A novel feature relearning method for automatic sleep staging based on single-channel EEG

Yujie Tao[1] · Yun Yang[1,2,3] · Po Yang[4] · Fengtao Nan[5] · Yan Zhang[1] · Yulong Rao[1] · Fei Du[5]

Abstract

Correctly identifying sleep stages is essential for assessing sleep quality and treating sleep disorders. However, the current sleep staging methods have the following problems: (1) Manual or semi-automatic extraction of features requires professional knowledge, which is time-consuming and laborious. (2) Due to the similarity of stage features, it is necessary to strengthen the learning of features. (3) Acquisition of a variety of data has high requirements on equipment. Therefore, this paper proposes a novel feature relearning method for automatic sleep staging based on single-channel electroencephalography (EEG) to solve these three problems. Specifically, we design a bottom–up and top–down network and use the attention mechanism to learn EEG information fully. The cascading step with an imbalanced strategy is used to further improve the overall classification performance and realize automatic sleep classification. The experimental results on the public dataset Sleep-EDF show that the proposed method is advanced. The results show that the proposed method outperforms the state-of-the-art methods. The code and supplementary materials are available at GitHub: https://github.com/raintyj/A-novel-feature-relearning-method.

Introduction

Sleep accounts for one-third of human life and is a critical link in human life. The quality of sleep affects many aspects of a person's physical health, mental health, and memory [1–3]. However, the task of sleep quality analysis is not only very demanding for physicians but also requires equipment with a high level of expertise. Polysomnography (PSG) is a powerful tool for sleep assessment that contains data such as electroencephalogram (EEG), electrooculogram (EOG), and electromyography (EMG). Physicians need to manually clas-

sify the collected PSG data records, which is a subjective and tedious process. Traditional machine learning methods based on manual extraction of statistical features usually include the following four steps: data preprocessing, feature extraction, feature selection, classification. Because of the need for strong professionalism to extract and select representative features, it is not friendly to researchers.

Step 1: Human body electrical signals are easily affected by other physiological electrical signals and the environment during the collection process, so preprocessing methods are needed to remove some noise effects. Methods such as multi-scale principal component analysis (PCA) [4], or wavelet transform [5], or notch filter and band-pass Butterworth filter [6], etc.

Step 2: Extract feature information from the polysomnography map, such as maximum value, median value, entropy, and energy [7].

Step 3: Use methods such as Best Subset Program (BSP) [8], Minimum Redundancy Maximum Correlation Algorithm (MRMR) [9], and recursive feature elimination algorithm based on support vector machine (SVM) [10] to select the best feature subset.

✉ Yun Yang
yangyun@ynu.edu.cn

1 National Pilot School of Software, Yunnan University, Kunming 650091, China

2 Kunming Key Laboratory of Data Science and Intelligent Computing, Kunming 650500, China

3 Key Laboratory in Software Engineering of Yunnan Province, Kunming 650500, China

4 Department of Computer Science, The University of Sheffield, Sheffield, UK

5 School of Information Science and Engineering, Yunnan University, Kunming 650500, China

Table 1 Characteristic of each stage, adapted from AASM

Stage	Unique feature	Similar feature
Wake	beta wave($>$**13 Hz**)	$>$50% alpha (8–13 Hz) activity, mixed (2–7 Hz) frequency activity
N1	"vertex sharp waves (5–14 Hz, $>$75uV)"	"50% of the epoch consists of relatively low voltage, mixed (2–7 Hz) activity, $<$50% alpha (8–13 Hz) activity "
N2	—	high voltage, mixed (2–7 Hz) activity, K complexes, sleep spindles
N3	"slow wave activity (**0.25–2 Hz**)"	high voltage, low frequency activity
REM	"sawtooth waves (2–6 Hz)"	relatively low voltage mixed (2–7 Hz) frequency EEG, alpha (8–13 Hz) activity

Step 4: Use various machine learning classification methods to perform sleep stages on selected feature combinations, such as decision tree [11], clustering [12,13], etc.

Deep learning has gradually become the mainstream method in recent years, because it requires no domain knowledge and can implement end-to-end systems excellently. Some researchers have found that the sleep process has a certain transitional regularity [14]. Based on this, it is common practice to enhance the central era information by taking advantage of the surrounding epochs. For instance, Tsinalis et al. [15] combined one preceding and following epochs as common input and converted multiple one-dimensional features into two-dimensional features for learning by a stacking layer in Convolutional Neural Networks (CNN), thus realizing automatic sleep staging. Li et al. [16] used a many-to-one strategy, took multi-epoch (3 epochs) raw EEG signals as its input, and relabeled the input. In addition, they set the threshold of softmax of their CCN-SE network according to the data distribution to alleviate the problem of class imbalance. Seo et al. [17] designed a network with a modified ResNet-50 and a two-layered BiLSTM to capture intra- and inter-epoch representative features, and compared the impact of inputting one, four (3 past), and ten (9 past) epochs data on the classification results. Their experiments show that the more input data epochs, the better the results, which proves that there is indeed a certain correlation between sleep epochs. In [18] and [19], researchers built multi-task network architectures based on the theory that most adjacent epochs have the same label. For each input epoch, the categories of surrounding epochs were additionally calculated, and joint decision-making was carried out by assigning different weights. However, the above methods are all based on the research of the correlation between the various stages, ignoring the similarity of the internal features of each stage, as shown in the summary of Table 1.

