# Multivariate Sleep Stage Classification using Hybrid Self-Attentive Deep Learning Networks

Ye Yuan$^{*\dagger}$, Kebin Jia$^{*\dagger}$, Fenglong Ma$^{\ddagger}$, Guangxu Xun$^{\ddagger}$, Yaqing Wang$^{\ddagger}$, Lu Su$^{\ddagger}$ and Aidong Zhang$^{\ddagger}$
$^*$College of Information and Communication Engineering, Beijing University of Technology, Beijing, China
Email: yuanye91@emails.bjut.edu.cn, kebinj@bjut.edu.cn
$^\dagger$Beijing Key Laboratory of Computational Intelligence and Intelligent System
Beijing University of Technology, Beijing, China
$^\ddagger$Department of Computer Science and Engineering, State University of New York at Buffalo, NY, USA
Email: fenglong,guangxux,yaqingwa,lusu,azhang@buffalo.edu

***Abstract*—Recently, significant efforts have been made to explore comprehensive sleep monitoring to prevent sleep-related disorders. Multivariate sleep stage classification has garnered great interest among researchers in health informatics. In this paper, we propose HybridAtt, a unified hybrid self-attentive deep learning network, to classify sleep stages from multivariate polysomnography (PSG) records. HybridAtt is an end-to-end model that explicitly captures the complex correlations among biomedical channels and the dynamic relationships over time. By constructing a new multi-view convolutional representation module, HybridAtt is able to extract hidden features from both channel-specific and global views of the heterogeneous PSG inputs. In order to enhance feature representation, a new fusion-based attention mechanism is also proposed to integrate the complementary information carried by each feature view. To evaluate the performance of our model, we carry out experiments on a benchmark PSG dataset. Experimental results show that the proposed HybridAtt model achieves better performance compared to ten baseline methods, demonstrating the effectiveness of HybridAtt in the task of sleep stage classification.**

## I. Introduction

Sleep is a natural resting state of body and mind, which covers around one-third of human lifespan. Due to the increasing pressures of work and unhealthy lifestyle, sleep disturbances become one of the serious health problems in modern societies. In order to conduct comprehensive physiological monitoring, overnight polysomnography (PSG) recordings are often utilized to analyze complex physiologic events during sleep [1]. PSG can be represented as heterogeneous multivariate time series which includes various physiological measurements used to monitor different body functions. In clinical practice, the collected PSG records are segmented into 30-second slots and visually inspected by well-trained experts. However, long term PSG visual inspection is extremely time-consuming and laborious for physicians, and requires highly-trained professionals to diagnose sleep issues. Thus, it has motivated researchers to develop automatic sleep stage classification systems that can efficiently perform PSG sleep analysis.

Recently, a variety of methods has been investigated for the classification of sleep stages using the PSG data [2], [3]. To aggregate and analyze the multivariate PSG records, traditionally, several researchers propose to combine handcrafted features with a classifier to build a multi-stage sleep stage classification system. More recently, in order to automatically learn meaningful representations for such multivariate biosignals, significant efforts have been made to explore feature extraction techniques using deep learning methods [4]–[6]. In the task of sleep stage classification, the features extracted by deep learning models, such as deep belief networks (DBN) [7], [8], convolutional neural networks (CNN) [9]–[11] and recurrent neural networks (RNN) [12], [13], have proven to be more robust than the handcrafted features due to better classification performance.

Despite many deep learning studies reporting promising results in sleep stage classification, some challenges still need to be addressed. One of the major challenges is that most deep learning models fail to explicitly incorporate the inherent correlations of multivariate biosignals. On one hand, there exist complex correlations among PSG channels which should be captured to identify sleep stages. On the other hand, the dynamic correlations among the data across different timestamps (i.e., slots) are also crucial to capturing sleep-related events. Moreover, the hidden patterns during sleep vary significantly across individuals, rendering it a challenging task to develop a cross-subject sleep stage classifier.

