-REM1 (N1), non-REM2 (N2), and non-REM3 (N3) [6,7].

Automatic sleep staging has wide practical value in clinical medicine. Researchers have used traditional machine learning algorithms based on time and frequency domain features, random forest (RF) [8], and support vector machines (SVM) [9] for sleep stage classification. In recent years, with the development of deep learning, it has been widely used to classify sleep stages with its powerful representational ability. [10–13]. However, there are some disadvantages to the realization of sleep staging using convolutional neural networks [14]. Generally speaking, there are three types of information fusion in EEG signals processing: spatial information fusion, temporal information fusion, and spatial and temporal information fusion [15]. 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. Traditional CNN architecture and 2D
time-frequency signal processing methods could well process the temporal information fusion of EEG signals. Unfortunately, the fixed convolution kernel cannot maintain the translation invariance of non-Euclidean distance signals or directly capture effective node information. Therefore, it cannot fuse and process rich spatial information. Jia et al. [16] considered using the graph structure to establish the functional connectivity and physical proximity of EEG brain regions.

Furthermore, brain regions are located in non-Euclidean space, and different brain regions control various functions of the human body [17]. For example, the brainstem has the function of maintaining circulation and respiration, and the telencephalon dominates the activities of the higher cortex, such as smell, hearing, and other systemic sensations, as well as body movement. Research has determined that the whole of the human brain function structure is strongly influenced by regional anatomical relationships, and the degree of interaction between regions varies [18]. When updating the current node representation of GCN, all adjacent nodes in the brain region will be randomly assigned static important weights. As a result, neighboring nodes closely related to the current node functions are randomly assigned larger or smaller weights, causing the importance of features of node functions to be incorrectly amplified or ignored [19].

Inspired by these recent works, we propose a multi-layer graph attention sleep stage classification network (MGANet) to distinguish sleep stages in the graph structure; the framework is shown in Figure 1. During training, the influence weights of neighbor channels in the EEG spatial domain and temporal domain are calculated, and the constructed graph structure can reflect the different interaction of functional connectivity between brain regions. Multilayer GATs use stacked layered space-time characteristics to capture information, and reduce the graph structure model parameters. This model successfully applies node-level attention and stage-level attention to sleep classification, dynamically establishing the interaction between channels, and combining RNN and GAT to extract the most representational spatial and temporal features.

![Figure 1. Overall architecture of the MGANet method.](image)

In general, the sleep staging model based on a multi-layer graph attention network has the following characteristics:

1. A node-level attention mechanism is applied to effectively calculate and dynamically update the importance of adjacent channels to the current channel, while considering the spatial location and functional connections of brain regions. (2) The parallel operation of multi-layer graph attention convolution of MGANet improves the computational efficiency of the model. When calculating output features and executing attention operations, all nodes and adjacent edges run in parallel. (3) A transition stage estimator is constructed to accurately classify sleep stages by expressing the features of the transition stage epoch through the stage-level attention vector. Experimental results show that MGANet achieves state-of-the-art performance in sleep staging. The remaining sections of this paper are organized as follows. Section 2 introduces the related work. Section 3 describes the materials and our method. Section 4 outlines the results and discussions, and Section 5 summarizes future work.
2. Related Work

In early studies, manual sleep stage classification by human sleep experts was tedious and time-consuming, with highly subjective classification results [20,21]. The application of traditional sleep staging methods includes the use of machine learning, such as RF and SVM. Tsinalis et al. [13] used the method of manual extraction of sleep spectrum features, and used stacked sparse autoencoders for integrated learning. Such methods require manual features extraction and a large amount of prior knowledge, and are gradually replaced by deep learning algorithms that automatically extract features.

Subsequently, a single-channel sleep staging model with the bidirectional long short term memory (BiLSTM) network was designed by Supratak et al. [10]. Later, Chambon et al. [11] used a hybrid neural network to extract time-invariant features and time-contextual features. Phan et al. [22] proposed a multi-variable and multi-modal algorithm, which automatically extracted sleep features by using multi-modal physiological signals and auxiliary classification of sleep rules. Chriskos et al. [23] processed sleep stage staging tasks as sequence-to-sequence tasks, and Jia et al. [24] established a sleep staging framework based on CNN to extract synchronous features of cerebral cortex interaction. Eldele et al. [25] proposed a model which consisted of feature extraction modules based on a multi-resolution convolutional neural network (MRCNN) and adaptive feature recalibration (AFR). A method of utilizing deep residual networks in mixed queue environments was proposed by Alexander et al. [26].

In addition, graph convolutional networks inherited from CNN are mainly divided into graph convolutional networks based on spectral structure and graph convolutional networks based on spatial structure [19]. Jia et al. [16] used graph structure to represent brain spatial connections for the first time, which adaptively learns the internal connections between different channels represented by an adjacency matrix. Subsequently, they proposed a multi-view spatial-temporal graph convolutional network with domain generalization, which constructed two brain views by utilizing functional connectivity and the physical proximity of brain regions [27].

In this paper, we propose a new general framework, which represents and fuses the deep semantic features of the input EEG data, and effectively learns the significant features and temporal dependence across epochs. The experimental results show that MGANet is superior to the most advanced baseline. In future work, it could be expanded to emotion recognition, speech processing, non-stationary signal analysis, and other fields [28–30].

3. Materials and Methods

MGANet constructs a new graph structure to describe the brain’s spatial position and temporal contextual information, the architecture of which is shown in Figure 1. We introduce the networks used to construct sleep graph attention networks, and directly outline their theoretical and practical benefits, as well as limitations, compared to previous work.