For sleep staging tasks, just like PSG, the diversity of data types will affect the classification effect. Jia et al. [20] input EEG, EOG, and EMG into an independent CNN branch network with multi-scale and residual connections (called SleepPrintNet), and perform feature fusion and classification. Amelia et al. [21] used two PSG channel data (located at two different positions on the scalp), and after data preprocessing and data enhancement by overlapping windows in order, they were input into the CNN + LSTM network for classification. Xu et al. [22] designed a lightweight convolutional neural network model to detect EEG fatigue status. They decomposed the five-channel EEG signal into multiple frequency bands and sent them to the convolutional network, respectively, and finally used the integrated learning method to weigh and vote on the network results. However, multimodal data often require the subject to wear more sensors. On the one hand, it has an impact on sleep itself; on the other hand, it is not conducive to the promotion of daily sleep monitoring.

To solve the above problems, this paper proposes a novel feature relearning method for automatic sleep staging based on single-channel EEG, which aims to reduce equipment requirements by using single-channel data and mining information within the stage. First, merge N1 and REM, which have strong similarities and a small amount of data [23]. In the first part of the method, a novel stacked network is used for four classifications. In the second part of the method, the CNN block network is used to classify N1 and REM. The two parts are cascaded to get the final sleep staging result. The contributions of this paper are summarized as follows: (1) We develop a bottom–up and top–down network combined with the attention mechanism and use the cascading step with an imbalanced strategy, which can mine feature information within the stage. (2) We achieve automatic sleep staging tasks, and avoid any prior knowledge in the automated process, saving time, and manpower. (3) We only use single-channel EEG data, which reduces the requirements for equipment and facilitates the promotion of daily sleep monitoring applications.

The rest of this article is organized as follows. In Section 'Methods', we introduce the network structure and the framework of the proposed method. Section 'Materials' reflects the experiment and analysis in detail. Section 'Results and discussion' discusses results and visualization. Finally, Section 'Conclusion and future work' presents the conclusion and the future.

Methods

In this section, we explain in detail the theoretical knowledge of the technology used, and describe the network structure and framework of the proposed method.

Convolutional neural networks

Convolutional neural networks are one of the representative algorithms of deep learning [24]. They are widely used thanks to their excellent adaptive representational learning and selection capability [25–27]. They mainly consist of a convolutional layer, a pooling layer, and a fully connected layer. The convolutional layer performs local feature extraction on the input information. The pooling layer, also known as the downsampling layer, mainly performs downsampling with reduced data dimensionality, thus reducing the computational complexity fully connected layer integrates local features. By stacking convolutional layers layer by layer, the low-level feature information may suffer loss [28]. Therefore, our backbone architecture is inspired by the complementarity of the high-level and low-level information of the pyramid network [29] to obtain richer, more comprehensive, and more reliable features.

Long short-term memory

Recursive neural networks (RNN) and their variants can help with current tasks by integrating past information and have excellent learning abilities on problems related to sequential data processing [30–32]. Compared with classical RNN network, long short-term memory (LSTM) is more favorable to deal with long-term dependency problem [33], which is attributed to the "gate" structure. As shown in Fig. 1, f represents the forgetting gate, i is the input gate, o represents the output gate. x_t, h_t, c_t represents the input, output, and cell state of the network at time t. σ represents the sigmoid function. The sigmoid activation function is used in the gate structure to determine the amount of information transferred. When the sigmoid output is 1, the door opens, allowing messages to pass through; when the sigmoid output is 0, the gate is closed, preventing messages from passing through.

In Eq. (1), w and b represent the weight matrix and bias vector of the forgetting gate, respectively. The forgetting gate calculates h_{t-1} and x_t splicing results through sigmoid function, and determines the degree of information retention in the cell state at the previous moment

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f). \tag{1}$$

The input gate indicates how much information the current network input needs to be to reserve to cells (as shown in Eq. (2)). The calculation method of the information retention

for candidate cells ΔC_t is shown in Eq. (3). And Eq. (4) expresses the degree of selection of forgetting and input to update the current cell state

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\Delta C_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{3}$$

$$C_t = f_t \times C_{t-1} + i_t \times \Delta C_t. \tag{4}$$

Equations (5)–(6) represent how much information in the current cell is used as the output of the hidden layer, where Eq. (5) is the calculation formula of the output gate

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t \times tanh(C_t). \tag{6}$$

The convolutional block attention module

The attention mechanism comes from the fact that humans give different levels of attention to different parts of things when observing them, which can be understood as allocating resources with different importance under limited conditions. Attention mechanism has been widely used in image classification [26], activity recognition [34], machine translation [35], and other fields in recent years. The Convolutional Block Attention Module (CBAM) [36] can perform adaptive attention learning in channel and space dimensions with high efficiency. The attention is computed as

$$\begin{aligned} F' &= M_c(F) \otimes F \\ F'' &= M_s(F') \otimes F', \end{aligned} \tag{7}$$

where F and F' represent the input and intermediate feature graphs, respectively. M_c is the channel attention map and M_s is the spatial attention map. $\otimes$ denotes element-wise multiplication.