To tackle the aforementioned challenges, we propose a hybrid self-attentive deep learning network (HybridAtt) to classify sleep stages from multivariate PSG records. The framework of HybridAtt is presented in Fig. 1. Specifically, to learn informative features from the heterogeneous PSG inputs, we first construct a multi-view convolutional encoder to extract features from both channel-specific and global perspectives, referred to as channel-view and global-view features, respectively. Based on the learned multi-view features, we then develop a new fusion-based hybrid attention mechanism, which consists of a channel-wise attention layer and a time-wise attention layer, to model the dual correlations of PSG channels and timestamps. Finally, we adopt a softmax layer using the obtained attentional hidden representation to train our proposed end-to-end deep learning model as a cross-subject classifier. We summarize the main contributions of this paper as follows:

- We propose HybridAtt, a unified hybrid self-attentive deep learning network, to learn informative representa-

978-1-5386-5488-0/18/$31.00 ©2018 IEEE

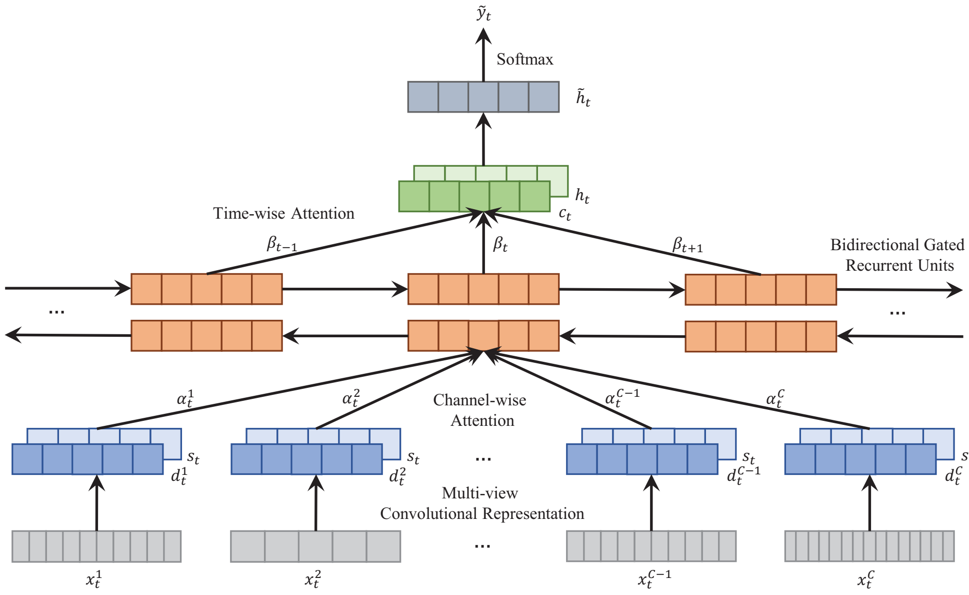

Fig. 1. Schematic illustration of the overall approach pipeline.

tions of multivariate PSG records for the task of cross-subject sleep stage classification. HybridAtt explicitly incorporates the dual correlations of PSG biomedical channels and timestamps.

- We propose a fusion-based attention mechanism which is able to combine the complementary information carried by the learned multi-view convolutional features, and hence can generate discriminative representations from the multivariate PSG data.
- We empirically show that the proposed HybridAtt outperforms ten baseline methods on a benchmark dataset.

## II. Methodology

In this section, we introduce the methodology of our proposed HybridAtt model with multivariate PSG inputs. We discuss the details of the main components in the following subsections.

### A. *Multi-view Convolutional Representation*

In practice, the collected PSG data often tend to be heterogeneous, referred to different sample rates, signal strengths, and rhythm patterns. In order to preserve the unique characteristics of each biomedical channel during the feature representation, the multi-view deep learning strategy [14] can be employed. The benefit of adopting multi-view deep learning is to extract features from multiple perspectives using different deep learning structure to improve the generalization performance. It has proven to be effective for several tasks including human activity recognition [15], EEG seizure detection [16], [17], and 3-D shape recognition [18]. Following the previous studies, in our model, we further extend this strategy by modifying the CNN structure to learn latent representations from channel-specific and global views, i.e., the channel-view hidden features $\boldsymbol{d}_t^{1:C}$ and global-view hidden features $\boldsymbol{s}_t$, respectively. Formally, given the input $\boldsymbol{x}_t^c \in \mathbb{R}^{n^{(c)}}$ in the $c$-th channel at timestamp $t$, we can obtain its channel-view representation $\boldsymbol{d}_t^c \in \mathbb{R}^p$ through a 1-D channel-CNN encoder, denoted as $\mathrm{CNN}_c$, as follows:

$$\boldsymbol{d}_t^c = \mathrm{CNN}_c(\boldsymbol{x}_t^c; \boldsymbol{\theta_c}), \tag{1}$$

where $\boldsymbol{\theta_c}$ denotes all the learnable parameters of $\mathrm{CNN}_c$. Similarly, the global-view representation $\boldsymbol{s}_t \in \mathbb{R}^p$ can be calculated through a 2-D global-CNN encoder (i.e., $\mathrm{CNN}_g$), as follows:

$$\boldsymbol{h}_g = \mathrm{CNN}_g(\boldsymbol{x}_t^{1:C}; \boldsymbol{\theta_g}), \tag{2}$$

where $\boldsymbol{\theta_g}$ denotes all the learnable parameters of $\mathrm{CNN}_g$. Here in Eq. (2), to derive a unified matrix input, we adopt linear interpolation to align the input vector of each channel into same dimension.