3.1. Preliminaries

Sleep stage. Epochs were divided into five classes after 30 s.

Graph attentional sleep network. This is defined as an undirected graph \( G_A = (N, C, A) \), where \( N, C, A \) denotes vertex, adjacent edge, and the adjacency matrix in GATs, respectively, mapping the brain as an electrode channel, channel connection, and adjacent relationship between channels, respectively. The EEG channel numbers are defined as \( |N| = n \), and the input sequence is defined as \( S_{m+1,N} = \{N_c, N_{c+1,1}, N_{c+1,2}, \cdots, N_{c+1,n}\} \), where \( N_{c+1,m+1,n} \) represents the center node \( N_c \)'s spatial features of the \( n^{th} \) channel within the \( m+1 \) epoch. The node matrix is updated as \( \hat{f}_S = (\hat{f}_{1,1}, \hat{f}_{2,1}, \cdots, \hat{f}_{m,1})^T \) according to graph attention convolution.
Recurrent neural sleep network. The RNN’s sleep time domain feature sequence is defined as \( \vec{S}_{t+1,N} = \{ i_{t+1,1}^N, i_{t+1,2}^N, \ldots, i_{t+1,n}^N \} \). \( i_{t+1,n}^N \) represents \( N_2 \)'s temporal features of the \( n \)th channel at \( t+1 \) point. The time-frequency features output are finally updated as \( \vec{f}_T = (\vec{f}_1', \vec{f}_2', \ldots, \vec{f}_t') \).

3.2. Spatial Node-Level Graph Attention Sleep Network

3.2.1. Multi-Head Graph Attention Learning

Firstly, we construct the human brain as a graph, with electrodes corresponding to the nodes of the graph and the connections between channels corresponding to the edges of the graph. In addition, the sleep features of each node are taken as its dynamic features, and appropriate weights are assigned to adjacent nodes, corresponding to the influence level of neighbors on the current channel.

In order to adaptively capture spatial location information through brain position nodes, the input is a node features sequence \( \vec{f}_S = (\vec{f}_1', \vec{f}_2', \ldots, \vec{f}_m') \). MGANet uses parallel attention operation to assign importance weightings to different channels in the same brain region according to the influence of the interaction between nodes, using the spatial-temporal features of adjacent channels. Specifically, a fixed number of neighborhood edges are selected around the vertex of each electrode, and the learnable weight coefficient matrix \( W_G \) of sleep space features is obtained. The attention coefficient \( u \), i.e., the importance of neighbor nodes \( N_k, k = 1, 2, \ldots, m \) to the center node \( N_c \), which is calculated by the attention factor \( a \) and \( W_G \) through the shared attention mechanism, and \( m \) is the first-order adjacent channel of \( N_c \).

\[
\alpha_{N_c, N_k} = \frac{\exp(\vec{a}^T [W_{N_c} \vec{f}_{N_k} || W_{N_k} \vec{f}_{N_c}])}{\sum_{l \in n} \exp(\vec{a}^T [W_{N_c} \vec{f}_{N_k} || W_{N_k} \vec{f}_{N_c}])}
\]

To make the coefficients across different nodes easily comparable, softmax normalization is used to leverage masked self-attention vector \( a \) of \( N_c \) and its neighborhoods’ \( N_k \) can be obtained through softmax normalization operation, as follows:

\[
\alpha_{N_c, N_k} = \text{softmax}_{N_k}(\mu_{N_c, N_k}) = \frac{\exp(\mu_{N_c, N_k})}{\sum_{l \in n} \exp(\mu_{N_c, N_l})}
\]

The LeakyReLU nonlinear function with negative slope \( \beta = 0.2 \) is applied to activate the attention vector \( \alpha_{N_c, N_k} \) as follows:

\[
\alpha_{N_c, N_k} = \frac{\exp(\text{LeakyReLU}(\alpha_{N_c, N_k}^T [W_{N_k} \vec{f}_{N_k} || W_{N_c} \vec{f}_{N_k}]))}{\sum_{l \in n} \exp(\text{LeakyReLU}(\alpha_{N_c, N_k}^T [W_{N_k} \vec{f}_{N_k} || W_{N_c} \vec{f}_{N_k}]))}
\]

where \( \vec{a}^T \) is the attention mechanism network transpose operation, \( W_{f_{N_k}} || W_{f_{N_c}} \) is the feature aggregation to be given weighting. In order to allocate the importance of different sleep features contained in each channel to adjacent nodes, the multi-node attention mechanism is introduced, and the weighted sum of multiple features contained in nodes \( N_c \) is then averaged, and the final spatial features \( \vec{f}_s \) are obtained after \( L \)th attention. The calculation process is shown as follows:
where $\sigma$ is the sigmiod activation function, and $a^{(l)}_{N_k,N_{k'}}$ is the $l$th attention coefficient after normalization, and $W^{(l)}$ is the attention weight matrix.

MGANet uses the fixed symmetric normalized Laplacian operator widely used in graph convolution as the propagation mode to dynamically learn the spatial structure through the adjacency matrix of the current channel, and introduces a multi-head attention mechanism to match and update the features of adjacent nodes of each node that has significant influence on itself. The model framework is shown in Figure 2.

\[ \vec{f}_s = \sigma \left( \frac{1}{L} \sum_{l=1}^{L} \sum_{N_k \in m} a^{(l)}_{N_k,N_{k'}} W^{(l)} \vec{f}_{N_k} \right) \] (4)

Figure 2. Multi-head graph attention learning structure.

