Deep Learning in Automatic Sleep Staging With a Single Channel Electroencephalography

Mingyu Fu1, Yitian Wang1, Zixin Chen2, Jin Li3, Fengguo Xu4, Xinyu Liu1 and Fengzhen Hou1*

1 School of Science, China Pharmaceutical University, Nanjing, China, 2 College of Engineering, University of California, Berkeley, Berkeley, CA, United States, 3 College of Physics and Information Technology, Shaanxi Normal University, Xi’an, China, 4 Key Laboratory of Drug Quality Control and Pharmacovigilance, China Pharmaceutical University, Nanjing, China

This study centers on automatic sleep staging with a single channel electroencephalography (EEG), with some significant findings for sleep staging. In this study, we proposed a deep learning-based network by integrating attention mechanism and bidirectional long short-term memory neural network (AT-BiLSTM) to classify wakefulness, rapid eye movement (REM) sleep and non-REM (NREM) sleep stages N1, N2 and N3. The AT-BiLSTM network outperformed five other networks and achieved an accuracy of 83.78%, a Cohen’s kappa coefficient of 0.766 and a macro F1-score of 82.14% on the PhysioNet Sleep-EDF Expanded dataset, and an accuracy of 81.72%, a Cohen’s kappa coefficient of 0.751 and a macro F1-score of 80.74% on the DREAMS Subjects dataset. The proposed AT-BiLSTM network even achieved a higher accuracy than the existing methods based on traditional feature extraction. Moreover, better performance was obtained by the AT-BiLSTM network with the frontal EEG derivations than with EEG channels located at the central, occipital or parietal lobe. As EEG signal can be easily acquired using dry electrodes on the forehead, our findings might provide a promising solution for automatic sleep scoring without feature extraction and may prove very useful for the screening of sleep disorders.

Keywords: deep learning, single channel electroencephalography, automatic sleep staging, bidirectional long short-term memory, attention mechanism

INTRODUCTION

Sleep is important for the optimal functioning of the brain and the body (Czeisler, 2015). However, a large number of people suffer from sleep related disorders, such as sleep apnea, insomnia and narcolepsy (Ohayon, 2002). Effective and feasible sleep assessment is essential for recognizing sleep problems and making timely interventions.

Sleep assessment is generally based on the manual staging of overnight polysomnography (PSG) signals, including electroencephalogram (EEG), electrooculogram (EOG), electromyogram (EMG), electrocardiogram (ECG), blood oxygen saturation and respiration (Weaver et al., 2005), by trained and certified technicians. According to the American Academy of Sleep Medicine (AASM) manual (Iber et al., 2007), sleep can be staged as wakefulness (WAKE), rapid eye movement (REM) sleep and non-REM (NREM) sleep, which is further divided into three stages, N1, N2 and N3. Usually, it takes about 2–4 h for a technician to mark an overnight (lasting about 8 h)
The application of deep neural networks for automatic sleep staging is soaring (Table 1). The PhysioNet Sleep-EDF Expanded (PSEE) dataset (Goldberger et al., 2000; Kemp et al., 2000) was the most widely employed dataset in related studies. As shown in Table 1, Tsinalis et al. (2016) and Phan et al. (2019) reported an accuracy of 74.0% and 81.9% respectively, for 5-class sleep staging of the PSEE dataset with a CNN algorithm, while Supratak found that the combination of CNN and BiLSTM increased the accuracy to 82.4% (Supratak et al., 2017). There are also some datasets aside from PSEE that are routinely employed in studies of automatic sleep staging with a single-channel EEG and deep learning algorithms. Hsu et al. (2013) built an RNN model on the PhysioNet Sleep-EDF (PSEE) dataset and achieved an accuracy of 87.2%. On the Montreal Archive of Sleep Studies (MASS) dataset, Phan et al. (2019) built a CNN model and achieved an accuracy of 83.6% while Supratak et al. (2017) built a CNN-LSTM model and obtained an accuracy of 86.2%. A CNN was also applied on the Sleep Heart Health Study (SHHS) dataset, yielding an accuracy of 87% (Sors et al., 2018). However, few works investigated whether the performance of sleep staging can be further improved by the combination of BiLSTM and the attention mechanism. Aside from that, there is a lack of comparison between the performance of deep learning based and conventional feature extraction based models.