In general, both the channel- and global-CNN encoders can be parameterized by a series of convolutional-nonlinear-pooling cells with several filter kernels. Instead of adopting standard CNN, in our model, we construct a new CNN structure to unleash the power of multi-view feature extraction shown in Fig. 2. In particular, on one hand, we attempt to convolve in parallel different sizes of feature kernels to handle multiple object scales. In this way, the feature learning module is able to cover a big area while keeps fine resolutions for small patterns in biosignals. On the other hand, we adopt max pooling and average pooling for $\mathrm{CNN}_c$ and $\mathrm{CNN}_g$, respectively. The idea here is to guide $\mathrm{CNN}_c$ to focus on the most important features of different channels while let $\mathrm{CNN}_g$ retain more general information among all the channels. Taking the advantage of multi-view structure, we can not only learn informative representations from PSG data, but also uniform the heterogeneous inputs for the following hybrid attention module. Note that the dimension of each cell relies on the configurations of CNN, which is given in Section 3.2.

### *B. Channel-wise Attention*

In the task of PSG-based sleep stage classification, there exist complex correlations among PSG channels. An ideal approach is to dynamically qualify the importance of information carried by each biomedical channel and relay on more informative ones to achieve better performance. Towards this end, we develop a new fusion-based channel-wise attention mechanism to adaptively capture the complex channel correlations from PSG data.

Given the multi-view features $\boldsymbol{d}_t^c$ and $\boldsymbol{s}_t$ obtained by Eq. (1) and Eq. (2), we propose to calculate a fusional rate $r_t^c \in \mathbb{R}$ for each channel $c$ at timestamp $t$, defined as:

$$r_t^c = \sigma(\boldsymbol{W}_{rg}^\top \boldsymbol{s}_t + \boldsymbol{W}_{rc}^\top \boldsymbol{d}_t^c + b_{rc}), \tag{3}$$

where $\boldsymbol{W}_{rg} \in \mathbb{R}^p$, $\boldsymbol{W}_{rc} \in \mathbb{R}^p$, and $b_{rc} \in \mathbb{R}$ are the parameters to be learned. Here in Eq. (3), we use the sigmoid function $\sigma(\cdot)$ to rescale the fusional rate to be in the range of $[0, 1]$, representing how much information carried by each CNN encoder should be fused. Based on the fusional rate, we can assign the attention energy $e_t^{g,c}$ for each channel $c$ using the integrated information, as follows:

$$e_t^{g,c} = \boldsymbol{W}_{ec}^\top((1 - r_t^c) \odot \boldsymbol{s}_t + r_t^c \odot \boldsymbol{d}_t^c) + b_{ec}, \tag{4}$$

where $\odot$ denotes the element-wise multiplication operator, $\boldsymbol{W}_{ec} \in \mathbb{R}^p$ and $b_{ec} \in \mathbb{R}$ are the weight vector and bias value, respectively. Then, we can derive the contribution score vector $\boldsymbol{\alpha}_t$ normalized by the softmax function, as follows:

$$\boldsymbol{\alpha}_t = \text{Softmax}([e_t^{g,1}, \cdots, e_t^{g,c}, \cdots, e_t^{g,C}]). \tag{5}$$

Finally, the output vector of the channel-wise attention $\tilde{\boldsymbol{x}}_t \in \mathbb{R}^{2p}$ can be calculated according to the contribution score vector $\boldsymbol{\alpha}_t$ using weighted aggregation:

$$\tilde{\boldsymbol{x}}_t = \boldsymbol{s}_t \oplus (\sum_{c=1}^{C} \alpha_t^{g,c} \odot \boldsymbol{d}_t^c), \tag{6}$$

where $\oplus$ is the concatenation operator. In this way, the proposed fusion-based attention mechanism can fully utilize the multi-view information carried by both two feature views, and thus generate more informative representations from multivariate PSG data.