The channel attention mechanism is designed to measure the importance of different channel information. First, the original feature graph F is compressed into two global information F_{avg} and F_{max} using average pooling and maximum pooling operations. Then, they are input into a shared network with two layers of the neural network, and the elements are summed. W_0 represents the first-layer neural network in which the number of neurons is C/r, r represents the reduction ratio, and the activation function of W_0 is Relu. W_1 represents the second-layer neural network with c neurons. Finally, the channel attention feature map is obtained by sigmoid operation, where σ represents the sigmoid function

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F))) + MLP(MaxPool(F)) \\ &= \sigma(W_1(W_0(F_{avg}^c)) + W_1(W_0 F_{max}^c)). \end{aligned} \tag{8}$$

The spatial attention mechanism is to measure the key information of different spatial positions. First, the original
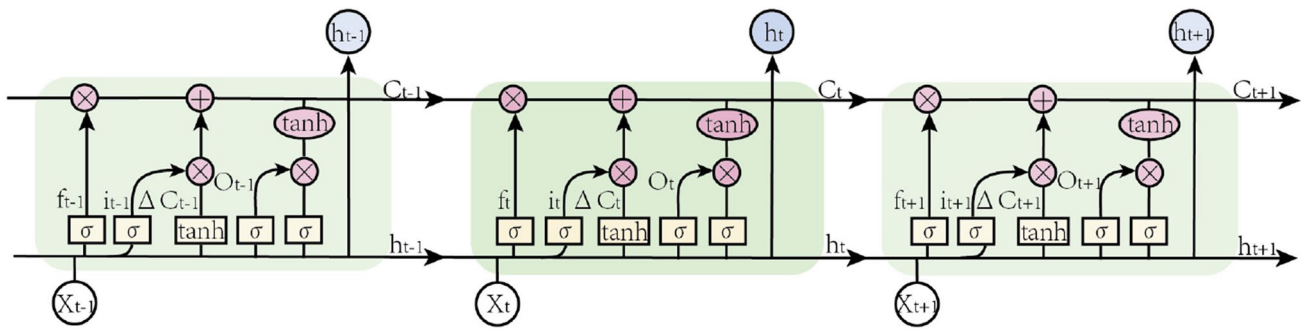
Fig. 1 LSTM structure diagram

feature graphs are averaged and maximized along the channel axis, and the two pooling results are concatenated in the channel dimension to obtain a 2D pooling result. Then, a kernel size = 7 filter is used for convolution operation. Finally, the sigmoid function is used to obtain the spatial attention weight matrix

$$M_s(F) = \sigma(f^{7\times7}([Avg\,Pool(F); Max\,Pool(F)]))$$
$$= \sigma(f^{7\times7}(F_{avg}^s); F_{max}^s). \tag{9}$$

Framework of the proposed method

This paper proposes a novel feature relearning method consisting of two parts, as shown in Fig. 2. The first part performs a four-class classification through three key modules : bottom–up, top–down, and feature fusion. In the bottom–up modules , multiple convolution operations are stacked to learn frequency information, and LSTM is used to extract the temporal information of frequency features of each layer. In the top–down modules , the upsampling method is used to strengthen information of the bottom layer. In the feature fusion modules , high-semantic features that help classification and shallow-level features that help localization [37] are fused by the idea of the attention mechanism, which can generate adaptive weights from different levels, rather than manually specifying. Considering the similarity between the N1 and REM and data imbalance, a random oversampling method is used in the second part to construct a new training dataset. Finally, the two parts are concatenated to divide the five different sleep stages.

Suppose there are N segments of 30 s single-channel EEG signals. The feature extraction workflow in part 1 is as follows:

$$x_i = f(x_{i-1}) \tag{10}$$
$$h_i = p(g(x_i)\|k(h_{i+1})), \tag{11}$$

where $f(x)$ turns x_t into feature vector x_{t+1} via CNN. $p(x)$ represents the convolutional block attention module, which can

automatically integrate and learn the importance of different levels of information. And $g(x)$ represents the operation of extracting time information from spatial features of each layer through the LSTM block. $\|$ is a concatenate operation that combines the features between the layers. $k(x)$ means the upsampling process of the left and right 0 paddings on the upper layer features.