Although deep learning algorithms have shown themselves promising in automatic sleep staging with a single-channel EEG, few studies investigated whether the performance of such algorithms is sensitive to the choice of EEG channel. Therefore, in this study, the PSEE dataset and the DREAMS Subjects (DRM-SUB) dataset (Devuvyst, 2005) were used. Both datasets have more than one channel of EEG and the DRM-SUB dataset was involved in many automatic sleep staging studies with conventional feature extraction (Hassan and Bhuiyan, 2016a, 2017; Ghimatgar et al., 2019; Shen et al., 2019). A neural network named AT-BiLSTM was proposed, which uses the neural attention mechanism of the BiLSTM to classify sleep stages. For comparison, five other networks, CNN, LSTM, BiLSTM, the combination of CNN and LSTM (CNN-LSTM), and the combination of CNN and BiLSTM (CNN-BiLSTM) were also trained and tested. Our aims are threefold: first, to investigate whether AT-BiLSTM can achieve the highest performance among these networks; second, to confirm whether RNN algorithms (i.e., LSTM and BiLSTM) outperform CNN in sleep staging with single channel EEG; third, to explore whether the method of making hybrid networks further improves the performance of sleep staging.

### MATERIALS AND METHODS

#### Datasets

The data analyzed in this study were obtained from two open-access datasets: the DRM-SUB dataset and the PSEE dataset. The DRM-SUB consists of 20 whole-night PSG recordings (lasting 7–9 h) obtained from 20 subjects (four males and 16 females, 20–65 years old). Three EEG channels located in different lobes (Cz-A1, Fp1-A1 and O1-A1) were included in DRM-SUB, with a sampling rate of 200 Hz. To investigate the impact of the choice of EEG derivations on the performance of automatic sleep staging, EEG...
TABLE 1 | An overview of the application of deep networks on sleep staging.

| Authors          | Dataset | Channel      | Model     | Accuracy |
|------------------|---------|--------------|-----------|----------|
| Tsinalis et al.  | PSEE    | Fpz-Cz       | CNN       | 74.0%    |
| Phan et al.      | PSEE    | Fpz-Cz       | CNN       | 81.9%    |
| Supratak et al.  | PSEE    | Fpz-Cz       | CNN-BILSTM| 82.4%    |
| Hsu et al.       | PSE     | Fpz-Cz       | RNN       | 87.2%    |
| Phan et al.      | MASS    | C4-A1        | CNN       | 83.6%    |
| Supratak et al.  | MASS    | F4-EOG (Left)| CNN-BILSTM| 86.2%    |
| Sors et al.      | SHHS    | C4-A1        | CNN       | 87.0%    |

TABLE 2 | Data distribution of sleep stages in both datasets.

| Dataset | Total epochs | WAKE (%) | N1 (%) | N2 (%) | N3 (%) | REM (%) |
|---------|--------------|----------|--------|--------|--------|--------|
| PSEE    | 41663        | 19.2     | 6.6    | 42.2   | 13.4   | 18.5   |
| DRM-SUB | 20265        | 17.6     | 7.3    | 40.7   | 19.4   | 14.9   |

signals from all three channels were used separately for the following analysis.

Twenty healthy subjects (10 males and 10 females, 25–34 years old) from the PSEE dataset were also included. There are two EEG channels (Fpz-Cz and Pz-Oz) available in the PSEE dataset, with a sampling rate of 100 Hz. For each subject, two PSGs of about 20 h each were recorded during two subsequent day-night periods at the subjects’ homes. In order to remain consistent with previous studies (Supratak et al., 2017), for each subject and each PSG, only the data from 30 min before sleep-onset (i.e., the first sleep epoch after light-off in the evening) and 30 min after the last sleep epoch in the morning were included. Both channels were investigated separately.

For both datasets, labels of sleep staging for each 30-s EEG epoch were provided by the data distributors according to AASM rules. Five staging classes, i.e., WAKE, N1, N2, N3, and REM were used in this study. The distribution of 30-s EEG epochs of both datasets is illustrated in Table 2.

Construction of the AT-BiLSTM Network

The proposed AT-BiLSTM network architecture for automatic sleep staging is illustrated in Figure 1. It is composed of two main components, three stacked BiLSTM layers for feature exacting and one attention layer to weight the most relevant parts of the input sequence. According to a preset parameter, called the input dimension m, each raw 30-s EEG epoch is divided into multiple vectors, which are fed into the BiLSTM part sequentially to construct a feature matrix. Then to emphasize the different importance of different vectors, an attention layer is applied in the intra-epoch feature learning and summarizes the outputs of the BiLSTM part with different weights. Finally, the probability of each sleep stage can be derived from a fully connected (FC) layer and a softmax layer.