### *C. Time-wise Attention*

In order to capture the dependencies of different timestamps, the aforementioned fusion-based attention mechanism can also be adopted in the time dimension, referred to as time-wise attention. Specifically, given the learned vector sequence from $\tilde{\boldsymbol{x}}_1$ to $\tilde{\boldsymbol{x}}_T$, we can obtain the hidden state $\boldsymbol{h}_t \in \mathbb{R}^{2q}$ through a 2-layer BGRU [19], as follows:

$$\boldsymbol{h}_{1:T} = \text{BGRU}(\tilde{\boldsymbol{x}}_{1:T}; \boldsymbol{\theta}_r), \tag{7}$$

where $\boldsymbol{\theta}_r$ denotes all the parameters of BGRU. The extracted hidden state $\boldsymbol{h}_t$ is the concatenation of both forward and backward hidden vectors, denoted as $\overrightarrow{\boldsymbol{h}}_t$, $\overleftarrow{\boldsymbol{h}}_t \in \mathbb{R}^q$, respectively.

To calculate the time-wise contribution score vector $\boldsymbol{\beta}_t$, we can reformalize the fusion-based attention mechanism from Eq. (3) to Eq. (5) as follows:

$$r_i = \sigma(\boldsymbol{W}_{rt}^\top \boldsymbol{h}_t + \boldsymbol{W}_{ri}^\top \boldsymbol{h}_i + b_{rt}),$$

$$e_{t,i} = \boldsymbol{W}_{et}^\top((1 - r_i) \odot \boldsymbol{h}_t + r_i \odot \boldsymbol{h}_i) + b_{et},$$

$$\boldsymbol{\beta}_t = \text{Softmax}([e_{t,1}, \cdots, e_{t,i}, \cdots, e_{t,T}]),$$

where $\boldsymbol{W}_{rt} \in \mathbb{R}^{2q}$, $\boldsymbol{W}_{ri} \in \mathbb{R}^{2q}$, $b_{rt} \in \mathbb{R}$, $\boldsymbol{W}_{et} \in \mathbb{R}^{2q}$, and $b_{et} \in \mathbb{R}$ are the learnable parameters. Subsequently, we derive a temporal context vector $\boldsymbol{c}_t \in \mathbb{R}^{2q}$ as the output of the time-wise attention:

$$\boldsymbol{c}_t = \sum_{i=1}^{T} \beta_{t,i} \odot \boldsymbol{h}_t. \tag{8}$$

### *D. Unified Training Procedure*

Given the context vector calculated by Eq. (8), we combine it with the current hidden state to derive an attentional representation $\hat{\boldsymbol{h}} \in \mathbb{R}^r$, defined as:

$$\hat{\boldsymbol{h}}_t = f(\boldsymbol{W}_h[\boldsymbol{c}_t \oplus \boldsymbol{h}_t] + \boldsymbol{b}_h),$$

where $\boldsymbol{W}_h \in \mathbb{R}^{r \times 4q}$ and $\boldsymbol{b}_h \in \mathbb{R}^r$ denote the learnable parameters. The attentional representation is then fed through the softmax layer for the final task of sleep stage classification, as follows:

$$\hat{\boldsymbol{y}}_t = \text{Softmax}(\boldsymbol{W}_s \hat{\boldsymbol{h}}_t + \boldsymbol{b}_s). \tag{9}$$

where $\boldsymbol{W}_s \in \mathbb{R}^{|\mathcal{C}| \times r}$ and $\boldsymbol{b}_s \in \mathbb{R}^{|\mathcal{C}|}$ are the learnable weight matrix and bias vector, respectively.

Each record $\boldsymbol{X}_t$ at timestamp $t$ contains a set of $C$-channel heterogeneous waveform vectors $\{\boldsymbol{x}_t^1, \boldsymbol{x}_t^2, \cdots, \boldsymbol{x}_t^C\}$ where $\boldsymbol{x}_t^c \in \mathbb{R}^{n^{(c)}}$.

To train a unified model, we adopt cross-entropy to measure the loss between the ground truth $\boldsymbol{y}_t$ and the $\hat{\boldsymbol{y}}_t$ obtained by Eq. 9. Formally, the final cost function of our end-to-end HybridAtt model is defined as:

$$J_{\text{HybridAtt}}(\boldsymbol{X}_1^{(1)}, \cdots, \boldsymbol{X}_{T^{(1)}}^{(1)}, \cdots, \boldsymbol{X}_1^{(M)}, \cdots, \boldsymbol{X}_{T^{(M)}}^{(M)})$$
$$= -\frac{1}{M} \sum_{i=1}^{M} \frac{1}{T^{(i)}} \sum_{t=1}^{T^{(i)}} \left[\boldsymbol{y}_t^\top \log \hat{\boldsymbol{y}}_t + (\boldsymbol{1} - \boldsymbol{y}_t)^\top \log(\boldsymbol{1} - \hat{\boldsymbol{y}}_t)\right].$$