According to AASM, dataset S1 is obtained by merging stage N3 and stage N4 into a single N3, deleting MOVE-MENT and UNKNOWN data. And Merge the N1 and REM into a single stage to obtain the dataset S2(WAKE, N1-REM, N2, N3). Part 1 and 2 are trained using dataset S2 and balanced dataset N1-REM, respectively, to obtain the final model.

Algorithm 1

Input: batch size of (x_i, y_i)
Output: classification results

Train
  for each $i \in [1, N]$ **do**
    Calculate the label using part 1.
    if label == N1-REM and y_i == N1-REM **then**
      Train part 2 using the balanced selection of
      dataset(x_i, y_i).
    else
      Label the sample as a label.
    end if
  end for

Test
  for a test instance, x with an unknown label **do**
    Predict the label using part 1.
    if label == N1-REM **then**
      Reclassify the sample as N1 or REM stage using
      part 2.
    else
      Label the sample as a label, end this test and
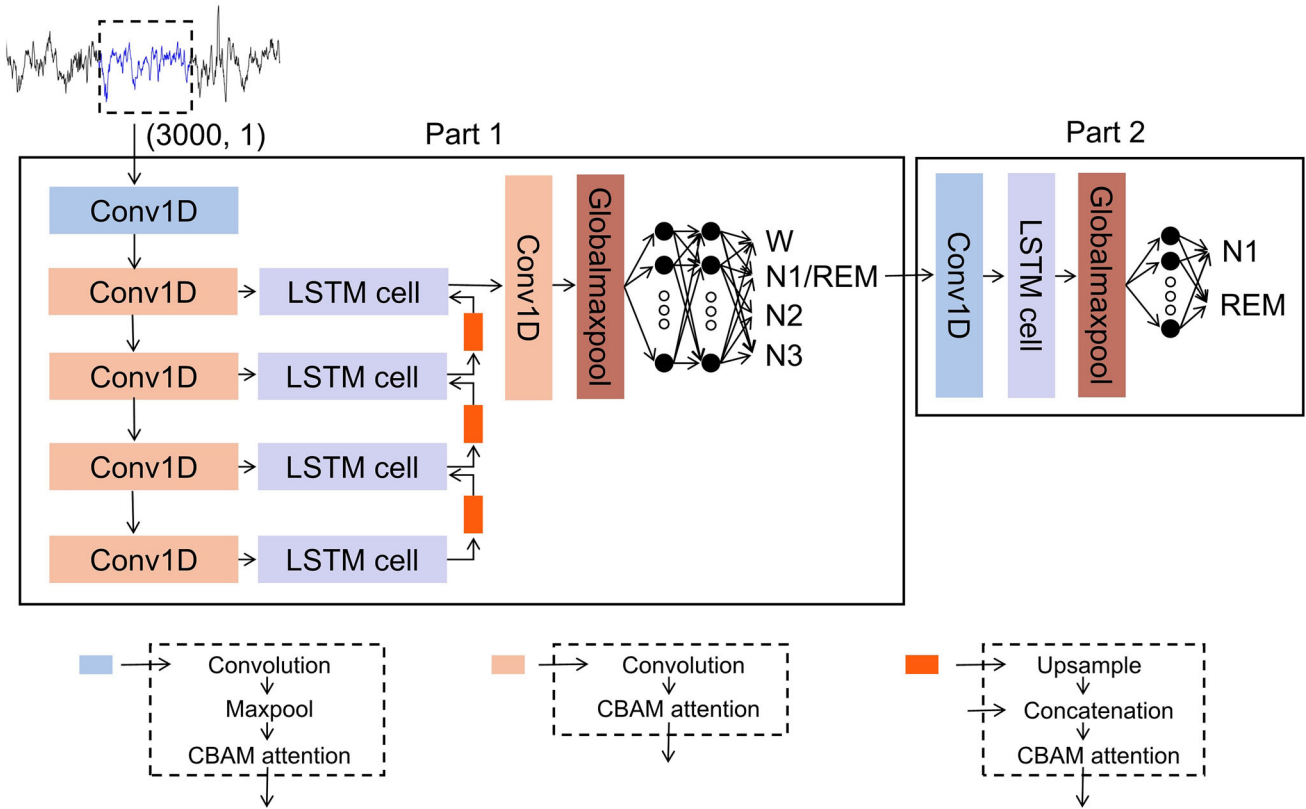   continue the following sample.

Fig. 2 Framework of the feature relearning method

Table 2 Statistical results on sleep-EDF dataset

Stage	W	N1	N2	N3	REM	Total
Number	8285	2804	17,799	5703	7717	42,308
Percentage	19.6%	6.6%	42.0%	13.5%	18.2%	–

Materials

In this section, we introduce the experimental dataset and evaluation metrics and conduct multiple sets of ablation experiments. All experiments are guaranteed to run five times in the same environment, and the average value is taken as the final result to ensure the authenticity of the experimental results.