3.2.2. Spatial Domain Attention

The spatial attention mechanism is introduced into MGANet to automatically update the attention coefficient to extract the dynamic spatial valuable information of signals, and its definition is as follows:

\[ A_S = P_s \cdot \sigma((f^{(l-1)}Q_1)Q_2(Q_3f^{(l-1)})^T + Q_s) \] (5)

where $P_s, Q_s, Q_1, Q_2, Q_3$ is an adjustable parameter, and the attention matrix is softmax normalized, as follows:

\[ A'_S(i,j) = \text{softmax}(A_S(i,j)) \] (6)

where $A_S(i,j)$ represents the correlation degree of sleep nodes $i$ and $j$ in the spatial domain.

After the spatial attention mechanism, the spatial features $\vec{f}_s$ are updated as follows:

\[ \vec{f}'_s = [(\vec{f}_1, \vec{f}_2, \cdots, \vec{f}_n) = (F_1, F_2, \cdots, F_n)A'_S] \] (7)

3.3. Temporal Node-Level Graph Attention Sleep Network

EEG is a typical distraction time signal with intense time dependence between adjacent sleep stages. Combining the gated recurrent unit (GRU) to MGANet is used to capture temporal dependent features and spectral information of the current node, and the attention mechanism is used to capture the dynamic time information between sleep stages.
3.3.1. Time-Frequency Feature Extraction

RNN represents the input epoch sequence as $S_{t+1} = \{i_{t+1,1}, i_{t+1,2}, \ldots, i_{t+1,n}\}$, recursively combining the time-domain features extracted at the previous moment with the current moment, thus dynamically updating the output data, that is $\tilde{f}_t = r(i_{t+1,n}, \tilde{f}_{t-1})$. Time-frequency domain features of the current node and nonlinear function combine two kinds of information sources, which are represented by $\tilde{f}_t$ and $r(\cdot, \cdot)$. The GRU update procedure is as follows:

$$
\begin{align*}
    u_t &= \sigma(W^{(u)}[\tilde{f}_t, i_{t+1,n}^N] + U^{(u)}[h_{t-1}, i_{t+1,n}^N] + b^{(u)}) \\
    d_t &= \sigma(W^{(d)}[\tilde{f}_t, i_{t+1,n}^N] + U^{(d)}[h_{t-1}, i_{t+1,n}^N] + b^{(d)}) \\
    \hat{h}_t &= \phi(W^{(h)}[\tilde{f}_t, i_{t+1,n}^N] + U^{(h)}(d_t \odot h_{t-1}) + b^{(h)}) \\
    h_t &= u_t \odot h_{t-1} + (1 - u_t) \odot \hat{h}_t
\end{align*}
$$

(8)

Input sleep sequence $S_{t+1}$, output updated sleep features $x_t$ through GRU update gate, $W^{(\cdot)}$ and $U^{(\cdot)}$ are the weight matrix of input and of the current moment $\tilde{f}_t$ and the previous moment $\tilde{f}_{t-1}$, respectively, with $b^{(\cdot)}$ representing the bias in the model. At the same time, the reset threshold vector $d_t$ is reset to the transition vector $\tilde{f}_t$ representing the hidden state vector of sleep as the intermediary, and the time-frequency features $\tilde{f}_t$ of the final output are obtained after nonlinear operations, such as sigmoid $\sigma(\cdot)$ and tanh $\phi(\cdot)$. The GRU network unit structure is shown in Figure 3.

![Figure 3. Gated recurrent unit network.](image)

3.3.2. Temporal Domain Attention

To better extract the temporal and frequency domain sleep features, the temporal attention mechanism is added to the RNN temporal sleep network, as follows:

$$
A_T = P_0 \cdot \sigma((f^{(r-1)}_{t})^T P_1) P_2 (P_3 f^{(r-1)}_{t}) + Q_0)
$$

(9)

where $P_0, Q_0, P_1, P_2, P_3$ are the adjustable parameters of the model, and $f^{(r-1)}_{t}$ is the input feature vector after the convolution of $r$th layer. The final time-domain feature expression is obtained after softmax normalization, as follows:

$$
A_T(i, j) = \text{softmax}(A_T(i, j))
$$

(10)
where $A_T(i, j)$ represents the time correlation degree of sleep nodes $i$ and $j$. After the time attention mechanism, the model updates to $f_t^A$ which is as follows:

$$f_t^A = (f_1, f_2, \ldots, f_{T-1}) = (f_1, f_2, \ldots, f_{T-1}) A_T$$  \hspace{1cm} (11)

In this paper, we present the calculation of the MGANet network layer in pseudo-code referring to Appendix A.3.

### 3.4. Stage-Level Attention Transitional Estimator

The immediate task by the transitional stage estimator (TSE) mainly comprises two parts: spatial graph features and temporal contextual relationships. It is challenging to classify transitional epochs because of the easily confused features of multiple stages. Therefore, we extracted the sleep features of the transition stage through EEG-learnable hidden relationships, and explicitly proposed an attention-guided transitional stage estimator.