## III. Experiments

### *A. Dataset Description*

The multivariate sleep dataset we use is the UCD dataset provided by St. Vincents University Hospital and University College Dublin. The UCD dataset is an open access dataset and can be download from the PhysioNet [20]. This dataset contains 14-channel multivariate PSG data collected from adult subjects, including electroencephalogram (EEG) at 128Hz, electromyogram (EMG) at 64Hz, electrooculogram (EOG) at 64Hz, and other biosignals related to patient movement, posture and breathing. In addition, according to the standard Rechtschaffen and Kales (R&K) rules [21], each 30-second

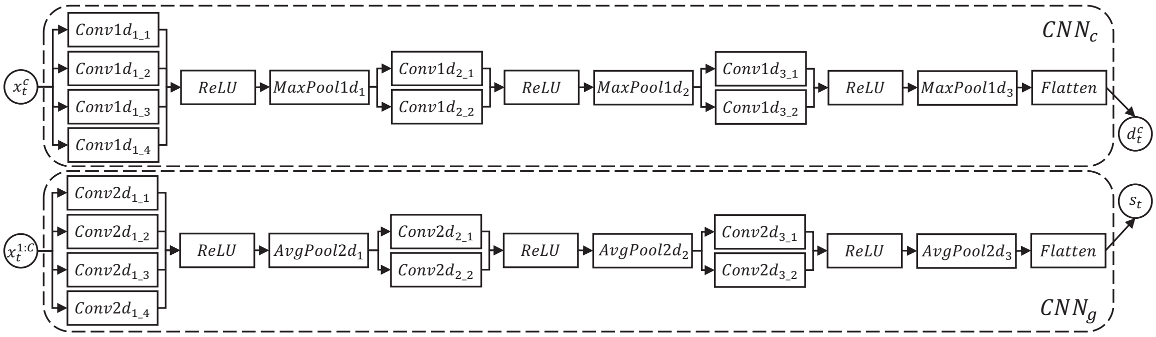

Fig. 2. CNN Structure of the multi-view convolutional representation module in HybridAtt.

slot is labeled as in one of the five sleep stages: W, REM, S1, S2, and S3 (merged from S3 and S4). Different from previous work that selected specific PSG channels [7], [22] or subject groups [8] using prior knowledge, we generate $287,840$ waveform vectors from all the 25 subjects and feed all the channels into our model. As the original 30-second long data contain 14 biomedical channels, we adopt short-time Fourier transform (STFT) using the blackman window as preprocessing, and derive the input slot with $27,300$ data points. Note that we only retain the slots belonging to the five sleep stages in our experiments.

### B. Experiment setup

*1) Baseline Approaches:* To validate the effectiveness of the proposed model, we compare it with several widely used biosignal feature learning baseline methods. We select the following ten existing approaches as baselines:

*Support vector machine (SVM) [23].* SVM is one of the most popular machine learning baselines. Since the standard linear SVM model is a binary classifier, we use one-vs-all SVM for the task of multi-class sleep stage classification. To avoid the curse of dimensionality, we utilize principal component analysis (PCA) to extract the features from all the channels, and then select top-$r$ related components as features to train the SVM model, referred to as PSVM.

*Deep Neural Networks (DNN) [24].* DNN is a commonly used baseline for deep learning. We first concatenate the waveform vectors of all the channels together, and then feed them into a 3-layer DNN with softmax.

*RNN.* RNN is another widely used deep learning baseline. Similar to DNN, we feed the concatenated vectors into the BGRU. The hidden representations produced by the BGRU are directly used for training using softmax.

*RNNAtt.* We incorporate attention mechanism into RNN. After the BGRU outputs the hidden vectors $\boldsymbol{h}_{1:T}$, RNNAtt adopts attention mechanism to obtain a context vector $\boldsymbol{c}_t$. Then, RNNAtt concatenates both $\boldsymbol{c}_t$ and $\boldsymbol{h}_t$ as an attentional representation for final training. Two existing strategies, namely location-based and concatenation-based attention [25], are employed, referred to as RNNAtt$_l$ and RNNAtt$_c$, respectively.

*CNN.* We first integrate the inputs from all the channels as a matrix, and then extract features through the multi-view convolutional architecture shown in Fig. 2. The learned hidden representations are directly used for training.