Dataset

The public sleep dataset Sleep-EDF used in this study comes from Physiobank [38,39]. Each record in the dataset contains various physiological data, among them, EEG is the most widely used [40], with a sampling frequency of 100Hz. The sleep experts manually classify all recorded data into one of eight categories according to the Rechtschaffen and Kales (R&K) [41,42]. As recommended by the American Academy of Sleep Medicine (AASM) [14], stages N3 and N4 are combined into one N3, Movement and Unknown data

are removed. Therefore, the sleep stages are divided into five categories, W, N1, N2, N3, and REM. In this study, the raw 30 s EEGs are used as input without any other preprocessing.

Table 2 summaries the number and proportion of each stage. It can be found that there is an imbalance problem in the dataset. The N2 category has the largest amount of data, ranging from 2 to 6 times that of other categories. This is also one of the reasons that prompted us to merge the minority classes in the first part of the method.

Evaluation metrics and experimental design

We use Precision (PR), Recall (RE), and F1-score (F1) to evaluate the classification performance of each sleep stage and Accuracy (ACC) and macro-average F1-score (MF1) to evaluate the overall performance of the classification. Precision represents the proportion of correctly predicted as positive samples in predicted. Recall represents the propor-

tion of positive samples that are correctly predicted in the true positive samples. F1-score is the harmonic average of Precision and Recall. Accuracy is the proportion of the total number of all categories predicted to be correct to the total number of samples. MF1 is the average value of F1 for all classes. The formulas of these evaluation metrics are as follows:

$$PR = \frac{TP}{TP + FP} \tag{12}$$

$$RE = \frac{TP}{TP + FN} \tag{13}$$

$$F1 = \frac{2 \times PR \times RE}{PR + RE} \tag{14}$$

$$ACC = \frac{\sum_{c=1}^{C} TP_c}{N} \tag{15}$$

$$MF1 = \frac{\sum_{c=1}^{C} F1_c}{N}, \tag{16}$$

where TP represents correct prediction, and FP and FN represent incorrect prediction. FP represents the prediction of another class as this class. FN represents the prediction of this class as another class. And C represents the number of categories.

In this study, an Adam optimizer is used to train our models. Set the initial learning rate to 0.001 and the batch size to 128. Avoiding over-fitting caused by too few iterations, the early stopping method is adopted to control the interruption of learning at an appropriate time. We use the fivefold cross-validation to test the performance of the method. The idea is to divide the whole dataset into K parts according to the classification proportion and take one of them in turn as the test dataset and the rest of the $K-1$ parts as the training dataset. Finally, the average of the K results is taken as the evaluation value.

Ablation study

To clarify the influence of different single channels on the results, the experimental design is shown in Table 3. Under the same settings, the classification performance of Fpz-Cz is generally better than that of Pz-Oz. The Fpz-Cz channel has better performance on the proposed model, possibly because the channel contains more useful information. Therefore, in the following ablation studies, we all use the Fpz-Cz channel as the experimental data.

To prove the structural design of part 1, we conducted a series of ablation studies to prove the validity of each module of the method, as shown in Table 4. We compared the results of removing the attention structure (second row) and top–down connection (third row) from the original model (first row). The results prove that deleting both modules harms the classification effect. The attention mechanism has reliable

Table 3 Classification results based on different channels

Channle	Classification	ACC	MF1
Fpz-Cz	4-stage	86.5	86.7
	5-stage	82.7	76.8
Pz-Oz	4-stage	83.7	83.5
	5-stage	79.7	72.5

Table 4 The F1 results of part 1 ablation experiments

Operations	Wake	N1-REM	N2	N3
Our method	90.0	80.4	87.8	88.7
No attention	89.7	79.4	86.9	87.0
No top-down	88.6	79.2	87.3	87.8

Table 5 The F1 results of part 2 ablation experiments

Cascading	Oversampling	Undersampling	N1	REM
✓	✓		41.6	75.8
✓		✓	40.3	73.8
✓			39.6	77.1
			37.5	77.6
	✓		35.2	74.4
		✓	37.0	69.6

attention to different levels of information, and the connection between the upper and lower levels can reduce the loss of information and improve classification accuracy.

To explore the impact of the second part of the method on the classification results, we organized the following comparative experiments, as shown in Table 5. "Cascading" means that it is classified into four first and then classified into two. Both "oversampling" and "undersampling" are common methods of data balancing. Oversampling means random repetition of minority samples, and undersampling means randomly deleting the number of samples in the majority class. The first row in Table 5 represents the composition and results of the second part in the original model. The second row means to change the data balance processing from random oversampling to random undersampling in the second part. The third row indicates that the second part of the model directly performs cascade classification without data balancing operations. The fourth and fifth rows represent the classification results with random oversampling and random undersampling balancing operations when there is no cascade. The results show that the cascading step with random oversampling has the best overall performance. It can be seen that, regardless of whether there is a data balancing operation, the cascade step improves the overall results of N1 and REM. The last two rows in the table represent the N1 and REM results for the five classes of data balancing operations

without the cascading step. The results showed a negative effect on N1 and REM. This may be because the first part of the method has enough ability to learn the overall features of the data, and the sampling operation will make the learning become over-fitting to the training data. It is worth mentioning that, under the cascading step, although the sampling operation improves the classification performance of the N1 stage, it reduces the REM stage to some extent. The possible reason is that the first step of the cascade ignores some of the similar features of REM, and the sampling operation makes the model pay too much attention to learning N1 information. Due to the increase in overall performance, we still retain these operations.