TSE firstly aggregates spatial features and temporal features as follows:

$$\overrightarrow{f} = \overrightarrow{f}_s || \overrightarrow{f}_t$$  \hspace{1cm} (12)

Then, the attention vector is represented as the features of a sleep transition stage at each epoch, and uses the softmax function to calculate the sleep stage probability distribution $P_f$ as follows:

$$P_f = \text{softmax}(F_G(f))$$  \hspace{1cm} (13)

where $F_G$ indicates the full connection layer used for decision-making and $G$ indicates the global average pooling operation. $P_f$ represents the confidence of each sleep stage, and the class probability jointly carries the unknown stage information. As the potential feature information, the probability of the sleep stage is calculated by a series of linear or non-linear functions, such as sigmoid, to represent its attention vector as follows:

$$A_f = \text{sigmoid}(F_C(P_f))$$  \hspace{1cm} (14)

Finally, class probabilities and their attention vectors are clearly represented by the Hadamard product multiplication operation as follows:

$$\overrightarrow{f}_s = A_f \otimes P_f$$  \hspace{1cm} (15)

The obtained $\overrightarrow{f}_s$ contains epoch features and transitional fusion information. The TSE’s network structure is shown in Figure 4.

![Figure 4](image-url)

**Figure 4.** TSE consists of global average pooling, full connection, softmax, and sigmoid.
3.5. Spatial Multi-Layer Graph Attention Convolution

We extend the definition of spatial graph attention convolution to nodes with multiple channels to effectively solve the problem of too large or too small weighting, caused by uneven importance distribution of adjacent nodes. After multi-head graph attention sleep feature learning, MGANet extracts spatial features \( f_s \) related to sleep; the spatial multi-layer graph architecture is shown in Figure 5.

![Figure 5. Spatial multi-layer graph attention convolution.](image)

We introduce a multi-layer graph attention convolution layer by layer using a stacked method to capture features. After extracting the center node’s features \( f_c \) and the neighbors’ features \( f_k \) in the \( l \)th layer, the input updates to \( f_c \) is \( \text{ReLU}(W^{(l)}f_s) \) and the \( (l+1) \)th layer is obtained through ReLU activation. Where \( W^{(l)} \) is the shared learnable weight matrix of \( l \)th layers. We obtained the final spatial convolution feature after the L attention layer by matrix stacking. Using a multi-layer network can improve the efficiency of model operations and reduce model parameters. The center node \( f_c \) of the layer \( l \) has \( m \) sleep features, which are aggregated with the attention vector \( \alpha_{ij} \) for \( (n-1) \) times, combined with the weight matrix \( W_l \) to finally obtain the center node \( f_c^{(n)} \) of layer \( n \) and its neighbor features \( f_1^{(n)} \), and the sleep features are extracted layer by layer using the stacking method.

4. Results and Discussion

4.1. Classification Performance

Our model is verified on the ISRUC [31,32] and SHHS [33,34] datasets. The obtained model performance is shown in Table 1. Dataset introduction and baseline settings refer to Appendices A.1 and A.2.

Table 1 shows the performance comparison of several baseline results. The proposed model has the best performance on the two datasets. The accuracy, MF1, and Kappa of the models 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.
Table 1. Performance comparison of the state-of-the-art approaches.

| Model               | ISRUC Accuracy | ISRUC MF1 | ISRUC Kappa | SHHS Accuracy | SHHS MF1 | SHHS Kappa |
|---------------------|----------------|-----------|-------------|---------------|----------|------------|
| MRCNN + AFR [25]   | 0.606          | 0.552     | -           | 0.689         | 0.557    | -          |
| CNN + BiLSTM [10]  | 0.788          | 0.779     | 0.730       | 0.719         | 0.588    | -          |
| MLP + LSTM [12]    | 0.779          | 0.713     | 0.758       | 0.802         | 0.779    | 0.792      |
| ResNet-50 [26]     | 0.782          | -         | 0.674       | 0.837         | -        | 0.754      |
| ARNN + RNN [22]    | 0.789          | 0.763     | 0.725       | 0.865         | 0.785    | 0.811      |
| STGCN [15]         | 0.821          | 0.808     | 0.769       | -             | -        | -          |
| proposed model     | 0.825          | 0.814     | 0.775       | 0.873         | 0.801    | 0.827      |

We bold the optimal result and underline the second-best result.

After comparing the deep learning method combining CNN and RNN, we found that the spatial location information of EEG could not be completely characterized, and the extracted features did not include regional connections between the brains. As shown in Table 1, the performance of traditional methods using CNN and RNN needs to be improved. Although the existing graph convolutional networks can fully extract the temporal context and spatial features, their spatial feature extraction does not consider the weight of each adjacent feature, resulting in some features being excessively important or neglected, resulting in differences between the extracted features and the real signal expression. Through the comparative test, it can be seen that the MSTGCN proposed can extract the spatial location information of EEG signals, and fully consider the importance of the interaction between adjacent channels to achieve the most advanced performance.

For each sleep stage, our model achieved the highest F1-score at the Wake, N1, and N2 stages, and the second highest level at the N3 and REM stages. Thus, for the most common confusion in the sleep staging task transition duration, namely Wake and N1, the proposed model’s proposed TSE achieves the most advanced level, with N1 performing 2% better than the second most advanced.

4.2. Number of Parameters and Training Duration

We compared the number of training parameters and training duration of several advanced models and MGANet models in each epoch in Table 2.