Given a 30-s EEG epoch \( X = [x_1, x_2, \ldots, x_N] \) with N data points, a moving window with input dimension of m is applied to X without overlap, leading to the matrix form of X, as shown in Equation 1, where \( n \) equals to \( N/m \) and \( X_t \) represents the vector in time step \( t \).

\[
\begin{pmatrix}
X_1 \\
X_2 \\
\vdots \\
X_t \\
\vdots \\
X_n \\
\end{pmatrix} = \begin{pmatrix}
x_1 & x_2 & \cdots & x_m \\
x_{m+1} & x_{m+2} & \cdots & x_{2m} \\
 \vdots & \vdots & \cdots & \vdots \\
x_{(t-1)m+1} & x_{(t-1)m+2} & \cdots & x_{tm} \\
 \vdots & \vdots & \cdots & \vdots \\
x_{(n-1)m+1} & x_{(n-1)m+2} & \cdots & x_{nm} \\
\end{pmatrix} \quad t \in [1, n] \tag{1}
\]

All the vectors are fed into the first BiLSTM layer, forward and backward respectively. For time step \( t \), the output of the forward or backward network, denoted as \( h^f_t \) or \( h^b_t \), can be obtained, respectively, according to Equation 2 or 3.

\[
h^f_t = \sigma(W_{fx}x_t + W_{fh}h^{f}_{t-1} + b_f) \tag{2}
\]

\[
h^b_t = \sigma(W_{bx}x_t + W_{bh}h^{b}_{t-1} + b_b) \tag{3}
\]

where \( \sigma \) is the logistic sigmoid function, \( W \) is the weight matrix (e.g., subscription “fx” in \( W \) represents the forward network of \( x_t \)) and \( b \) is the bias vector of the network (\( b_f \) and \( b_b \) represents the bias vector of forward and backward network, respectively).

The weighted sum of \( h^f_t \) and \( h^b_t \), denoted as \( h_t \), is computed as the output of the first BiLSTM layer following Equation 4.

\[
h_t = W_{fh}h^f_t + W_{bh}h^b_t + b_h \tag{4}
\]

The output of the previous BiLSTM layer is fed into the next layer in the same way. The third layer gives the final output of the BiLSTM part, which is weighted by the attention layer before feeding into the FC layer. Considering that EEG signal in different time steps should contribute differently to the classification task, it is rational to give strong weights to the more discriminative parts and vice versa. Formally, the attention weight \( a_t \) at the time step \( t \) is computed according to Formula (5) – (6).

\[
a_t = \frac{\exp(u_t^T u_w)}{\sum_t \exp(u_t^T u_w)} \tag{6}
\]

\[
u_t = \tanh(w_u h_t + b_u) \tag{5}
\]
FIGURE 1 | Illustration of the proposed AT-BiLSTM network architecture for automated sleep staging. The network consists of a BiLSTM part, an attention layer, a full-connected (FC) layer and a softmax layer. The input of the network is a raw 30-s EEG time series and the output is the probability of each sleep stage. The dashed rectangle on the EEG time series represents a vector of EEG signals at a time step.

In Formula (5)–(6), \( u_t \) represents the state of the hidden layer obtained from a simple neural network, \( u_w \) represents a weight vector randomly initialized, \( a_t \) represents the similarity between \( u_t \) and \( u_w \) obtained by softmax function.

\[
S_t = \sum_1^t a_t h_t
\]

(7)

By weighting and summing the output of the BiLSTM part, the attention vector, denoted as \( s_t \), can be obtained and fed into FC layer, preceding to the softmax layer which finally yields the probability of each sleep stage.

Construction of Baseline Networks
Apart from the proposed AT-BiLSTM network, we also constructed five baseline networks, including three single networks, i.e., CNN, LSTM and BiLSTM, and two hybrid networks, i.e., CNN-LSTM and CNN-BiLSTM.

Single Networks

Figure 2A illustrated the CNN topology used in this study, which is fed with a matrix reconstructed from a raw 30-s EEG epoch according to Equation 1. It consists of three convolution blocks and three max pooling layers. Each convolutional block contains a one-dimensional convolutional layer and a rectified linear unit (ReLU) activation layer. The input matrix is padded with zeros to ensure that the number of rows in the matrix is constant during the convolutional process. The output of CNN is fed into a FC layer, then activated by softmax function to obtain the sleep stage probability.

Two scenarios were considered in single RNN network. In the first scenario, three layers of LSTMs were stacked, also followed by a FC layer and a softmax layer. The second scenario employed stacked BiLSTM layers instead of the LSTM layers.

Hybrid Networks With CNN and RNN

As shown in Figure 2B,C, a CNN part followed by an RNN part was adopted in the hybrid networks, in order to make use of RNN for further processing the features extracted by CNN. The structures of the CNN part and RNN part are the same with the single networks aforementioned.