*CRNN.* CRNN is a RNN variant combined with a CNN, which is widely used in several view-related tasks. We first derive the multi-view features using the aforementioned CNN, then feed the representations to BGRU to train an end-to-end model.

*CRNNAtt.* CRNNAtt employs attention mechanism after the feature extraction of CRNN. Similarly, we perform the same process as RNNAtt, referred to as CRNNAtt$_l$ and CRNNAtt$_c$, respectively.

*ChannelAtt [17].* ChannelAtt focuses on soft-selecting critical channels from multivariate biosignals. Compared to RNNAtt, ChannelAtt adopts a new global attention mechanism in the channel domain instead of the time domain. Different from the original model using fully-connected layer for feature extraction, we use the proposed CNN structure as the feature encoder to train the model.

*2) Our Approaches:* We show the performance of the following two approaches in the experiments.

*HybridAtt$_l$.* HybridAtt$_l$ is a reduced model that employs the location-based attention mechanism in Hybrid to classify sleep stages.

*HybridAtt$_f$.* This model uses the proposed fusion-based attention mechanism when calculating both the channel-wise and time-wise score vectors.

*3) Evaluation criteria:* Since the evaluation task belongs to a classification problem, F1 score and Accuracy are used to validate our model. We also employ the area-under-the-curve of receiver operator characteristic (AUCROC) and precision-recall (AUCPR) scores to numerically evaluate the quality of each method. In addition, as the dataset contains multiple classes, here we show both the Macro-F1 and MicroF1 scores. MacroF1 score biases the metric towards the least populated classes, while MicroF1 score biases towards to the most populated classes. Note that the AUC-ROC and AUC-PR scores are both based on the Macro metric.

To evaluate our model as a cross-subject classifier, considering the computational expense, we perform 5-fold subject-independent cross validation and report the average test performance for each method. The ratio of training, validation and test sets is $0.7 : 0.1 : 0.2$. Note that, the models are never

TABLE I
CONFIGURATIONS OF THE MULTI-VIEW CONVOLUTIONAL REPRESENTATION MODULE IN HYBRIDATT.

| Type | Kernel size | Stride | Padding |
|---|---|---|---|
| $Conv1d_{1\_1}$ | $8 \times 8$ | 2 | 3 |
| $Conv1d_{1\_2}$ | $16 \times 8$ | 2 | 7 |
| $Conv1d_{1\_3}$ | $32 \times 8$ | 2 | 3 |
| $Conv1d_{1\_4}$ | $64 \times 8$ | 2 | 7 |
| $MaxPool1d_{1}$ | 6 | 4 | 1 |
| $Conv1d_{2\_1}$ | $3 \times 16$ | 1 | 1 |
| $Conv1d_{2\_2}$ | $5 \times 16$ | 1 | 2 |
| $MaxPool1d_{2}$ | 3 | 2 | 1 |
| $Conv1d_{3\_1}$ | $3 \times 16$ | 1 | 1 |
| $Conv1d_{3\_2}$ | $5 \times 16$ | 1 | 2 |
| $MaxPool1d_{3}$ | 3 | 2 | 1 |
| $Conv2d_{1\_1}$ | $1 \times 8 \times 8$ | $1, 2$ | $0, 3$ |
| $Conv2d_{1\_2}$ | $1 \times 16 \times 8$ | $1, 2$ | $0, 7$ |
| $Conv2d_{1\_3}$ | $1 \times 32 \times 8$ | $1, 2$ | $0, 3$ |
| $Conv2d_{1\_4}$ | $1 \times 64 \times 8$ | $1, 2$ | $0, 7$ |
| $AvgPool2d_{1}$ | $1 \times 6$ | $1, 4$ | $0, 1$ |
| $Conv2d_{2\_1}$ | $3 \times 3 \times 16$ | $1, 1$ | $1, 1$ |
| $Conv2d_{2\_2}$ | $5 \times 5 \times 16$ | $1, 1$ | $2, 2$ |
| $AvgPool2d_{2}$ | $1 \times 3$ | $1, 2$ | $0, 1$ |
| $Conv2d_{3\_1}$ | $3 \times 3 \times 16$ | $1, 1$ | $1, 1$ |
| $Conv2d_{3\_2}$ | $5 \times 5 \times 16$ | $1, 1$ | $2, 2$ |
| $AvgPool2d_{3}$ | $14 \times 3$ | $14, 2$ | $0, 1$ |

TABLE II
CLASSIFICATION PERFORMANCE COMPARISONS ON THE UCD DATASET IN THE FREQUENCY DOMAIN.