Table 2. Number of parameters and training duration for each epoch.

| Model               | Parameter | Training Duration |
|---------------------|-----------|-------------------|
| Supratak et al. [10]| $1.7 \times 10^7$ | 95.2              |
| Dong et al. [12]    | $1.7 \times 10^8$ | 268               |
| Chambon et al. [11] | $2.0 \times 10^5$ | 9.98              |
| Phan et al. [22]    | $1.2 \times 10^5$ | 367               |
| Chriskos et al. [23]| $4.1 \times 10^5$ | 75.8              |
| proposed model      | $1.5 \times 10^5$ | 69.7              |

As can be seen from Table 2, 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 $1.5 \times 10^5$, which is still on the same order of magnitude as other state-of-the-art methods. Furthermore, 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

To visualize the classification performance of the proposed model, we compared the sleep classification results of the MGANet model with sleep hypnograms manually scored by sleep experts, as shown in Figure 6. The differences between MGANet and expert manual scoring were mostly in the N1 and REM stages. Although the proposed model
effectively improved the classification accuracy of the N1 stage; the class imbalance of N1 stage causes the classification accuracy to drop.

**Figure 6.** The scored hypnogram by sleep experts and proposed MGANet. (a) Hypnogram manually scored by sleep experts. (b) Hypnogram automatically scored by the proposed MGANet.

Figure 6 shows that the MGANet projected for 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 repeatedly cycle between the waking to light sleep stages during this period. Around the 330 epoch, the model confuses REM with N2, a period of human dreaminess. Some incorrect 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 inaccurately score these stages when manually scoring. Most of the results are a misjudgment of N2 as N1, or REM as N2, which are also widespread inconsistencies in other baseline models.

**4.4. Ablation Experiments Results**

In order to compare the performance of the proposed model in each module, we designed the following ablation experiments. The variable settings and experimental results are shown in Table 3 and Figure 7. The baseline is set as the spatial multi-layer graph attention convolution module, variable A is the RNN convolution, variable B is the spatial-temporal attention, and variable C is TSE. Table 3 shows that the performance of the model using a 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. It could be that the first is the cause of the lack of time domain signal feature extraction and the second is why the model of the network layer number is less.
Table 3. Ablation model settings.

| Model Variable | Accuracy% | F1-Score% | Kappa% |
|----------------|-----------|-----------|--------|
| Baseline       | 80.9      | 79.6      | 75.4   |
| Baseline + A   | 81.5      | 80.3      | 76.2   |
| Baseline + B   | 81.0      | 79.4      | 75.5   |
| Baseline + C   | 81.4      | 80.3      | 76.1   |
| Baseline + A + B| 81.9     | 80.6      | 76.7   |
| Baseline + A + C| 82.4     | 81.1      | 77.3   |
| Baseline + B + C| 81.6     | 80.2      | 76.3   |
| Baseline + A + B + C (MGANet) | **82.5** | **81.4** | **77.5** |

We bold the optimal result and underline the second-best result.

Figure 7 shows eight confusion matrices obtained from ablation experiments, demonstrating the effectiveness of the proposed model in improving the classification of transition stages. In the past decades, many researchers have discussed the problem of stage-class imbalance [35] in sleep datasets, which seriously affects the classification performance of models. The transitional N1 stage accounts for only about 5–10% of a sleep cycle, while REM accounts for about 20–25% [36]. Dataset class imbalance tends to exacerbate the confusion degree in the transition stage. For this reason, traditional methods use oversampling and undersampling to expand or reduce the amount of data to achieve the purpose of data balance, but such methods can easily cause data overfitting [37]. In addition, some researchers have used weighted cross entropy loss function to balance the data. When the model performance is poor, the learning speed becomes fast; conversely, the learning rate decreases and the model performance cannot be further improved. Because the transitional stage of sleep confuses the characteristics of multiple stages, it makes it difficult to stage sleep. The current experiments have shown that the performance of the model is significantly improved after adding TSE, and it has also been proved to be helpful for accurate classification of confusion stages. Finally, the results show that the proposed MGANet adds variables A, B, and C at the same time, and the model has the best performance; its accuracy, F1 score, and kappa were 82.5, 81.4, and 77.5, respectively. Suboptimal results can be obtained using RNN and transition stage estimation methods.
When the number of attention items are larger, such as $h = 16$ and $h = 20$, the performance of the model can achieve the best effect. The number of heads determines the influence of adjacent channels on the weighting of the current channel, and selecting an appropriate number of heads can achieve the best effect of the model. The number of heads determines the influence of adjacent channels on the weighting of the current channel, and selecting an appropriate number of heads can achieve the best effect of the model.

When the number of heads $h = 2, 4, 8, 16$, or $20$, the accuracy and F1-score were obtained to explore the optimal number of heads used in Figure 8. $h = 8$ achieved the highest accuracy and F1-scores of 82.5 and 81.4, respectively.

4.5. Heads Numbers

We also discussed the number of heads of multi-head attention proposed, and designed it so that when the number of heads $h = 2, 4, 8, 16$, or $20$, the accuracy and F1-score were obtained to explore the optimal number of heads used in Figure 8. $h = 8$ achieved the highest accuracy and F1-scores of 82.5 and 81.4, respectively.

In Figure 8, the head numbers of MGANet represent the standard for distinguishing features of center nodes. Target features are extracted from multiple angles to avoid feature redundancy or single features. As shown in Figure 8, the model achieves best performance when $h = 8$ is used. When the number of attention heads is small, such as $h = 2$ or $h = 4$, the accuracy of the model is only 80.3% and 80.7%, respectively. Due to the lack of aggregation features in each head, it results in ineffective multi-attention mechanisms. When the number of attention items are larger, such as $h = 16$ and $h = 20$, the performance gradually decreases, which may be caused by the increase of aggregation features leading to over-fitting of the model. The number of heads determines the influence of adjacent channels on the weighting of the current channel, and selecting an appropriate number of heads can achieve the best effect of the model.
4.6. Speed Training

We compared the training rates of the proposed model with STGCN [17] and MSTGCN [27] on 80 epochs. The experimental results are shown in Figure 9. The proposed model converges faster and the loss is closer to 0 than others.