Datasets Splitting Strategy

Machine learning algorithms require independent training and test sets for model training and performance evaluation. Also, k-fold cross validation is preferred in application. Generally, there are two types of training data partitioning for clinic data: subject-wise and epoch-wise (Figure 3). For the subject-wise method, all the subjects were split into k folds equally and onefold is taken as the test set in turn while the remains as the training set. For the epoch-wise method, all the 30-s EEG epochs from all the subjects were merged and then split into k equal folds for each stage randomly. That is, for each sleep stage, all the 30-s EEG epochs from all the subjects were merged and then split into k equal folds for each stage randomly. Consequently, the epochs of a subject may appear in both the training and test set, violating the independence between the training and test set and contributing to a virtual high performance. Thus, in the present study, the subject-wise method with fivefold cross validation was adopted. The model was trained using the training set and evaluated using the test set. Finally, all evaluation results were combined.

Experimental Setting and Network Optimization

Using the first fold as the test set, the network parameters, such as the input dimension, the number of hidden units in each LSTM/BiLSTM/convolutional layer, and the filter/stride size of each convolutional layer and pooling layer, were determined by a grid-search to minimize the errors of networks with Python 3.6 and TensorFlow v1.15.0 (Abadi et al., 2016). The standard cross-entropy loss was used as the loss function in model training due to its good performance in measuring the errors of networks with
FIGURE 2 | Structure of the baseline networks for sleep staging: (A) the CNN network, (B) the CNN-LSTM network and (C) CNN-BiLSTM network. The CNN network consists of a CNN part, a full-connected layer and a softmax layer. In the CNN part, there are three convolution layers and three max pooling layers. Each convolution layer has 256 filters with a size of $7 \times 1$ each and each pooling layer has one filter of size $5 \times 1$. A rectified linear unit (ReLU) follows the convolution layer and precedes the pooling layer. The CNN part in panels (B,C) has the same topology with panel (A). For the LSTM/BiLSTM part, there are three stacked LSTM/BiLSTM layers with each layer consists of 256 memory cells. The target for all the networks was the probability of each sleep stage.

FIGURE 3 | Schematic diagram for the dataset splitting of training and test set: (A) subject-wise method; (B) epoch-wise method. For the subject-wise method, all the 30-s EEG epochs from a subject will be included in the training set or the test set as a whole while for the epoch-wise method, the epochs of a subject may appear in both the training and test set.

Each network was trained for 30 epochs with a mini batch size of 64 sequences. As a result, the input dimension $m$ was set as 5, the number of hidden units as 256, and the stride size for both convolution layers and max pooling layers as $1 \times 1$. The filter size of each convolutional layer and max pooling layer in CNN were set to $1 \times 7$ and $1 \times 5$ respectively.

For backpropagation, the adaptive moment estimation (ADAM) algorithm was adopted because it solves the optimization problem in non-stationary conditions and
works faster than the standard gradient descent algorithm and the root mean square propagation (Kingma and Ba, 2017). The main hyper-parameters used for ADAM algorithm were set as: learning rate (\(\alpha = 0.001\)), gradient decay factor (\(\beta_1 = 0.9\)), squared gradient decay factor (\(\beta_2 = 0.999\)), and epsilon (\(\varepsilon = 10^{-8}\)) for numerical stability. Moreover, a dropout layer before the last FC layer was used to avoid over-fitting and its dropout rate was set to 0.2, leading to 20% of the weights dropped during the training phase.

Performance Metrics

Overall metrics, including accuracy, macro F1-score (MF1) and Cohen’s kappa (\(\kappa\)) were used to evaluate the performance of each model. Performance on individual sleep stages was also assessed via class-wise precision and sensitivity.

Cohen’s kappa coefficient is a statistical measure of inter-rater agreement for categorical items (Cohen, 1960). When two persons (algorithms or raters) try to evaluate the same data, Cohen’s Kappa coefficient, \(\kappa\), is used as a measure of agreement between their decisions. In this study, it measures the amount of agreement between the output of the proposed algorithm and the provided labels of sleep stages.

Another metric used for performance evaluation here is the area under the receiver operating characteristics (ROC) curve, called AUC. The ROC curve is a graphical tool and demonstrates the classification performance by plotting the true positive rate (TPR) against the false positive rate (FPR) at different classification thresholds (Zweig and Campbell, 1993). Furthermore, it provides a convenient way for selecting the threshold that provides the maximum classification TPR while not exceeding a maximum allowable FPR level (Kim et al., 2019). For an \(n\)-class classification task, \(n\) ROC curves can be obtained by splitting the task into \(n\) binary classification tasks. For each binary classification task, its AUC value can be used as a class-wise measure of performance and the macro-average AUC of these tasks can be regarded as an overall metric for the performance evaluation.