Results and discussion

In this section, we select and compare the research results of the past 5 years, and conduct visual analysis.

Experimental results

Table 6 shows the confusion matrix of 42,308 test datasets on the Fpz-Cz channel, which correctly classified 34,985 datasets. The bold represents the number of samples that are correctly classified in each category. The sum of each column in the confusion matrix is the number of predicted samples, and the sum of each row is the number of real samples. The last three columns of each row are the performance metrics of each class. It can be seen from the confusion matrix results that the classification effect of Wake and N3 stages are the best. We speculate that this phenomenon is related to the unique wave frequencies of the two stages in Table 1. For the N2 stage, because the amount of data is several orders of magnitude more than other stages (as shown in Table 2), the classification effect is also very good. Although the N1 and REM phases have feature waves that are different from other phases, on the one hand, the frequency of the feature waves and other waves has overlapped, and on the other hand, their data volume is small (as shown in Table 2), so the effect is the worst. From the N1 and REM rows in the confusion matrix, it can be seen that they are easily predicted to be each other, and are also in the N2 stage, but are rarely predicted to the N3 stage, which is inseparable from the overlap of similar features in Table 1.

Comparison with other approaches

Table 7 describes the comparison between our method and the other methods in terms of classification method, feature extraction method, the channel used, accuracy, and macro-average f1-score. The bold represents the best results. It can be seen from Table 7 that our method has achieved ACC:

Table 6 Confusion matrix using the proposed method

	Predicted					Per-class Metrics		
	W	N1	N2	N3	REM	PR	RE	F1
W	**7351**	552	165	23	194	91.2	88.8	90.0
N1	390	**1349**	562	6	492	36.7	48.1	41.6
N2	131	671	**15,658**	564	775	87.6	88.0	87.8
N3	18	4	655	**5023**	3	89.4	88.1	88.7
REM	167	1100	842	4	**5604**	79.2	72.6	75.8

82.7, 79.7, and MF1: 76.8, 72.5 on the Fpz-Cz channel and the Pz-Oz channel, respectively.

The experimental results on the Pz-Oz channel are slightly weaker than the Fpz-Cz channel. We speculate that it is because the Fpz-Cz channel contains more information that is useful for classification.

The work of [43] is based on the multi-tapered spectrogram decomposition and semi-supervision, their performance is lower than other algorithms. Literature [16] focuses on stage correlation and uses deep learning methods for automatic feature extraction and classification, but only uses the convolution module. [44] converted the description text in the AASM manual into semantic features, and performed weighted fusion with the frequency domain and time domain features of the EEG signal for automatic classification. However, the way researchers divide the one-dimensional signal into multiple segments and compose the two-dimensional signal as input, it is easy to lose the local information of the segment to a certain extent, so the effect is slightly worse. Although our results are not as good as [16], it can be seen that we use less data, which is also one of our advantages. In contrast, our method can achieve a good classification effect under the condition of 30 s single-channel EEG data. The results show that our method of relearning features can achieve results comparable to multi-epoch input.

Visualization verification

We randomly select a test data file and perform visual analysis. As shown in Fig. 3, we draw the confidence of the method prediction as a boxplot diagram. The red line in the figure represents the median. The triangle represents the average value. The two horizontal lines on the top and bottom of the box chart represent the maximum and minimum probability.

Obviously, the median and the average probability of being correctly classified in the N1 stage is the smallest, followed by REM, N2, N3, and W. And the N1 and REM stages have the largest probability distribution range, and the classification effect is the most unstable. We speculate that this phenomenon is related to the distribution of the dataset and similar features, as shown in Tables 1 and 2.

Table 7 Performance comparison on Sleep-EDF

Citations	Methods	Feature type	Employed Channels	Overall results	
				ACC	MF1
[43]/2018	Gaussian Mixture Model	Hand-crafted	Pz-Oz EEG	73.2	–
[44]/2020	GAN + CNN + BiLSTM + AdaBoost	Learned	Fpz-Cz EEG	82.5	76.3
[16]/2021	CNN	Learned	Fpz-Cz EEG (90 seconds)	87.5	76.2
			Fpz-Cz EEG	81.3	65.5
Our method	cascaded CNN + LSTM	Learned	Fpz-Cz EEG	82.7	76.8
			Pz-Oz EEG	79.7	72.5
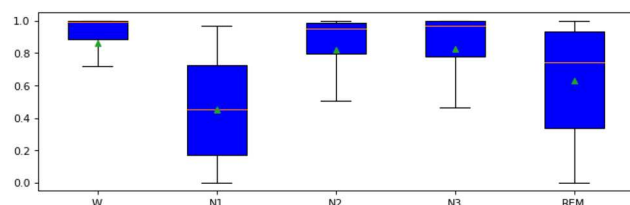
Fig. 3 Confidence distribution of sample classification