![Graph showing comparison of training losses with STGCN and MSTGCN over 80 epochs.](image)

**Figure 9.** Comparison of training losses with STGCN and MSTGCN over 80 epochs.

As shown in Figure 9, the losses of the proposed model decreased the fastest among the first 10 epochs, reaching the lowest losses compared with the other two methods. Based on the speeding training in Figure 9, we consider that the multi-head graph attention convolution used by MGANet aggregates features, calculates the corresponding weights, and obtains the most refined features for model training. Therefore, the running time and running efficiency are better than other methods. It is worth noting that the proposed model converges faster and loses less than other methods.

4.7. Attention Channel Visualization

We visualized the MGANet constructed through nodes and edges in the brain, as shown in Figure 10. A graph structure was constructed of the human brain. The MGANet was used to correspond to the current node and that of its neighbors to establish the adjacent relationship, extract the dynamic spatial features, and display the schematic diagram of multi-orientation construction.

Figure 10 shows that the multiple neighbor nodes are aggregated by adjacent edges carrying information to form the feature representation of the current node. For the central node, the non-Euclidean distance and weight of different neighbor nodes are significantly different, so it is necessary to consider the description of their mutual influence. The dynamic update feature is faultlessly implemented in the MGANet network.
Figure 10. Visualization of MGANet spatial features. (a) Topside and bottom side. (b) Frontal side and backside. (c) Medial view of left and right. (d) Lateral view of left and right.

5. Conclusions

In this paper, we proposed a novel and state-of-the-art algorithm for a multi-layer spatial graph attention convolutional network for sleep staging. The proposed method uses graph attention convolution to process non-Euclidean sequences in the spatial dimension, RNN to process spectrum information and time dependence in the temporal dimension, and the attention mechanism causes the model to focus on attributing the most appropriate weightings to adjacent channels. The proposed transitional stage estimator makes the model more suitable for the sleep transition classification stage and solves the problem of poor classification accuracy caused by the confusion of transition stages. The proposed MGAN not only outperforms the advanced works, but also achieves the most advanced performance on the ISRUC-S3 and SHHS.

Most EEG research has been 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 datasets; hence, the proposed model cannot be well generalized and is not suitable for clinical diagnosis at present. In future research, we will build a semi-supervised model and an unsupervised model, and will use the easily obtained unlabeled data to classify sleep stages to improve generalization ability.

Author Contributions: Conceptualization, Q.W.; methodology, Q.W.; software, Q.W.; resources, Q.W.; writing—original draft preparation, Q.W.; writing—review and editing, Y.S.; supervision, H.G. and S.T.; project administration, Y.G.; funding acquisition, Y.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grants No. 61673222) and Jiangsu postgraduate practical innovation program (No. SJCX22_0333).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: In this article, we used SHHS and ISRUC data sets. Sleep Heart Health Study (SHHS) is a multicenter cohort study of the cardiovascular and other consequences of sleep-disordered breathing. The data download link: https://sleepdata.org/datasets/shhs, accessed on 10 May 2021. ISRUC-S3 contains 126 recordings of 118 unique subjects recorded between 2009 and 2013 at the Center for Sleep Medicine, University Hospital, Coimbra, Portugal. The third panel we used contained recordings of physiological signals from 10 healthy subjects, 9 males and 1 female. The data download link is as follows: https://sleeptight.isr.uc.pt/ISRUC_Sleep/, accessed on 30 May 2021.

Acknowledgments: The authors wish to thank the editor and the anonymous reviewers for their valuable suggestions on improving this paper. We thank the support in part by the National Natural Science Foundation of China (No. 61673222) and Jiangsu postgraduate practical innovation program (No. SJCX22_0333).

Conflicts of Interest: The authors declare no conflict of interest.
Appendix A

Appendix A.1. Dataset and Experiment Settings

We evaluated our framework using two public datasets: SHHS-2 [22] and ISRUC-S3 [23].

Sleep Heart Health Study (SHHS): SHHS is a multicenter cohort study of the cardiovascular and other consequences of sleep-disordered breathing. Subjects had a variety of medical conditions, including pulmonary disease, cardiovascular disease, and coronary heart disease. Our study using the second part of the SHHS included the performance of two EEG channels, two EOG channels, and one EMG channel over 2651 nights. To minimize the impact of these disorders, we selected subjects who were considered regular sleepers (e.g., apnea-hypopnea index or AHI < 5) based on studies [23]. In the end, we selected 329 subjects, 52.8% of whom were female.

ISRUC-S3: ISRUC-S3 contains 126 recordings of 118 unique subjects recorded between 2009 and 2013 at the Center for Sleep Medicine, University Hospital, Coimbra, Portugal. The third panel we used contained recordings of physiological signals from ten healthy subjects, nine males and one female. Sleep stages and sleep events for each PSG were scored by two technicians according to AASM guidelines, dividing each recording into 30 epochs. The ISRUC dataset contains six EEG channels, two EOG channels, three EMG channels, and one ECG channel. For both data sets, we applied the following pre-processing steps. (1) We excluded any unknown stages that were not part of any sleep stage. (2) We combined stages N3 and N4 into one stage (N3) according to AASM criteria. (3) We only included 30 min of wakefulness before and after sleep to increase attention to sleep stages.