RESULTS

Table 3 shows the overall performance of different networks on the PSEE dataset. The proposed AT-BiLSTM network outperforms other networks with overall accuracy, \(\kappa\), MF1 and MAUC of 83.78%, 0.766, 82.14% and 97.45% on channel Fpz-Cz, respectively and an overall accuracy, \(\kappa\), MF1 and MAUC of 80.79%, 0.731, 96.33% on channel Pz-Oz, respectively. The AT-BiLSTM network performs better than the other networks overall. For the single networks, the RNN-based networks outperform the CNN network while the results of BiLSTM and LSTM are comparable. The hybrid networks further improve the overall performance compared to the single models. Moreover, AT-BiLSTM achieves better precision and sensitivity on N3 and REM than the hybrid networks with CNN and RNN, although they have a comparable performance on stages Wake, N1 and N2. Furthermore, better performance is found in Fpz-Cz than Pz-Oz channel, regardless of the network topology used, indicating EEG derived from the frontal lobe is more valuable than those from the parietal lobe in sleep staging.

Table 4 shows the performance of different networks on the DRM-SUB dataset. The AT-BiLSTM network still outperforms other networks, suggesting its good generalization in sleep staging. Consistent with the results in PSEE dataset, the frontal EEG channel (Fp1-A1 here) achieves the best performance. The results are in line with a recent work, which found that EEG signals from an Fp1-A1 channel yielded higher accuracy values in automatic sleep staging than those of a Cz-A1 or O1-A1 channel (Ghimatgar et al., 2019).

Figure 4 illustrates the hypnograms labeled manually by a clinical technician of sleep and by the trained AT-BiLSTM model. The corresponding EEG recoding was obtained from the first person in PSEE dataset (SC4001E0), who spent 7 h during sleep. Noting that the subject is located in the test set for the trained model. The accuracy of the automatic sleep staging for this subject is 87.30%, showing considerable reliability of the proposed AT-BiLSTM network. Most of the wrong classifications were made during the transitions from one stage to another.

Table 5 shows the class-wise performances obtained on the PSEE dataset. For most stages, better performance is achieved by the AT-BiLSTM model than the baseline networks and Fpz-Cz channel outperforms the Pz-Oz one. Although classification accuracy of stage N1 is significantly lower than that of the other stages, which might due to the small percentage of N1 during sleep, it is higher than those reported in previous studies (Hsu et al., 2013; Supratak et al., 2017). Similar findings can be found on the DRM-SUB (Table 6).

Furthermore, ROC curves were used to compare the performances of the proposed AT-BiLSTM model
for different sleep stages with the frontal channels in both datasets (Figure 5). As shown in Figure 5, AT-BiLSTM is sufficient to identify WAKE, N3 and REM, but insufficient to identify N1.

Table 7 illustrates the results of a comparison between the proposed AT-BiLSTM model and the state-of-the-art works using the same dataset of DRM-SUB (Hassan and Bhuiyan, 2016a,b; Hassan and Subasi, 2017; Ghimatgar et al., 2019; Shen et al., 2019). With the same dataset, same EEG channel and same dataset splitting strategy, the proposed AT-BiLSTM model achieves the highest accuracy.

**DISCUSSION**

In this study, we proposed an AT-BiLSTM network for automatic sleep staging with single-channel EEG. The main findings were: (1) the frontal EEG derivations contribute to better
### TABLE 6 | The class-wise performance obtained on the DRM-SUB dataset (value in bold represents for the best among all the networks).