Conclusion and future work

In this paper, a novel feature relearning method is proposed. We design bottom–up and top–down model structures to effectively learn the features of each stage and use the cascading step to improve the classification performance further. The method implements an end-to-end staging, eliminating the need for professional knowledge. Only a single-channel EEG signal is used, which reduces the need for acquisition equipment, which shows that our method will be conducive to the practical application of sleep staging. The experimental results on the public dataset Sleep-EDF show that the proposed method is advanced. Although the complexity of the method has increased due to the hierarchical connection, the experimental results prove that the top–down connection achieves the complementation of information and comprehensively improves the classification performance to a certain extent.

In future work, we will explore a more concise model and conduct a more detailed study on the similarity of sleep stages, especially between N1 and other stages.

Availability of data and materials We evaluated our model with the Sleep-EDF dataset. The data set is publicly available on the Internet http://www.physionet.org/physiobank/database/sleep-edfx/.

Declarations

Conflict of interest The authors declare that there is no conflict of interest or no competing interests regarding the publication of this article.

Code availability The code and supplementary materials are available at GitHub: https://github.com/raintyj/A-novel-feature-relearning-method.

References

1. Frandsen R, Nikolic M, Zoetmulder M, Kempfner L, Jennum P (2015) Analysis of automated quantification of motor activity in rem sleep behaviour disorder. J Sleep Res 24(5):583–590
2. Tempesta D, Socci V, De Gennaro L, Ferrara M (2018) Sleep and emotional processing. Sleep Med Rev 40:183–195
3. Rauchs G, Desgranges B, Foret J, Eustache F (2005) The relationships between memory systems and sleep stages. J Sleep Res 14(2):123–140
4. Alickovic E, Subasi A (2018) Ensemble svm method for automatic sleep stage classification. IEEE Trans Instrum Meas 67(6):1258–1265
5. Xiao-ping C, Wei-xing H, Jing Y (2008) Sleep stage classification based on wavelet transformation and approximate entropy. ZHONGGUO ZUZHI GONGCHENG YANJIU YU LINCHUANG KANGFU 12(9):1701
6. Khalighi S, Sousa T, Pires G, Nunes U (2013) Automatic sleep staging: A computer assisted approach for optimal combination of features and polysomnographic channels. Expert Syst Appl 40(17):7046–7059

7. Güneş S, Polat K, Yosunkaya Ş (2010) Efficient sleep stage recognition system based on eeg signal using k-means clustering based feature weighting. Expert Syst Appl 37(12):7922–7928

8. Krakovská A, Mezeiová K (2011) Automatic sleep scoring: A search for an optimal combination of measures. Artif Intell Med 53(1):25–33

9. Herrera Luis Javier, Mora Antonio Miguel, Fernandes C, Migotina Daria, Guillén Alberto, Rosa Agostinho C (2011) Symbolic representation of the eeg for sleep stage classification. In *2011 11th International Conference on Intelligent Systems Design and Applications*, pages 253–258. IEEE

10. Koley B, Dey Debangshu (2012) An ensemble system for automatic sleep stage classification using single channel eeg signal. *Computers in biology and medicine*, 42(12):1186–1195

11. Liang S-F, Kuo C-E, Yu-Han H, Cheng Y-S (2012) A rule-based automatic sleep staging method. J Neurosci Methods 205(1):169–176

12. Diykh M, Li Y, Wen P (2016) Eeg sleep stages classification based on time domain features and structural graph similarity. IEEE Trans Neural Syst Rehabil Eng 24(11):1159–1168

13. Yang Y, Jiang J (2018) Adaptive bi-weighting toward automatic initialization and model selection for hmm-based hybrid meta-clustering ensembles. IEEE transactions on cybernetics 49(5):1657–1668

14. Iber C, Ancoli-Israel S, Chesson A, Quan SF (2007) For the american academy of sleep medicine. the aasm manual for the scoring of sleep and associated events: rules, terminology and technical specifications. *Westchester, IL: American Academy of Sleep Medicine*

15. Tsinalis Orestis, Matthews Paul M, Guo Yike, Zafeiriou Stefanos (2016) Automatic sleep stage scoring with single-channel eeg using convolutional neural networks. *arXiv preprint* arXiv:1610.01683

16. Li F, Yan R, Mahini R, Wei L, Wang Z, Mathiak K, Liu R, Cong F (2021) End-to-end sleep staging using convolutional neural network in raw single-channel eeg. Biomed Signal Process Control 63:102203

17. Seo H, Back S, Lee S, Park D, Kim T, Lee K (2020) Intra-and inter-epoch temporal context network (iitnet) using sub-epoch features for automatic sleep scoring on raw single-channel eeg. Biomed Signal Process Control 61:102037