The experimental configuration adopted in this paper is as follows: The hardware environment is Intel Xeon Silver 4116 CPU, 16G memory, NVIDIA CTX3060Ti GPU; the software environment is Win10 system, Python3.8, and the deep learning framework is Tensorflow 1.14.0. The proposed model was compared with six baseline models, and experimental data preprocessing and evaluation settings were consistent with other baseline models for fairness of comparison. Specifically, the experimental process was evaluated on multiple channels using 10-fold cross-validation, with 20% of the training set as the test set. Parameter Settings in the experiment are shown in Table A1.

Table A1. Proposed model experiment parameter setting.

| Parameter     | Batch Size | Learning Rate | Reduction Rate | Optimizer | Epochs | $\beta_1$ | $\beta_2$ | $\lambda$ |
|---------------|------------|---------------|----------------|-----------|--------|-----------|-----------|-----------|
| Value         | 32         | 0.0002        | 2              | Adam      | 80     | 0.9       | 0.999     | $10^{-4}$ |

Appendix A.2. Baseline

We compared the proposed model with six other advanced models in terms of accuracy, F1-score, and Kappa. The comparison of experimental results between the proposed MGANet model and other baseline models is shown in Table 2. The highest performance is highlighted in bold and the second highest performance is underlined. Baseline model settings are as follows:

Supratak et al., 2017 [10]: A CNN was used to obtain temporal invariant features, and BiLSTM acquired contextual features for sleep staging.

Dong et al., 2017 [12]: A mixed neural network of multi-layer perceptron (MLP) and long and short-term memory (LSTM) to classify sleep stages.

Phan et al., 2019 [22]: A neural network converts sleep stage into sequence to sequence, used bidirectional RNN and RNN.

Eldele et al., 2021 [25]: Multi resolution CNN with adaptive feature recalibration is combined with multi head self-attention mechanism to achieve sleep classification.

Alexander et al., 2021 [26]: In the mixed queue environment, sleep classification is carried out by using deep residual network.

Jia et al., 2021 [27]: Proposed a multi-view spatial-temporal graph convolutional network with domain generalization to classify sleep stages.
Appendix A.3. Pseudocode

Algorithm A1: MGANet for sleep staging

Input: An information network: $G_s = (N, C, A)$

Output: Node vector fusion information representations $\mathbf{f}_n$

Initialization: $i = 0$, training epochs $I$ and labeled nodes $y_l$

While $i \leq I$ do

for a vertex $y_k \in V$ do

$\mathbf{f}_{nk} \leftarrow (\mathbf{W}_{nk}^g \mathbf{f}_g + \mathbf{W}_{nk}^f \mathbf{f}_f) / \text{Learn sleep vector for each spatial feature.}$

$\mathbf{f}_{nk} \leftarrow (\mathbf{F}_2, \mathbf{F}_3, \ldots, \mathbf{F}_{i}) / \text{Learn attention spatial feature representation.}$

$\mathbf{f}_{nk} \leftarrow / \text{Learn the sleep vector for each temporal feature by Equation (8).}$

$\mathbf{f}_{nk} \leftarrow (\mathbf{F}_2, \mathbf{F}_3, \ldots, \mathbf{F}_{i-1}) / \text{Learn attention temporal feature representation.}$

end

$u \leftarrow a(W_{nk}^g \mathbf{f}_g + W_{nk}^f \mathbf{f}_f) / \text{Calculate attention coefficient.}$

$\mathbf{p}_n \leftarrow \text{softmax}(\mathbf{F}_c(G(f))) / \text{sleep stage probability distribution.}$

$i \leftarrow i + 1$

end

References

1. Carskadon, M.A.; Rechtschaffen, A. Monitoring and staging human sleep. *Princ. Pract. Sleep Med.* 2011, 5, 16–26.

2. Killgore, W. Effects of sleep deprivation on cognition. *Prog. Brain Res.* 2010, 185, 105–129. [PubMed]

3. Acharya, R.; Faust, O.; Kannathal, N.; Chua, T.; Laxminarayan, S. Non-linear analysis of EEG signals at various sleep stages. *Comput. Methods Programs Biomed.* 2016, 122, 324–333. [CrossRef] [PubMed]

4. Berry, B.; Brooks, R.; Gamaldo, C. The non-linear analysis of EEG signals at various sleep stages. *Comput. Methods Programs Biomed.* 2005, 80, 37–45. [CrossRef] [PubMed]

5. Rechtschaffen, A. A manual for standardized terminology, techniques and scoring system for sleep stages in human subjects. In *Rules, Terminology and Technical Specifications*; American Academy of Sleep Medicine: Darien, IL, USA, 2012; Volume 76, p. 2012.

6. Berry, R.; Budhiraja, R.; Gottlieb, D.J. 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.* 2012, 8, 597–619. [CrossRef] [PubMed]

7. Ghimatgar, H.; Kazemi, K.; Helfroush, M.S.; Aarabi, A. An automatic single-channel EEG-based sleep stage scoring method based on hidden Markov Model. *J. Neurosci. Methods* 2019, 26, 1587–1597. [CrossRef] [PubMed]

8. Memar, P.; Faradji, F. A novel multi-class EEG-based sleep stage classification system. *IEEE Trans. Neural Syst. Rehabil. Eng.* 2017, 26, 84–95. [CrossRef]

9. Alickovic, E.; Subasi, A. Ensemble SVM method for automatic sleep stage classification. *IEEE Trans. Instrum. Meas.* 2018, 67, 1258–1265. [CrossRef]