| | UCD Dataset (frequency Domain) | | | | |
|---|---|---|---|---|---|
| Method | AUCROC | AUCPR | MacroF1 | MicroF1 | Accuracy |
| PSVM | 0.8177 | 0.5767 | 0.5204 | 0.5854 | 0.6193 |
| DNN | 0.7213 | 0.5224 | 0.3542 | 0.4331 | 0.5262 |
| RNN | 0.6228 | 0.3350 | 0.2663 | 0.3970 | 0.5091 |
| RNNAtt$_l$ | 0.6172 | 0.3305 | 0.2457 | 0.3734 | 0.5002 |
| RNNAtt$_c$ | 0.6234 | 0.3335 | 0.2554 | 0.3712 | 0.5010 |
| CNN | 0.8732 | 0.6725 | 0.5925 | 0.6492 | 0.6590 |
| CRNN | 0.8660 | 0.6454 | 0.5693 | 0.6395 | 0.6634 |
| CRNNAtt$_l$ | 0.8570 | 0.6281 | 0.5810 | 0.6486 | 0.6683 |
| CRNNAtt$_c$ | 0.8671 | 0.6418 | 0.5849 | 0.6528 | 0.6791 |
| ChannelAtt | 0.8705 | 0.6818 | 0.6517 | 0.7070 | 0.7152 |
| HybridAtt$_l$ | 0.8719 | 0.6669 | 0.6342 | 0.6962 | 0.7070 |
| HybridAtt$_f$ | **0.8854** | **0.6886** | **0.6639** | **0.7231** | **0.7328** |

trained on data from the test subjects, and we adopt the same subject combination for all the models in each fold in order to fairly compare the performance.

*4) Implementation Details:* We implement all the approaches with Pytorch. The training process is done locally using NVIDIA Titan Xp GPU. During the training phase, we minimize the cost function by utilizing the Adadelta optimization algorithm [26]. We also use momentum ($\rho = 0.95$), weight decay ($L2$ penalty with the coefficient $0.001$), and dropout strategies (the dropout rate is $0.5$) for all the approaches. Furthermore, the configurations of our multi-view convolutional representation module is shown in Table I, and we set the same $p = 128$, $q = 128$, and $r = 128$ for baselines and our models.

### C. Experimental results

In this subsection, we compare the performance of our proposed HybridAtt model with the aforementioned baselines in the task of sleep stage classification. The experimental results are listed in Table II. We can observe that our proposed HybridAtt networks outperform all the baselines on the benchmark dataset.

Given the results of the baselines, PSVM performs better than DNN and the RNN-based models. The reason is that representing signals in the frequency domain would provide more powerful information, and SVM can hence learn a more distinct hyper-lane to separate each sleep stage in the vector space. This observation can also be found from the performance of DNN where it achieves better results benefiting from the handcrafted spectral features. Not surprisingly, CNN-based models work well on this task, which justifies the effectiveness of the proposed multi-view feature representation using convolutional operators. We can also observe that the attention-based CRNN models get better results than the plain CRNN model. This is because attention mechanism can help model to focus on more useful information carried by sequential hidden features and hence improve the classification performance. The ChannelAtt model adopting channel-aware attention performs better than the time-aware attention models. It illustrates that there exist more useful connections among PSG channels, and it is reasonable to employ attention mechanism to capture those connections for feature representation. Furthermore, taking the hybrid attention strategy into consideration, our proposed HybridAtt model yields better performance of sleep stage classification.

From the results of our models, we can see that the HybridAtt$_f$ model outperforms the other methods on all five evaluation measurements. Based on the overall performance comparisons, we can conclude that the single-dimension attention networks may lose critical information, and hence do not work well dealing with multivariate PSG data. By incorporating the hybrid attention structure, the proposed fusion-based hybrid attention mechanism achieves better results compared with the location-based hybrid attention mechanism. We arrive at a conclusion that our proposed HybridAtt model indeed learns informative representations to improve the performance of sleep stage classification.

## IV. CONCLUSIONS

Multivariate sleep stage classification has become a hot research topic in a variety of medical applications in healthcare. In this paper, we propose a hybrid self-attentive deep learning network, named HybridAtt, to classify sleep stages of heterogeneous PSG records. The proposed HybridAtt is an end-to-end model that combines multi-view convolutional representation with hybrid self attention mechanism to extract representative features from multivariate biosignals. In order to unleash the power of multi-view feature learning, we construct a new CNN structure to learn latent representations from channel-specific and global views. A new fusion-based hybrid attention mechanism, consisting of channel-wise and time-wise attention layers, is proposed to capture the dual correlations of PSG channels and timestamps. It can also integrate the complementary information carried by both learned feature views. Experimental results on a benchmark PSG dataset justify the effectiveness of our proposed HybridAtt model.