| EEG signal | Networks       | Precision | Sensitivity | Class-wise AUC |
|------------|----------------|-----------|-------------|----------------|
|            |                | W  N1 N2 N3 REM | W  N1 N2 N3 REM | W  N1 N2 N3 REM |
| Fp1-A1     | AT-BiLSTM      | 88.48 45.92 84.98 89.08 68.26 | 89.34 25.88 85.12 85.18 83.06 | 99.50 88.80 93.05 96.69 96.93 |
|            | CNN            | 83.54 40.82 79.76 87.14 63.02 | 84.22 11.54 85.92 78.52 78.96 | 97.85 82.77 87.05 86.85 94.96 |
|            | LSTM           | 83.63 44.16 82.58 88.26 68.14 | 89.92 19.18 84.82 84.74 75.28 | 99.35 87.06 92.33 96.40 96.33 |
|            | BLSTM          | 85.24 43.82 83.28 86.82 66.28 | 86.78 17.36 85.31 86.83 77.74 | 98.90 85.83 92.01 96.39 96.07 |
|            | CNN-LSTM       | 87.12 42.26 84.52 86.26 66.44 | 87.04 23.34 84.52 87.66 79.91 | 99.26 88.22 91.94 96.61 96.13 |
|            | CNN-BiLSTM     | 86.62 46.92 83.88 86.74 66.38 | 89.98 20.74 84.54 85.88 79.86 | 99.21 86.30 92.04 96.28 95.54 |
| Cz-A1      | AT-BiLSTM      | 88.02 42.02 84.98 87.56 69.22 | 90.96 22.22 86.32 86.28 79.91 | 99.40 89.24 93.62 97.10 96.87 |
|            | CNN            | 83.68 23.44 75.18 86.14 65.12 | 86.55 8.92 85.36 73.28 72.68 | 98.18 82.81 88.39 90.85 94.77 |
|            | LSTM           | 87.24 43.88 83.44 87.76 69.26 | 89.54 22.34 86.12 86.26 77.04 | 99.14 88.59 92.39 97.06 96.09 |
|            | BiLSTM         | 86.46 39.68 83.82 88.58 67.12 | 89.24 20.62 85.36 83.38 78.42 | 99.27 88.56 92.57 96.77 96.14 |
|            | CNN-LSTM       | 84.34 48.72 81.44 89.58 70.64 | 91.34 14.16 87.66 82.46 78.06 | 98.99 88.43 92.42 96.49 96.15 |
|            | CNN-BiLSTM     | 86.22 42.96 82.32 87.98 70.88 | 90.36 23.68 86.78 84.26 76.58 | 99.30 88.68 92.48 96.66 96.29 |
| O1-A1      | AT-BiLSTM      | 88.86 46.58 78.76 83.20 62.56 | 91.36 18.53 81.94 78.67 72.74 | 99.55 89.76 92.40 96.67 96.18 |
|            | CNN            | 86.21 38.78 73.48 82.92 52.02 | 84.38 7.56 80.62 72.40 65.38 | 98.25 82.82 87.94 92.30 91.63 |
|            | LSTM           | 83.66 28.76 74.96 83.98 55.44 | 90.44 12.12 81.88 75.96 62.66 | 99.17 84.52 91.31 96.09 93.75 |
|            | BiLSTM         | 89.92 39.72 75.06 80.52 56.02 | 87.72 13.14 73.94 81.02 62.96 | 99.24 84.52 91.28 95.89 93.33 |
|            | CNN-LSTM       | 90.52 42.81 76.32 82.25 59.18 | 88.76 13.56 82.34 78.86 68.73 | 99.23 87.46 91.39 95.13 95.46 |
|            | CNN-BiLSTM     | 88.04 45.96 76.26 82.44 60.84 | 89.92 11.72 82.18 79.62 67.68 | 99.26 86.43 91.41 96.40 95.44 |

### FIGURE 5 | ROC curves for sleep stages using the proposed AT-BiLSTM models trained with (A) Fpz-Cz channel of PSEE dataset and (B) Fp1-A1 channel of DRM-SUB dataset.

### TABLE 7 | Comparison of sleep staging performance on the DRM-SUB dataset between the proposed method and previous works based on conventional feature extraction.

| Authors                  | Year  | Methodology                                                                                     | Dataset splitting strategy | Channel     | Accuracy  |
|--------------------------|-------|-------------------------------------------------------------------------------------------------|---------------------------|-------------|-----------|
| Hassan and Bhuyan, 2016a | 2016  | Tunable Q-factor wavelet transform, random forest (Hassan and Bhuyan, 2016a)                    | Epoch-wise                | Fp1-A1      | 72.28%    |
| Hassan and Bhuyan, 2016b | 2016  | Implementation of ensemble empirical mode decomposition in conjunction with random under sampling boosting (Hassan and Bhuyan, 2016a) | Epoch-wise                | Fp1-A1      | 74.99%    |
| Hassan and Subasi, 2017  | 2017  | Tunable Q-factor wavelet transform, bagging (Hassan and Subasi, 2017)                            | Epoch-wise                | Fp1-A1      | 78.95%    |
| Shen et al.              | 2019  | Essence features extraction method (Shen et al., 2019)                                           | Subject-wise              | Cz-A1       | 80.90%    |
| Ghimatgar et al.          | 2019  | Features in time domain, frequency domain, cepstral domain, wavelet features, autoregressive coefficients and non-linear features with Hidden Markov Model (Ghimatgar et al., 2019) | Subject-wise              | Fp1-A1      | 81.22%    |
| Proposed method          |       | Raw EEG signal and AT-BiLSTM                                                                    | Subject-wise              | Fp1-A1      | 81.72%    |
Two EEG datasets, i.e., PSEE and DRM-SUB, with different EEG derivations were used in our study. To clarify the influence of the EEG channel on automatic sleep staging, here we applied the proposed method to all the EEG channels in both datasets. The results obtained from both datasets are similar: the model adopting frontal derivation behaved better than those from other lobes. Such a finding indicated that the performance of sleep scoring was sensitive to the selection of EEG channel and the derivations from the frontal region are the optimal choices. Physiologically, the prefrontal cortex is deactivated and reactivated during the sleep cycle, indicating its involvement in the wake–sleep cycle (Maquet et al., 1996). With the development of wearable EEG devices, EEG signals can be easily obtained using dry electrodes on the forehead (Hassan and Bhuiyan, 2016a); the proposed method would be promising in supporting people monitoring sleep.

In recent years, many automated sleep staging methods based on deep neural networks used CNNs for feature extraction and RNNs to capture temporal information. These approaches have significantly improved the accuracy of sleep staging (Hassan and Bhuiyan, 2016a; Boostani et al., 2017; Sors et al., 2018). In general, for the sequence-to-label model based on RNN, only the output vector at the last time step is retained for classification, e.g., via a softmax layer (Phan et al., 2017). However, it is reasonable to combine the output vectors of different time steps by some weighting schemes. Intuitively, those parts of the input sequence which are essential to the classification task at hand should be associated with strong weights, and those with less importance should be weighted correspondingly less. Ideally, these weights should be automatically learned by the network. This can be accomplished with an attention layer (Luong et al., 2015). Besides, previous works demonstrated that the performance of classification or regression can be further improved by stacking multiple BiLSTM in neural networks (Liu et al., 2017; Wang et al., 2018; Liu et al., 2018). Aside from that, we found the overall performance of the RNN based model to be better than that of the CNN models in automatic sleep staging, which might indicate that the RNNs are promising in capturing the temporal nature of an EEG time series. From such a perspective, the highest performance achieved by the proposed AT-BiLSTM might further confirm the role of stacking layers and attention mechanism in feature extracting of time series.

In this study, all experiments were performed on a server configured with 64 CPUs [Intel(R) Xeon(R) CPU @ 2.10 GHz], 64 GB memory, a GPU (NVIDIA GeForce GTX 1,080 Ti) and a Windows Server 2016 system. A CNN network has the lowest computational cost as its training time for each batch was 0.16 s on average. LSTM and CNN-LSTM networks take similar times (8.46 and 8.60 s respectively) for each batch in training. The computational cost of BiLSTM based networks is twice that of LSTM based networks because they must calculate the input sequence in two directions and set up double parameters. Moreover, approximately 1.3 s more is required for each batch with the attention layer.

Our study demonstrated that a deep learning approach without manual feature extraction can also provide sufficient accuracy for sleep staging, which is even better than conventional methods based on manual feature extraction. Therefore, the proposed method is a promising choice for computer-aided detection of sleep stages and similar 1-D signal classification problems. In conclusion, our findings provide a possible solution for automatic sleep scoring without manual signal preprocessing and feature extraction. With the development of wearable EEG devices, such a solution would be valuable in the screening of sleep disorders at home for the general population.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: Dreams Subjects: https://zenodo.org/record/2650142#.X6tbymgzZdg. Sleep-EDF Database Expanded: https://physionet.org/content/sleep-edfx/1.0.0/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the institutional review board of two open-access datasets, i.e., the Sleep-EDF Expanded dataset available at Physionet and the DREAMS Subjects dataset. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

FH, XL, FX, and JL designed this study. MF and YW analyzed the data. MF, FH, and ZC wrote the article. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the National Natural Science Foundation of China (Grant No. 11974231) and the Double First-Class University Project of China.
Ronzhina, M., Janousek, O., Kolárová, J., and Nováková, M. (2012). Sleep scoring using artificial neural networks. Sleep Med. Rev. 16, 251–263. doi: 10.1016/j.smrv.2011.06.003

Schuster, M., and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Trans. Signal. Process. 45, 2673–2681. doi: 10.1109/78.650093

Seifpour, S., Niknazar, H., Mikaeili, M., and Nasrabadi, A. M. (2018). A new automatic sleep staging system based on statistical behavior of local extrema using single channel EEG signal. Expert Syst. Appl. 104, 277–293. doi: 10.1016/j.eswa.2018.03.020

Shen, H. M., Xu, M. H., Guez, A., and Li, A. (2019). An accurate sleep stages classification method based on state space model. IEEE Access. 4, 1–12. doi: 10.1109/ACCESS.2019.2939038