10. Supratak, A.; Dong, H.; Wu, C. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* 2017, 25, 1998–2008. [CrossRef]

11. Chambon, S.; Galtier, M.N.; Arnal, P.J.; Wainrib, G.; Gramfort, A. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Trans. Neural Syst. Rehabil. Eng.* 2018, 26, 758–769. [CrossRef]

12. Dong, H.; Supratak, A.; Pan, W. Mixed neural network approach for temporal sleep stage classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* 2017, 26, 324–333. [CrossRef]

13. Tsinalis, O.; Matthews, P.M.; Guo, Y. Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders. *Ann. Biomed. Eng.* 2016, 44, 1587–1597. [CrossRef] [PubMed]

14. Sekkal, R.N.; Bereki-Reguig, F.; Ruiz-Fernandez, D. Automatic sleep stage classification: From classical machine learning methods to deep learning. *Biomed. Signal Process. Control* 2022, 77, 103751. [CrossRef]

15. Honey, C.J.; Thivierge, J.P.; Sporns, O. Can structure predict function in the human brain? *Neuroimage* 2010, 52, 766–776. [CrossRef]

16. Liang, Z.; Zhou, R.; Zhang, L. EEGFuseNet: Hybrid unsupervised deep feature characterization and fusion for high-dimensional EEG with an application to emotion recognition. *IEEE Trans. Neural Syst. Rehabil. Eng.* 2021, 29, 1913–1925. [CrossRef] [PubMed]

17. Veličković, P.; Cucurull, G.; Casanova, A. Graph Attention Networks. *Int. Conf. Learn. Represent.* 2018, 1050, 4.

18. Allam, J.P.; Samantaray, S.; Behara, C. A customized deep learning algorithm for drowsiness detection using single-channel EEG signal. In *Artificial Intelligence-Based Brain-Computer Interface*; Academic Press: Cambridge, MA, USA, 2022; pp. 189–201.
21. Bik, A.; Sam, C.; de Groot, E.R. A scoping review of behavioral sleep stage classification methods for preterm infants. *Sleep Med.* 2022, 90, 74–82. [CrossRef]

22. Phan, H.; Andreotti, F.; Cooray, N. SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Trans. Neural Syst. Rehabil. Eng.* 2019, 27, 400–410. [CrossRef]

23. Chriskos, P.; Frantzidis, C.A.; Gkivogkli, P.T. Automatic sleep staging employing convolutional neural networks and cortical connectivity images. *IEEE Trans. Neural Netw. Learn. Syst.* 2019, 31, 113–123. [CrossRef]

24. Jia, Z.; Cai, X.; Zheng, G. SleepPrintNet: A multivariate multimodal neural network based on physiological time-series for automatic sleep staging. *IEEE Trans. Artif. Intell.* 2020, 1, 248–257. [CrossRef]

25. Eldele, E.; Chen, Z.; Liu, C. An attention-based deep learning approach for sleep stage classification with single-channel eeg. *IEEE Trans. Neural Syst. Rehabil. Eng.* 2021, 29, 809–818. [CrossRef] [PubMed]

26. Alexander, N.O. Automatic sleep stage classification with deep residual networks in a mixed-cohort setting. *Sleep* 2021, 44, 161.

27. Jia, Z.; Lin, Y.; Wang, J. Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* 2021, 29, 1977–1986. [CrossRef]

28. Khare, S.K.; Bajaj, V. Time–frequency representation and convolutional neural network-based emotion recognition. *IEEE Trans. Neural Netw. Learn. Syst.* 2020, 32, 2901–2909. [CrossRef]

29. Arias-Vergara, T.; Klumpp, P.; Vasquez-Correia, J.C.; Nöth, E.; Orozco-Arroyave, J.R.; Schuster, M. Multi-channel spectrograms for speech processing applications using deep learning methods. *Pattern Anal. Appl.* 2021, 24, 423–431. [CrossRef]

30. Lopac, N.; Hržić, F.; Vuksanović, I.P.; Lerga, J. Detection of Non-Stationary GW Signals in High Noise From Cohen’s Class of Time–Frequency Representations Using Deep Learning. *IEEE Access* 2021, 10, 2408–2428. [CrossRef]

31. Khalighi, S.; Sousa, T.; Santos, J.M. ISRUC-Sleep: A comprehensive public dataset for sleep researchers. *Comput. Methods Programs Biomed.* 2016, 124, 180–192. [CrossRef]

32. Guillot, A.; Sauvet, F.; During, E.H. Dreem open datasets: Multi-scored sleep datasets to compare human and automated sleep staging. *IEEE Trans. Neural Syst. Rehabil. Eng.* 2020, 28, 1955–1965. [CrossRef]

33. Zhang, G.Q.; Cui, L.; Mueller, R. The National Sleep Research Resource: Towards a sleep data commons. *J. Am. Med. Inform. Assoc.* 2018, 25, 1351–1358. [CrossRef]

34. Quan, S.F.; Howard, B.V.; Iber, C.; Kiley, J.P.; Nieto, F.J.; O’Connor, G.T.; Rapoport, D.M.; Redline, S.; Robbins, J.; Samet, J.M.; et al. The sleep heart health study: Design, rationale, and methods. *Sleep* 1997, 20, 1077–1085.

35. Foulkes, W.D. Dream reports from different stages of sleep. *J. Abnorm. Soc. Psychol.* 1962, 65, 14. [CrossRef]

36. Carskadon, M.A.; Dement, W.C. Normal human sleep: An overview. *Princ. Pract. Sleep Med.* 2005, 4, 13–23.

37. Batista, G.E.; Prati, R.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor. Newsl.* 2004, 6, 20–29. [CrossRef]