## Acknowledgment

This work has been sponsored by the National Science Foundation of China under Grants No. 81871394, the Beijing Municipal Education Committee Science Foundation under No. KZ201610005007, and Beijing Laboratory of Advanced Information Networks under Grants No. 040000546618017. The authors would like to thank the anonymous reviewers, and NVIDIA Corporation for the donation of the Titan Xp GPU.

## References

[1] B. Şen, M. Peker, A. Çavuşoğlu, and F. V. Çelebi, "A comparative study on classification of sleep stage based on eeg signals using feature selection and classification algorithms," *Journal of medical systems*, vol. 38, no. 3, p. 18, 2014.

[2] R. Boostani, F. Karimzadeh, and M. Nami, "A comparative review on sleep stage classification methods in patients and healthy individuals," *Computer methods and programs in biomedicine*, vol. 140, pp. 77–91, 2017.

[3] K. A. I. Aboalayon, M. Faezipour, W. S. Almuhammadi, and S. Moslehpour, "Sleep stage classification using eeg signal analysis: a comprehensive survey and new investigation," *Entropy*, vol. 18, no. 9, p. 272, 2016.

[4] S. Najdi, A. A. Gharbali, and J. M. Fonseca, "Feature transformation based on stacked sparse autoencoders for sleep stage classification," in *Doctoral Conference on Computing, Electrical and Industrial Systems*. Springer, 2017, pp. 191–200.

[5] Y. Yuan, G. Xun, K. Jia, and A. Zhang, "A novel wavelet-based model for eeg epileptic seizure detection using multi-context learning," in *Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on*. IEEE, 2017, pp. 694–699.

[6] Y. Yuan, G. Xun, Q. Suo, K. Jia, and A. Zhang, "Wave2vec: Deep representation learning for clinical temporal data," *Neurocomputing*, 2018.

[7] M. Längkvist, L. Karlsson, and A. Loutfi, "Sleep stage classification using unsupervised feature learning," *Advances in Artificial Neural Systems*, vol. 2012, p. 5, 2012.

[8] J. Zhang, Y. Wu, J. Bai, and F. Chen, "Automatic sleep stage classification based on sparse deep belief net and combination of multiple classifiers," *Transactions of the Institute of Measurement and Control*, vol. 38, no. 4, pp. 435–451, 2016.

[9] A. Supratak, H. Dong, C. Wu, and Y. Guo, "Deepsleepnet: a model for automatic sleep stage scoring based on raw single-channel eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1998–2008, 2017.

[10] O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou, "Automatic sleep stage scoring with single-channel eeg using convolutional neural networks," *arXiv preprint arXiv:1610.01683*, 2016.

[11] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2018.

[12] E. P. Giri, M. I. Fanany, and A. M. Arymurthy, "Combining generative and discriminative neural networks for sleep stages classification," *arXiv preprint arXiv:1610.01741*, 2016.

[13] M. Zhao, S. Yue, D. Katabi, T. S. Jaakkola, and M. T. Bianchi, "Learning sleep stages from radio signals: a conditional adversarial architecture," in *International Conference on Machine Learning*, 2017, pp. 4100–4109.

[14] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Information Fusion*, vol. 38, pp. 43–54, 2017.

[15] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "Deepsense: A unified deep learning framework for time-series mobile sensing data processing," in *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017, pp. 351–360.

[16] Y. Yuan, G. Xun, K. Jia, and A. Zhang, "A multi-view deep learning method for epileptic seizure detection using short-time fourier transform," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2017, pp. 213–222.

[17] Y. Yuan, G. Xun, F. Ma, Q. Suo, H. Xue, K. Jia, and A. Zhang, "A novel channel-aware attention framework for multi-channel eeg seizure detection via multi-view deep learning," in *Biomedical & Health Informatics (BHI), 2018 IEEE EMBS International Conference on*. IEEE, 2018, pp. 206–209.

[18] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 945–953.

[19] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[20] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[21] E. A. Wolpert, "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects." *Archives of General Psychiatry*, vol. 20, no. 2, pp. 246–247, 1969.

[22] M. Manzano, A. Guillén, I. Rojas, and L. J. Herrera, "Deep learning using eeg data in time and frequency domains for sleep stage classification," in *International Work-Conference on Artificial Neural Networks*. Springer, 2017, pp. 132–141.

[23] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[24] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[25] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 1903–1911.

[26] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.