Silveira, T. L., Kozakevicius, A. J., and Rodrigues, C. R. (2017). Single-channel EEG sleep stage classification based on a streamlined set of statistical features in wavelet domain. Med. Biol. Eng. Comput. 55, 334–352. doi: 10.1007/s11517-016-1519-4

Sors, A., Bonnet, S., Mirek, S., and Vercueil, L. (2018). A convolutional neural network for sleep stage scoring from raw single-channel EEG. Biomed. Signal. Process. Proc. 42, 107–114. doi: 10.1016/j.bspc.2017.12.001

Stephansen, J. B., Olesen, A. N., Olsen, M., Ambati, A., Leary, E. B., Moore, H. E., et al. (2018). Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. Nat. Commun. 9, 1. doi: 10.1038/s41467-018-07229-3

Supratak, A., Dong, H., Wu, C., and Guo, Y. (2017). “DeepSleepNet: a model for automatic sleep scoring based on raw single-channel EEG,” in IEEE Transactions on Neural Systems and Rehabilitation Engineering, (Piscataway, NJ: IEEE), 99. doi: 10.1109/TNSRE.2017.2721116

Tian, P., Hu, J., Qi, J., Ye, X., Che, D., Ding, Y., et al. (2017). A hierarchical classification method for automatic deep sleep scoring using multiscale entropy features and proportion information of sleep architecture. Biocybern. Biomed. Eng. 37, 263–271. doi: 10.1016/j.bbe.2017.01.005

Tripathy, R., and Acharya, U. R. (2018). Use of features from RR-time series and EEG signals for automated classification of sleep stages in deep neural network framework. Biocybern. Biomed. Eng. 38, 890–902. doi: 10.1016/j.bbe.2018.05.005

Tsinalis, O., Matthews, P. M., and Guo, Y. (2015). Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders. Ann. Biomed. Eng. 44, 1587–1597. doi: 10.1007/s10439-015-1444-y

Tsinalis, O., Matthews, P. M., Guo, Y., and Zafeiriou, S. (2016). Automatic sleep stage scoring with single-channel EEG using convolutional neural networks. arXiv [Preprint]. Available online at: https://arxiv.org/abs/1610.01683 (accessed May 10, 2020).

Wang, C., Yang, H., and Meinel, C. (2018). Image captioning with deep bidirectional LSTMs and multi-task learning. ACM Trans. Multim. Comput. 14, 1–20. doi: 10.1145/3113432

Wang, Y., Loparo, K. A., Kelly, M. R., and Kaplan, R. F. (2015). Evaluation of an automated single-channel sleep staging algorithm. Nat. Sci. Sleep 7, 101–111. doi: 10.2147/NSS.S77888

Weaver, E. M., Woodson, B. T., and Steward, D. L. (2005). Polysomnography indexes are discordant with quality of life, symptoms, and reaction times in sleep apnea patients. Otolaryngol. Head Neck Surg. 132, 255–262. doi: 10.1016/j.otohns.2004.11.001

Yildirim, O., Baloglu, U. B., and Acharya, U. R. (2019). A deep learning model for automated sleep stages classification using PSG signals. Int. J. Environ. Res. Public Health 16, 599–599. doi: 10.3390/ijerph16040599

Younes, M. (2017). The case for using digital EEG analysis in clinical sleep medicine. Sleep Sci. Pract. 1.2. doi: 10.1186/s41606-016-0005-0

Zhang, J., and Wu, Y. (2017). A new method for automatic sleep stage classification. IEEE Trans. Biomed. Circuits Syst. 5, 1097–1110. doi: 10.1109/TBCAS.2017.2719631

Zhang, J., and Wu, Y. (2018). Complex-valued unsupervised convolutional neural networks for sleep stage classification. Comput. Methods Prog. Biomed. 164, 181–191. doi: 10.1016/j.cmpb.2018.07.015

Zhang, X., Liang, X., Zhiyuli, A., and Zhang, S. (2019). AT-LSTM: an attention-based LSTM model for financial time series prediction. IOP Conf. Ser. Mater. Sci. Eng. 569,052037. doi: 10.1088/1757-899X/569/5/052037

Zhu, G., Li, Y., and Wen, P. P. (2014). Analysis and classification of sleep stages based on difference visibility graphs from a single-channel EEG signal. IEEE J. Biomed. Health 18, 1813–1821. doi: 10.1109/JBHI.2014.2303991

Zweig, M. H., and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin. Chem. 39, 561–577. doi: 10.1093/clinchem/39.4.561

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Fu, Wang, Chen, Li, Xu, Liu and Hou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.