18. Phan Huy, Andreotti Fernando, Cooray Navin, Chén Oliver Y, De Vos Maarten (2018) Joint classification and prediction cnn framework for automatic sleep stage classification. *IEEE Transactions on Biomedical Engineering*, 66(5):1285–1296

19. Phan Huy, Andreotti Fernando, Cooray Navin, Chén Oliver Y, De Vos Maarten (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

20. Jia Z, Cai X, Zheng G, Wang J, Lin Y (2020) Sleepprintnet: A multivariate multimodal neural network based on physiological time-series for automatic sleep staging. IEEE Transactions on Artificial Intelligence 1(3):248–257

21. Casciola Amelia A, Carlucci Sebastiano K, Kent Brianne A, Punch Amanda M, Muszynski Michael A, Zhou Daniel, Kazemi Alireza, Mirian Maryam S, Valerio Jason, McKeown Martin J, et al. (2021) A deep learning strategy for automatic sleep staging based on two-channel eeg headband data. *Sensors*, 21(10):3316

22. Xu Haoyan, Xu Xiaolong (2019) Lightweight eeg classification model based on eeg-sensor with few channels. In *2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, pages 464–473. IEEE

23. Zhou J, Wang G, Liu J, Duanpo W, Weifeng X, Wang Z, Ye J, Xia M, Ying H, Tian Y (2020) Automatic sleep stage classification with single channel eeg signal based on two-layer stacked ensemble model. IEEE Access 8:57283–57297

24. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. nature 521(7553):436–444

25. Yang Yun, Hu Yuanyuan, Zhang Xingyi, Wang Song (2021) Two-stage selective ensemble of cnn via deep tree training for medical image classification. *IEEE Transactions on Cybernetics*

26. Xue Gang, Liu Shifeng, Ma Yicao (2020) A hybrid deep learning-based fruit classification using attention model and convolution autoencoder. *Complex & Intelligent Systems*, pages 1–11

27. Li T, Zhang Y, Wang T (2021) Srpm-cnn: a combined model based on slide relative position matrix and cnn for time series classification. Complex & Intelligent Systems 7(3):1619–1631

28. He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian (2016) Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778

29. Lin Tsung-Yi, Dollár Piotr, Girshick Ross, He Kaiming, Hariharan Bharath, Belongie Serge (2017) Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125

30. Duan Z, Yang Y, Zhang K, Ni Y, Bajgain S (2018) Improved deep hybrid networks for urban traffic flow prediction using trajectory data. Ieee Access 6:31820–31827

31. Liao W, Ma Y, Yin Y, Ye G, Zuo D (2021) Improving abstractive summarization based on dynamic residual network with reinforce dependency. Neurocomputing 448:228–237

32. Xiao Z, Xin X, Xing H, Luo S, Dai P, Zhan D (2021) Rtfn: A robust temporal feature network for time series classification. Inf Sci 571:65–86

33. Chung Junyoung, Gulcehre Caglar, Cho KyungHyun, Bengio Yoshua (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint* arXiv:1412.3555

34. Zhang H, Xiao Z, Wang J, Li F, Szczerbicki E (2019) A novel iot-perceptive human activity recognition (har) approach using multihead convolutional attention. IEEE Internet Things J 7(2):1072–1080

35. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv:1409.0473

36. Woo Sanghyun, Park Jongchan, Lee Joon-Young, Kweon In So (2018) Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19

37. Zhao R, Xia Y, Wang Q (2021) Dual-modal and multi-scale deep neural networks for sleep staging using eeg and ecg signals. Biomed Signal Process Control 66:102455

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

39. Goldberger Ary L, Amaral Luis AN, Glass Leon, Hausdorff Jeffrey M, Ivanov Plamen Ch, Mark Roger G, Mietus Joseph E, Moody George B, Peng Chung-Kang, Stanley H Eugene (2000) Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220

40. Wulff Katharina, Gatti Silvia, Wettstein Joseph G, Foster Russell G (2010) Sleep and circadian rhythm disruption in psychiatric and neurodegenerative disease. *Nature Reviews Neuroscience*, 11(8):589–599

41. Rechtschaffen Allan (1968) A manual of standardized terminology, technique and scoring system for sleep stages of human subjects. *Public Health Service*

42. Hobson J Allan (1969) A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects: A. rechtschaffen and a. kales (editors). (public health service, u.s.

government printing office, washington, d.c., 1968, 58 p., $4.00). *Electroencephalography and Clinical Neurophysiology*, 26(6):644

43. Munk Andreas Muff, Olesen Kristoffer Vinther, Gangstad Sirin Wilhelmsen, Hansen Lars Kai (2018) Semi-supervised sleep-stage scoring based on single channel eeg. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2551–2555. IEEE

44. Xiang Hongxin, Zeng Ting, Yang Yun (2020) A novel sleep stage classification via combination of fast representation learning and semantic-to-signal learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE