Channel Contribution in Deep Learning Based Automatic Sleep Scoring—How Many Channels Do We Need?

Changqing Lu, Shreyasi Pathak, Gwenn Englebienne, and Christin Seifert

***Abstract*— Machine learning based sleep scoring methods aim to automate the process of annotating polysomnograms with sleep stages. Although sleep signals of multiple modalities and channels should contain more information according to sleep guidelines, most multi-channel multi-modal models in the literature showed only a little performance improvement compared to single-channel EEG models and sometimes even failed to outperform them. In this paper, we investigate whether the high performance of single-channel EEG models can be attributed to specific model features in their deep learning architectures and to which extent multi-channel multi-modal models take the information from different channels of modalities into account. First, we transfer the model features from single-channel EEG models, such as combinations of small and large filters in CNNs, to multi-channel multi-modal models and measure their impacts. Second, we employ two explainability methods, the layer-wise relevance propagation as post-hoc and the embedded channel attention network as intrinsic interpretability methods, to measure the contribution of different channels on predictive performance. We find that i) single-channel model features can improve the performance of multi-channel multi-modal models and ii) multi-channel multi-modal models focus on one important channel per modality and use the remaining channels to complement the information of the focused channels. Our results suggest that more advanced methods for aggregating channel information using complementary information from other channels may improve sleep scoring performance for multi-channel multi-modal models.**

***Index Terms*— Channel contribution, deep learning, EEG, EOG, EMG, multi-channel multi-modal sleep scoring.**

I. Introduction

SLEEP stage annotations assist clinicians in detecting sleep disorders and formulating treatment plans for patients. Sleep stages are scored based on polysomnograms (PSGs) consisting of activity recordings of various parts of human body, e.g. electroencephalograms (EEGs), electrooculograms (EOGs) and electromyograms (EMGs). For annotation, PSGs of approximately 8-h sleep are segmented into 30-s epochs and annotated by sleep technicians following standardized guidelines. The Rechtchaffen and Kales standard (R&K manual) [1] and the American Academy of Sleep Medicine rules (AASM manual) [2] are the two most widely used guidelines, distinguishing between seven stages[1]: Wake, Non-REM1 (N1), Non-REM2 (N2), Non-REM3 (N3), Rapid Eye Movement (REM), Movement and Unscored. Each sleep stage is characterized by distinctive time- and frequency-domain patterns. Table I provides a summary of these specific patterns as defined in the AASM manual [2].

Sleep scoring is traditionally performed manually by sleep technicians. To reduce the manual effort and time for annotation, automatic sleep scoring approaches have been developed. In general, automatic sleep scoring approaches can be categorized into traditional machine learning approaches and deep learning approaches. The former (e.g., [3], [4], [5]) relied on manually defined features and applied traditional machine learning models to classify sleep stages based on these features. The latter (e.g., [6], [7], [8]) captured temporal and sequential features from raw sleep signals or transformed frequency representations (e.g., spectrograms) automatically using end-to-end deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). In our work, we focus on deep learning approaches, as they are more generalizable when applied to highly heterogeneous data sets [9].

Most early deep learning approaches were based on single-channel EEG (e.g., [6], [7]), as it contains the most information [2]. Khalighi et al. [10] showed that incorporating multiple modalities and channels (i.e., EOGs and EMGs) could improve the performance. Yet, surprisingly, most current multi-channel multi-modal models obtained very little performance improvement compared to single-channel EEG models and sometimes even failed to outperform them (cf., Table II). Additionally, it was found in [11] that while adding EEG channels improved the performance, using more than 6 EEG channels did not improve it further. We propose the following hypotheses: i) some single-channel EEG models successfully added particular model features into their deep learning architectures, which improved the performance, such as combining small and large filters in temporal learning to respectively

Manuscript received 28 July 2022; revised 15 November 2022; accepted 27 November 2022. Date of publication 6 December 2022; date of current version 1 February 2023. *(Corresponding author: Changqing Lu.)*

Changqing Lu and Gwenn Englebienne are with the Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, 7522NB Enschede, The Netherlands (e-mail: c.lu@utwente.nl; g.englebienne@utwente.nl).

Shreyasi Pathak is with the Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, 7522NB Enschede, The Netherlands, and also with the Institute for Artificial Intelligence in Medicine, University of Duisburg-Essen, 45147 Essen, Germany (e-mail: s.pathak@utwente.nl).

Christin Seifert is with the Institute for Artificial Intelligence in Medicine, University of Duisburg-Essen and University Hospital Essen, 45147 Essen, Germany (e-mail: christin.seifert@uni-due.de).

Digital Object Identifier 10.1109/TNSRE.2022.3227040

[1]The R&K manual distinguishes between eight stages, with stage N3 further split into stages N3 and N4. Following prior work, we use the AASM manual in our study.

TABLE I
SUMMARY OF EEG, EOG AND EMG TIME- AND FREQUENCY-DOMAIN PATTERNS FOR FIVE MAIN SLEEP STAGES IN THE AASM MANUAL [2]

Stages	EEG					EOG	EMG
	Delta (<4Hz)	Theta (4-7Hz)	Alpha (8-13Hz)	Beta (>13Hz)	Time-domain		
Wake			x	x		Eye movement (0.5-2Hz)	Variable amplitude but usually higher than during sleep stages
N1		x	x		Vertex waves	Slow eye movement	Lower amplitude than in stage Wake
N2		x			K-complexes Sleep spindles	Usually no eye movement, but slow eye movements may persist	Lower amplitude than in stage Wake and may be as low as in stage REM
N3	x				High amplitude Sleep spindles	Usually no eye movement	Lower amplitude than in stage N2 and sometimes as low as in stage REM
REM		x	x		Sawtooth waves	Rapid eye movement	Lower chin EMG tone; the lowest amplitude among all stages

capture time- and frequency-domain features [6], but the utilities of these model features have not been tested in the multi-channel multi-modal setting; ii) although all modalities and channels contain information, the information of certain channels of modalities may be sufficient to obtain accurate predictions.

To verify the hypotheses proposed above, in this paper, we investigate in two directions: i) whether multi-channel multi-modal models can be improved by adding particular model features from high-performing single-channel EEG models and ii) which channels contribute to a high-performing multi-channel multi-modal model. Specifically, our contributions are:

1) We evaluate the impacts of particular model features proposed for high-performing single-channel EEG models in the multi-channel multi-modal setting on a public benchmark data set, SleepEDF-13.
2) We incorporate the model features that improve the performance into a multi-channel multi-modal model and evaluate it on two public benchmark data sets, SleepEDF-13 (39 PSGs, small) and SHHS-1 (5,793 PSGs, large), obtaining state-of-the-art results.
3) We apply the layer-wise relevance propagation (LRP) [12], a post-hoc explainability method for model agnostic, to extract channel importance. We also adopt an embedded channel attention network (eCAN), motivated by [13] and [14], which intrinsically incorporates channel importance to deep sleep scoring models. We compare the results from both methods.
4) Based on the observations obtained from the interpretability experiments, we hypothesize that incorporating all channels is not necessary to obtain acceptable performance and verify it in a reverse ablation study.

The remainder of the paper is organized as follows. Section II presents the related work on deep learning based automatic sleep scoring and reviews the methods for extracting channel importance from deep learning models. Section III introduces data sets and data preprocessing. The experiment for evaluating single-channel model features in the multi-channel multi-modal setting and the accordingly improved multi-channel multi-modal model structure are described in Section IV. Afterwards, we present two interpretability methods to analyze channel contribution in Section V. Section VI provides the experiment setup for evaluating our multi-channel multi-modal model. All results are presented and discussed in Section VII. Finally, we conclude and outline the directions for future work in Section VIII.

II. RELATED WORK

In this section, we review deep learning based automatic sleep scoring approaches and the prior work on channel contribution analysis.

A. Automatic Sleep Scoring

We distinguish between single-channel EEG and multi-channel multi-modal deep sleep scoring models. Table II provides an overview of them in terms of data sets, used features, approaches and performance.

1) Single-Channel EEG Models: The most classic architecture is a combination of convoluational neural networks (CNNs) and recurrent neural networks (RNNs) (e.g., [6], [7], [16], [17]). Usually, CNNs are used to extract temporal features from sleep epochs and RNNs are employed to capture transition information from sleep sequences. To further improve the scoring performance, particular model features were added to this base architecture. Supratak et al. [6] employed filters of small and large sizes in the first layer of CNNs to capture both time- and frequency-domain features. Sors et al. [15] used deep CNNs to extract complex patterns of sleep epochs, because feature complexity can be increased by deeper layers [27]. Mousavi et al. [7] applied attention mechanisms in RNNs to focus on the important parts of sleep sequences when considering context information. Moreover, to enable the model to consider temporal and sequential features evenly in the final classification of sleep stages, Supratak et al. [6] also built a residual connection that concatenated features from temporal encoding layers. Additionally, there were also different model architectures proposed, such as using large-scale CNNs to capture transition information from neighbouring sleep epochs instead of using RNNs [15], [20] and learning sleep features not only from raw signals but also from their frequency representations [8], [18].

TABLE II
OVERVIEW OF THE STATE-OF-THE-ART DEEP LEARNING BASED AUTOMATIC SLEEP SCORING MODELS. WE DISTINGUISH BETWEEN SINGLE-CHANNEL EEG AND MULTI-CHANNEL MULTI-MODAL MODELS. THE '**FEATURES**' COLUMN INDICATES WHETHER RAW SIGNALS OR TRANSFORMED FREQUENCY REPRESENTATIONS WERE USED IN CORRESPONDING PUBLICATIONS

Paper	Year	Dataset	PSGs	Channels	Features	Approach	Evaluation	Accuracy
Supratak et al. [6]	2017	SleepEDF-13	39	1EEG	raw	CNN-RNN	20-fold CV	82.0
		MASS	62	1EEG			31-fold CV	86.2
Sors et al. [15]	2018	SHHS-1	5,728	1EEG	raw	CNN	50-20-30	86.8
Mousavi et al. [7]	2019	SleepEDF-13	39	1EEG	raw	CNN-RNN	20-fold CV	84.3
Supratak et al. [16]	2020	SleepEDF-13	39	1EEG	raw	CNN-RNN	20-fold CV	85.4
		MASS	62	1EEG			31-fold CV	87.5
Seo et al. [17]	2020	SleepEDF-13	39	1EEG	raw	CNN-RNN	20-fold CV	83.9
		MASS	62	1EEG			31-fold CV	86.3
		SHHS-1	5,791	1EEG			50-20-30	86.7
Wang et al. [8]	2020	SleepEDF-13	39	1EEG	frequency	CNN	20-fold CV	86.1
Sun et al. [18]	2020	MASS	147	1EEG	raw+frequency	CNN-RNN	leave-one-out CV	86.1
Eldele et al. [19]	2021	SleepEDF-13	39	1EEG	raw	CNN	20-fold CV	85.6
		SHHS	329	1EEG			20-fold CV	86.6
Firillo et al. [20]	2021	SleepEDF-13	39	1EEG	raw	CNN	20-fold CV	84.0
Paisarnsrisomsuk et al. [21]	2018	SleepEDF-13	39	2EEGs+1EOG	raw	CNN	4-fold CV	81.2
Phan et al. [22]	2018	SleepEDF-13	39	1EEG+1EOG	frequency	CNN	20-fold CV	82.3
		MASS	200	1EEG+1EOG+1EMG			20-fold CV	83.6
Chambon et al. [11]	2018	MASS	61	6EEGs+2EOGs+3EMGs	raw	CNN	5-fold CV	83.0
Sun et al. [18]	2020	MASS	147	1EEG+1EOG+1EMG	raw+frequency	CNN-RNN	leave-one-out CV	87.8
Jia et al. [23]	2020	MASS	62	20EEGs+2EOGs+3EMGs+1ECG	frequency	Graph CNN	31-fold CV	88.9
Pathak et al. [24]	2021	SHHS-1	5,793	2EEGs+2EOGs+1EMG	raw	CNN-RNN	81-9-10	85.0
Phan et al. [25]	2021	SleepEDF-13	39	1EEG+1EOG	raw+frequency	CNN-RNN	20-fold CV	86.3
		MASS	200	1EEG+1EOG+1EMG			20-fold CV	87.6
		SHHS-1	5,791	1EEG+1EOG+1EMG			70-30	89.1
Jia et al. [26]	2021	SleepEDF-13	39	1EEG+1EOG	raw	CNN	20-fold CV	87.5
This paper	2022	SleepEDF-13	39	2EEGs+1EOG+1EMG	raw	CNN-RNN	20-fold nested CV	87.2
		SHHS-1	5,793	2EEGs+2EOGs+1EMG			81-9-10	89.1

2) Multi-Channel Multi-Modal Models: Most of the multi-channel multi-modal models also used classic CNN-RNN architectures but with some additional spatial learning modules added to incorporate the sleep features of multiple modalities and channels. Paisarnsrisomsuk et al. [21] employed large-scale CNNs to extract both temporal and sequential information from 2 EEGs and 1 EOG and found that adding EOG signals increased the accuracy by 1%. A similar result was observed by Phan et al. [22] who generated the spectrograms of sleep signals and trained multi-task CNNs to create joint predictions for the current and neighbouring sleep epochs. Chambon et al. [11] proposed a spatio-temporal CNN architecture and used linear spatial filters to increase the signal-to-noise ratio. Pathak et al. [24] also designed a spatial-temporal-sequential model to extract sleep features from multi-channel multi-modal data input and verified that EEG is the most important modality using post-hoc interpretability methods. In [25], Phan et al. created a multi-view sequential model via learning joint representations from both raw signals and time-frequency images. Recently, more advanced architectures were developed to particularly model correlations among modalities and among channels within a modality. For instance, Jia et al. [23] employed graph CNNs to capture intrinsic connections among EEG channels. In another paper [26], they designed a multi-modal attention module which helped detect the relevance between EEG and EOG signals.

In summary, although some studies showed that adding the information of multiple modalities and channels could improve the performance, the improvement was rather small (cf., Table II). In addition, there was also evidence that the improvement vanished after adding many EEG channels [11]. Hence, it is important to understand how multi-channel multi-modal models use the information from different channels of multiple modalities. To that end, we built a state-of-the-art multi-channel multi-modal model and analyzed channel contribution in detail. More specifically, we first tested the impacts of particular model features developed for single-channel EEG models in the multi-channel multi-modal setting. Then, we designed an architecture aggregating the promising model features and investigated channel importance using interpretability methods. Since advanced architectures (i.e., [23], [26]) were designed for particular data[2] and focused on modelling the relevance among modalities and channels instead of understanding channel importance, we stayed with classic CNN-RNN architectures in our study to analyze channel contribution.

B. Analyzing Channel Contribution

We propose to analyze channel contribution in a deep sleep scoring model by assessing the importance of channel information acknowledged by that model. To the best of our knowledge, no particular study has investigated channel importance in multi-channel multi-modal sleep scoring, although some literature (e.g., [6], [7]) showed that the scoring performance varied based on different channels.

In related domains, Bohle et al. [28] used the layer-wise relevance propagation (LRP) to assist clinicians in explaining the area importance of MRIs for diagnosing Alzheimer's

[2]GraphSleepNet [23] needs a relatively large amount of channels to build the graph structure, which are not available in SleepEDF-13 and SHHS-1. If SalientSleepNet [26] considers the relevance from EMGs and from different channels within a modality, the model complexity will increase rapidly.

TABLE III
OVERVIEW OF SLEEPEDF-13 AND SHHS-1, SHOWING THE AMOUNT OF 30-S EPOCHS AND CLASS FREQUENCIES

Data set	Sleep Stages					Total
	Wake	N1	N2	N3	REM	
SleepEDF-13	8,285 (19.6%)	2,804 (6.6%)	17,799 (42.1%)	5,703 (13.5%)	7,717 (18.2%)	42,308
SHHS-1	1,691,288 (28.8%)	217,583 (3.7%)	2,397,460 (40.9%)	739,403 (12.6%)	817,473 (13.9%)	5,863,207

disease. Hu et al. [13] developed a Squeeze-and-Excitation block (SE-Net) to recalibrate channel-wise feature responses by modelling inter-channel dependencies. Wang et al. [14] proposed a similar model, the efficient channel attention net (ECA-Net), where an extra CNN block was employed to detect channel importance. Bastidas and Tang [29] also implemented a channel attention network (CAN) to allocate large attention weights to important channels in image predictions.

In this work, we used two kinds of methods to assess channel importance in multi-channel multi-modal sleep scoring in different directions. We applied the LRP [12] where the importance score of a channel is concluded by its relevance score for predictions. We also employed an embedded CAN, motivated by [13] and [14], to intrinsically measure channel importance, i.e. the importance score of a channel is learned and allocated via an extra neural network.

III. DATA SETS AND PRE-PROCESSING

We based our experiments on two public benchmark data sets: the SleepEDF-13 data set containing 39 PSGs and the SHHS-1 data set containing 5,793 PSGs. Table III provides an overview of them.

A. SleepEDF-13

SleepEDF-13 [30], [31] is a small data set initially used for two studies: investigating age effect in health subjects (SC) and investigating Temazepam effects on sleep (ST). Following prior work (e.g., [6], [7]), we used the 20 SC subjects in our study. For each subject, except one, 2 PSGs are available, resulting in 39 PSGs. Each PSG consists of 2 EEGs (channels Fpz-Cz and Pz-Cz), 1 EOG (horizontal) and 1 EMG. The EEG and EOG signals are sampled at 100 Hz, while the EMG signals are sampled at 1 Hz. Sleep epochs are manually annotated as one of the 8 sleep stages (Wake, N1, N2, N3, N4, REM, Movement and Unscored), according to the R&K manual [1].

B. SHHS-1

Sleep Heart Health Study (SHHS) [32] is a large data set used for sleep-disordered breathing research and consists of the data collected during two patient visits. Following prior work (e.g., [15], [24]), we used the subjects from the first visit (SHHS-1). Overall, 5,793 PSGs are collected from 5,793 subjects, where 2 EEGs (channels C3-A2 and C4-A1), 2 EOGs (left and right) and 1 EMG are recorded. The EEG and EMG signals are sampled at 125 Hz, while the EOG signals are sampled at 50 Hz. Similar to SleepEDF-13, the R&K manual [1] is used for annotating SHHS-1, also resulting in 8 sleep stages.

C. Data Pre-Processing

For both data sets, we merged stages N3 and N4 into stage N3 to comply with the AASM manual [2] and removed the Movement and Unscored epochs which are irrelevant for sleep scoring. In addition, following [6], we excluded long wake periods that are located 30 minutes before and after sleep periods for SleepEDF-13.

We preprocessed the signals in both data sets as follows. First, we resampled the signals at smaller sampling rates to the highest sampling rate among all signals in that data set (i.e., resampling the EMG signals in SleepEDF-13 to 100 Hz and the EOG signals in SHHS-1 to 125 Hz) such that all modalities in a data set share an identical feature extraction mechanism in deep sleep scoring models. Second, following [11], [24], [33], we filtered the EEG and EOG signals of both data sets to 0.16-30 Hz and the EMG signals to 10-30 Hz[3] and standardized the signals of every channel to mean 0 and standard deviation 1.

IV. IMPROVING MULTI-CHANNEL MULTI-MODAL MODEL

In this section, we present the improved multi-channel multi-modal model that is based on the promising model features developed for single-channel EEG models.

A. Evaluating Single-Channel Model Features

To apply particular model features that have been successfully used by single-channel EEG models for performance improvement to multi-channel multi-modal models, we first tested their utilities in the multi-channel multi-modal setting. Based on the review presented in Section II-A.1, we selected the four model features presented in Table IV as candidates. The assumptions for these model feature choices are as follows. Adding large filters to capture frequency-domain patterns enables the model to capture distinctive frequency features, e.g. the Delta waves in stage N3. Increasing feature complexity helps detect the sleep stages whose time-domain patterns are indistinguishable, e.g. stages N1 and REM. A focus on the important parts of sleep sequences improves the extraction of transition information thus benefits to associated transition stages. Moreover, an even attention on temporal and sequential features avoids the loss of temporal information in sequential learning. We selected the model by Pathak et al. [24] as the baseline,[4] because it was based on classic CNN-RNN architectures and obtained state-of-the-art performance. Additionally, we performed a reverse ablation study, i.e. adding

[3]Note that, the bandwidth selection for sleep scoring is not uniformly done in related work. We followed recent papers [11], [24], [33] to filter the EMG signals, while 0-100 Hz were selected in other work [34].

[4]For all model variants in this section, we excluded the spatial learning module of the baseline model, as detecting correlations among channels within a modality was not focused in this experiment.

TABLE IV
TESTED MODEL FEATURES FROM SINGLE-CHANNEL EEG MODELS AND THEIR FUNCTIONS, FOR IMPROVING MULTI-CHANNEL MULTI-MODAL MODELS

No.	Feature	Function
i	small and large filters in the first layer of CNNs	to capture time- and frequency- domain features of sleep signals respectively
ii	deeper layers in CNNs	to increase feature complexity
iii	attention mechanisms embedded in RNNs	to concentrate on the important parts of sleep sequences
iv	residual connection that concatenates features of CNNs and RNNs	to consider temporal and sequential features evenly in the final classification of sleep stages

model modifications one at a time, to test their impacts. Our experiment consists of three steps:

1) Model features in CNNs (cf., Table IV, i and ii): to test their impacts independently, we added them to the CNNs of the baseline model separately, one at a time, and measured the performance.
2) Model feature in RNNs (cf., Table IV, iii): we applied sequential learning with and without attention mechanisms on the model obtained from the first step and measured the performance.
3) Model feature in the whole architecture (cf., Table IV, iv): we measured the performance on the model obtained from the second step with and without the residual connection to verify its impact.

We ran the experiment on SleepEDF-13 using all four accessible channels (cf., Section III-A) and evaluated all model variants under the nested cross validation scheme (cf., Section VI-C). To determine the significance of adding the model features from single-channel EEG models to multi-channel multi-modal models, we employed statistical hypothesis testing. Specifically, we assumed that the sleep data of the 20 subjects in SleepEDF-13 is independent and identically distributed, thus the evaluation metrics computed over the data of each subject follow a Gaussian distribution [35]. Then, we reported a sequence of 20 macro F1-scores (cf., Section VI-B) per model variant, obtained from the 20-fold cross validation in the outer loop of the evaluation scheme. Afterwards, we compared the model variants pairwise on their respective sequences using an one-sided Welch's t-test, where the null hypothesis was set that the performance improved by adding a model feature is smaller than or equal to zero. We set a significance level to 0.05 for the test: if the p-value is smaller than 0.05, the added model feature is improving the performance of multi-channel multi-modal sleep scoring models.

The results for evaluating the model features from single-channel EEG models in the multi-channel multi-modal setting are presented in Fig. 1. Overall, all four model features are shown statistically significant, since they all achieved p-values smaller than 0.05 in the Welch's t-tests. We thus concluded that all four tested model features from single-channel EEG models are useful to improve multi-channel multi-modal models under classic CNN-RNN architectures.

B. Final Multi-Channel Multi-Modal Model

Based on the experiment presented in the previous section, we introduce our improved multi-channel multi-modal sleep scoring model here. Our model consists of four components: a *temporal learning* module used to extract temporal features, a *spatial learning* part embedded in the first layer of temporal learning to incorporate channel information within a modality, a *sequential learning* module applied to capture sequential features from sleep sequences and a *residual connection* employed to concatenate CNN and RNN features. The final classification of sleep stages is performed on the obtained feature representations via a fully-connected layer with the SoftMax activation function. Fig. 2 shows the full structure of the improved multi-channel multi-modal model on the SHHS-1 data set which contains 2 EEGs, 2 EOGs and 1 EMG. Note that, each modality $m \in \{\text{EEG}, \text{EOG}, \text{EMG}\}$ can have more than one signal which is referred as a channel throughout this paper. We describe the four components above in more detail as follows.

1) Temporal Learning: The first convolutional layer has two pipelines, one with small filter size and the other with large filter size, to respectively capture time- and frequency-domain features from raw sleep signals. Additional convolutional layers are added to extract complex underlying features. Specifically, each pipeline of CNNs consists of four convolutional layers and two max-pooling layers. Each convolutional layer is followed by a batch normalization layer [36] and a rectified linear unit (ReLU) activation layer (i.e., $\text{ReLU}(x) = \max(0, x)$). Details on the number of filters, filter sizes, stride and pooling sizes are shown in Figure 2. Following [6], we set the smaller filter size in the first convolutional layer to half the sampling rate, as distinctive time-domain features (e.g., K-complex) usually appear in 0.5-s ranges in sleep epochs. We set the larger filter size to 4 times the sampling rate to better detect the frequency components of these signals. Different from [6], we set the stride size in the first convolutional layer to 1 instead of a large value to prevent the information loss of basic features. Accordingly, we applied larger pooling sizes in the max-pooling layers to filter out more representative features and avoid overfitting. At the end of CNNs, the features extracted from time- and frequency-domain pipelines are concatenated as the final temporal feature representations of sleep epochs. We also employed two dropout layers [37] of probability 0.5 as regularization techniques to help prevent overfitting in the training process.

2) Spatial Learning: Li et al. [38] have shown that low temporal relevance can exist among EEG channels in Non-wake stages. To detect and incorporate channel information, i.e. spatial correlations among channels within a modality, we integrated a spatial block in the first temporal convolutional layer, following [24], [39]. First, we reshaped the signals of multiple channels of a modality m into an input of shape, $C_m \times D$, where C_m is the number of channels in this modality and D is the number of data points. Then, we passed them to the first convolutional layer of temporal learning including C_m input channels and 64 output channels. This layer learns

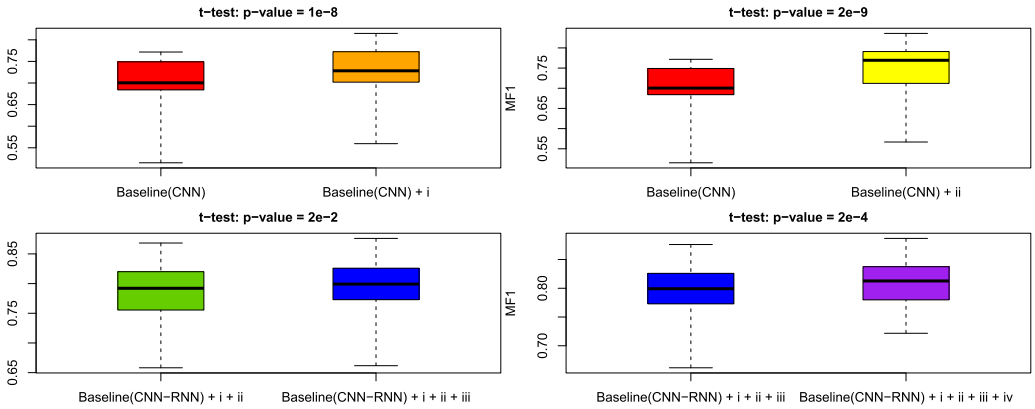


Fig. 1. Results for evaluating the model features from single-channel EEG models in the multi-channel multi-modal setting. Box plots show the distributions of the 20 macro F1-scores obtained for corresponding model variants. The p-values of the Welch's t-tests are also provided.

temporal features from raw sleep signals and then spatially aggregates the feature maps learnt from each channel. Compared to [24] where spatial learning was applied directly on raw sleep signals, our spatial learning block is applied on temporal feature maps, which has the advantage that distinctive patterns (e.g., sawtooth waves) existing in raw sleep signals will not be changed before they are identified.

3) Sequential Learning: To concentrate on the important sleep epochs of a sleep sequence, we employed the encoder-decoder sequential learning module by Luong et al. [40] with attention mechanisms in order to learn transition information from sleep sequences, as the stage of a sleep epoch is determined by both its own features and the information of neighbouring epochs [2]. Specifically, there are two phases: an encoding phase used to capture context information of sleep sequences and a decoding phase used to predict sleep stages epoch by epoch. The encoder employs two bidirectional long short-term memory layers (Bi-LSTM) with 256 hidden units to learn the context dependencies of sleep sequences containing multi-channel multi-modal CNN features in both forward and backward directions. The decoder uses a block composed of two long short-term memory layers (LSTM), an attention module and a fully-connected layer (FC) to predict sleep stages iteratively. Consider a sequence of n sleep epochs. The specific computation to predict the sleep stage for an epoch t can be expressed as follows:

$$\begin{aligned}
\bar{h}_s &= \text{mean}(\text{LSTM}_f, \text{LSTM}_b) = \text{Bi-LSTM}(f_{\text{MM-CNN}}),\\
h_{d,t} &= \text{LSTM}(h_{d,t-1}, y_{t-1}),\\
sc_{t,i} &= \tanh\left(W_d h_{d,t} + W_s \bar{h}_{s,i}\right),\ a_{t,i} = \frac{\exp\left(sc_{t,i}\right)}{\sum_{l=1}^{n} \exp\left(sc_{t,l}\right)},\\
c_t &= \sum_{i=1}^{n} a_{t,i} \bar{h}_{s,i},\ y_t = \text{FC}(c_t || h_{d,t} || RC_t),
\end{aligned}$$

where $t \in \{1, 2, \ldots, n\}$, $f_{\text{MM-CNN}}$ is the multi-channel multi-modal CNN features from temporal and spatial learning, $\bar{h}_s$ is the source hidden states from the encoder for all n epochs in the sequence, $h_{d,t}$ is the hidden state of the decoder at epoch t and y_t is the predicted output for epoch t. Moreover, $a_{t,i}$ is the SoftMaxed alignment score between $h_{d,t}$ and $\bar{h}_{s,i}$ and calculated by the additive similarity function [41] where W_d and W_s are trainable weights. c_t is the context vector for epoch t. RC_t is the residual connection over $f_{\text{MM-CNN}}$ and $||$ denotes the concatenation operation. For every sequence, the decoder input for the first epoch, y_0, was set to the true label of the last epoch of the previous sequence. Exception to this rule is the first sequence of a new PSG, where a zero label was initialized as the starting decoder input.

4) Residual Connections: We added a residual connection that concatenates CNN features to RNN features in order to consider temporal and sequential information evenly in the final classification of sleep stages. The residual connection employs a fully-connected layer to map CNN features into a feature vector, RC_t, which shares the same dimension of RNN features. Then, both features are concatenated side-by-side to address: i) data imbalance arising in the sequential learning as data balancing techniques discussed in Section IV-C were only employed in the training process for CNNs and not for RNNs and ii) possible information loss of temporal features when the model was trained for sequential features.

C. Addressing Class Imbalance

In PSGs, stages N1 and N3 usually occur much less frequently, yielding imbalanced data sets (cf., Table III). Moreover, complex deep neural networks are often biased to detecting majority classes better than minority classes [42]. To guarantee that all classes can be learnt equally, we employed two data balancing techniques: applying the weighted loss function (WLF) in the training process and oversampling (OS) the instances of minority classes [43]. For WLF, we calculated the categorical cross entropy loss with the weighted function, $W_c = 1 - N_c/N$, to assign higher loss on minority classes, where W_c is the weight for class c, N_c is the number of instances in class c and N is the total number of instances in all classes. For OS, we duplicated the whole batch of instances of minority classes multiple times until their number of instances were close to the number of instances of the majority class. Then, we randomly duplicated single instances from minority classes again to make the

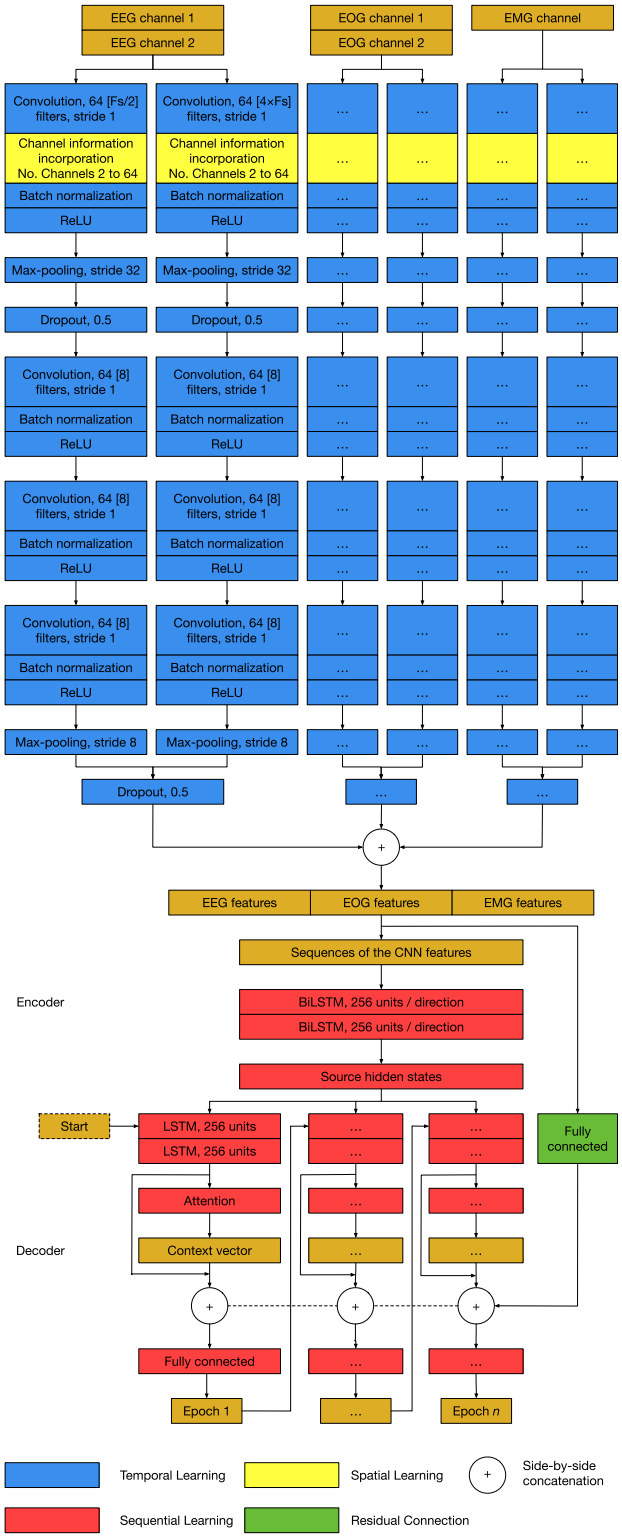


Fig. 2. Improved multi-channel multi-modal model structure, consisting of four components: temporal learning, spatial learning, sequential learning and residual connection. The specifications and parameters of the components are included. Note that, the spatial learning component is an embedded part of the first layer of temporal learning; the yellow blocks highlight its function. Dashed lines deriving from the residual connection indicate that those CNN features are concatenated to RNN features side-by-side for every epoch. *Figure best viewed in color.*

number of instances in all classes exactly the same. Note that, we only applied data balancing techniques during training the temporal and spatial learning components (i.e., in CNNs), as the arrangements of sleep sequences would be invalidated in the sequential learning phase (i.e., in RNNs) if we apply data balancing techniques there.

V. Channel Contribution Analysis

In this section, we present two explainability methods to analyze channel contribution in multi-channel multi-modal sleep scoring. The layer-wise relevance propagation (LRP) [12] is a post-hoc explainability method for model agnostic and extracts information from a trained deep neural network. Although widely applied, post-hoc explainability methods might not be faithful to the underlying model [44]. In contrast, the embedded channel attention network (eCAN), motivated by [13] and [14], learns channel importance intrinsically. In our study, we focused on channel importance in the CNNs of our model structure to exclude context information interactions from neighbouring sleep epochs. Furthermore, we proposed a hypothesis based on the obtained results and subsequently employed channel exclusion experiments to verify our conclusion. All experiments were performed on SHHS-1, as it contains a broad range of research subjects (5,793 subjects) and more channels of sleep signals than SleepEDF-13 (cf., Section III).

A. Layer-Wise Relevance Propagation

We employed the LRP [12] to compute the relevance scores of sleep signals of different channels to represent their importance on predictions. Since we used the ReLU activation layers in CNNs, which are always positive and monotonically increasing, we employed the propagation rule by Montavon et al. [45] to allocate the relevance scores from a current layer k to a preceding layer j. This rule has a positive and a negative contribution term. We focused on the positive one, because it shows channel importance straightforwardly. The relevance scores were then calculated as follows:

$$R_j = \Sigma_k \frac{a_j w_{jk}^+}{\Sigma_j a_j w_{jk}^+} R_k,$$

where R_j and R_k are the relevance scores of the neurons at layer j and k, a_j is the activations of the neurons at layer j and w_{jk}^+ denotes the positive connections of the neurons between layer j and k. Note again, this propagation rule has a constraint that the activations in every preceding layer (including the input data) must be non-negative. However, our input is sleep signals and thus can be negative. To address this problem, we adapted the original rule to

$$R_j = \Sigma_k \frac{(a_j w_{jk})^+}{\Sigma_j (a_j w_{jk})^+} R_k$$

by considering a_j and w_{jk} as a whole. In this way, if the product of the input data and the associated weights in the first layer is positive, the input data has a positive contribution to the output of this layer and is counted. In the experiment, the relevance score of a channel to a prediction was defined as the sum of the relevance of all signal points in the data input of that channel. The channel importance for a particular sleep

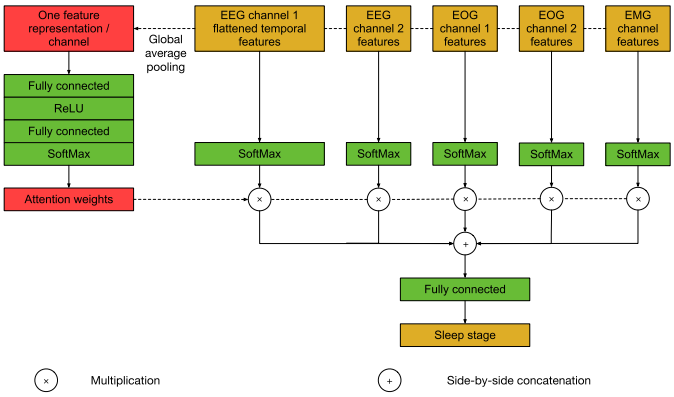


Fig. 3. Structure of the embedded channel attention network (eCAN). Dashed lines connect each channel to the channel attention module and indicate parallel operations over the intermediate data.

stage was obtained by averaging the channel relevance scores over all predicted sleep epochs of that stage.

B. Embedded Channel Attention Network

Our eCAN (cf., Fig. 3) uses a channel attention module which takes the sleep features of all channels as inputs and outputs the attention weights per channel. We used the same CNNs as outlined in Section IV-B to extract temporal features but removed the spatial learning component to capture the individual contribution from each channel. Specifically, the extracted temporal features of each channel of the modalities obtained from CNNs, in a shape of $d \times w$ where d and w respectively denote the depth and width of the feature representations, were first flattened and passed into a global average pooling layer to generate one representative feature per channel. Then, the generated features of all channels were input to a block of two fully-connected layers and a ReLU activation layer, to compute an attention weight for each channel. Note that, this block has the same number of input and output neurons as the number of channels. Next, both the features of each channel and the attention weights were self-normalized using the SoftMax function. The normalized attention weights were multiplied with the normalized features of corresponding channels, resulting in attention weighted features for every channel which were finally passed into another fully-connected layer with the SoftMax activation function for sleep stage classification. We trained the eCAN using the same data balancing techniques as introduced in Section IV-C. Here, the channel importance for a particular sleep stage was obtained from the trained model by averaging the channel attention weights over all predicted sleep epochs of that stage.

C. Verification Using Reverse Ablation

To verify the channel contribution results derived from the LRP and the eCAN, we also performed a reverse ablation study. Similar to the eCAN, we used the same CNNs in Section IV-B and removed the spatial learning component. Then, we excluded one channel of the data input at a time and trained the model on remaining channels. We reported performance decreases in terms of per-class F1-scores for particular stages to illustrate the importance of the excluded channel.

In addition, we also performed the same experiment on the whole model structure including the sequential learning and residual connection components, i.e. on CNNs & RNNs & RC, to investigate the influence of sequential features on compensating for the information loss of the excluded channel. We still focused on performance decreases to identify the positive contribution of a channel, i.e. the information added by incorporating that channel.[5]

VI. Experimental Setup

In this section, we introduce the training scheme, model parameters, evaluation metrics and evaluation designs for our multi-channel multi-modal sleep scoring model.

A. Training Scheme and Model Parameters

We used a two-step training scheme to address the class imbalance problem. In the first step, we pre-trained CNNs (i.e., temporal and spatial learning) via minimizing the categorical cross entropy loss between model predictions and the ground truth. We used one of the two data balancing techniques, WLF and OS, as discussed in Section IV-C. The CNN features were passed into a fully-connected layer with the SoftMax activation function for sleep stage classification. This step enables our model to capture the time-invariant information of a sleep epoch precisely and learn minority classes equally to the majority class. In the second step, we froze the parameters of CNNs and trained RNNs (i.e., sequential learning), the residual connection and the final fully-connected layer. We used the categorical cross entropy loss here again. Note that, in this step, we did not use any data balancing technique.

For both steps, we used early stopping with a patience of 16. We used Adam [46] as the optimizer and set the learning rate to 10^{-4}, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ in both training steps. Following [24], we set the mini-batch size to 192 segments of 30-s sleep epochs, as a sleep cycle usually lasts around 96 minutes. We expected that one mini-batch training can cover all classes of sleep stages. For training RNNs and the residual connection, we set the mini-batch size to 24 and the sequence length to 8. However, the number of epochs in a PSG may not be exact multiples of 8. To still include the last epochs in the training, validation and test set, we padded them with the starting epochs of the same PSG. Our models were implemented using PyTorch and the source code is publicly available: https://github.com/Bobby-Lu/Analyzing-channel-contribution-in-multi-channel-multi-modal-sleep-scoring.

B. Evaluation Metrics

We used commonly used evaluation metrics to report the performance of our multi-channel multi-modal sleep scoring

[5]When the performance increased after excluding a channel, i.e. the channel had a negative influence (e.g., adding noise) on predictions, we set its contribution to zero.

model (e.g., [6], [7]): accuracy (Acc), macro F1-score (MF1), Cohen's kappa (κ) and per-class F1-score (pF1). Among them, MF1 is the harmonic mean of precision and recall and reflects the detection performance on minority classes. κ measures the agreement between a model and the ground truth. pF1 shows the detection performance for specific classes. We list the formulas to calculate them as follows:

$$Acc = \frac{\Sigma_{c=1}^{C} TP_c}{N}, \quad MF1 = \frac{\Sigma_{c=1}^{C} pF1_c}{C}, \quad \kappa = \frac{p_o - p_e}{1 - p_e},$$
$$p_e = \Sigma_{c=1}^{C} \frac{n_{cg}}{N} \frac{n_{cp}}{N}, \quad pF1_c = \frac{2}{Pr_c^{-1} + Re_c^{-1}},$$

where c is one class of sleep stages and C is the number of classes. TP_c is the number of true positives of class c. N is the total number of epochs. $pF1_c$ is the per-class F1-score of class c. p_o is the relative agreement between the ground truth and the predictions; p_e integrates the hypothetical probability of chance agreement, obtained from the number of sleep epochs in the ground truth for a class, n_{cg}, and the number of epochs in the predictions for that class, n_{cp}. Pr_c is the precision of class c and Re_c is the recall of class c.

C. Evaluation Designs

We used different evaluation designs for SleepEDF-13 and SHHS-1, because the two data sets differ greatly in size. The SleepEDF-13 data set only contains 20 subjects with 39 PSGs. Hence, we used a nested cross validation scheme. The outer loop is a 20-fold cross validation corresponding to 20 subjects used to estimate the global performance. This means, in every outer fold k, we left out one of the 20 subjects as the test subject at a time. At the end, we combined the results of all 20 test subjects. Every inner loop is a 10-fold cross validation used to optimize the trainable weights of the models. We trained 10 models on 10 training-validation data combinations and tested them on the data of subject k. Finally, we combined the results of 200 sets (i.e., 20 outer loops $\times$ 10 inner loops) and calculated performance metrics on the global confusion matrix. For the SHHS-1 data set, we randomly shuffled the 5,793 subjects and split the data set into training (81%), validation (9%) and test (10%) following [24]. We trained our model on the training set, used early stopping on the validation set and reported model performance on the test set. Note that, the subjects were always kept separate to prohibit information leakage.

VII. Results & Discussion

In this section, we present and discuss the results for the performance of our multi-channel multi-modal sleep scoring model and for the channel contribution obtained by the LRP and the eCAN methods. We also show the results of channel exclusion experiments.

A. Sleep Scoring Performance

Table V gives an overview of the performance results. Our best model variant achieved an accuracy of 87.2% & 89.1%, a macro F1 score of 82.1% & 81.4% and a Cohen's kappa of 0.82 & 0.85 on the small SleepEDF-13 and large SHHS-1 data sets, respectively. We observe that adding transition information from sleep sequences helps complement the insufficiency of temporal information from sleep epochs thus improves the detection of sleep stages, especially for stages N1, N3 and REM. The two data balancing techniques, WLF and OS, showed comparable impacts on the performance, whereas the latter is much more computationally expensive in the training process. Nevertheless, the minority classes, stages N1 and N3, were still difficult to detect; the prediction of the other three stages achieved a F1-score around 90% for respective classes.

Compared to the state-of-the-arts, our model outperformed all single-channel EEG and multi-channel multi-modal models that were based on classic CNN-RNN architectures. Moreover, comparing our model to SalientSleepNet [26] which used the advanced U-Net architecture, we observe close albeit slightly lower performance (i.e., 87.5% Acc vs. 87.2% Acc), showing that classic CNN-RNN architectures are still competitive for multi-channel multi-modal sleep scoring. SalientSleepNet relied heavily on inter-modality attention modules to minimize the redundancies in data streams, which may be extended to build on our conclusion (cf., Section VIII) to reduce inter-channel redundancies as well. However, note that, the focus of this paper is to investigate channel contribution and not to propose a novel multi-channel sleep scoring model that outperforms the state-of-the-arts.

To conclude, adding particular model features from single-channel EEG models (cf., Table IV) improves multi-channel multi-modal models. The improved model outperforms previous single-channel EEG and multi-channel multi-modal models. Compared to the best performing single-channel EEG models, the advantage of incorporating the information of multiple modalities and channels is around 2% Acc on both SleepEDF-13 and SHHS-1. The result suggests that the information of part of the modalities and channels may be sufficient to obtain accurate predictions for sleep scoring.

B. Channel Contribution

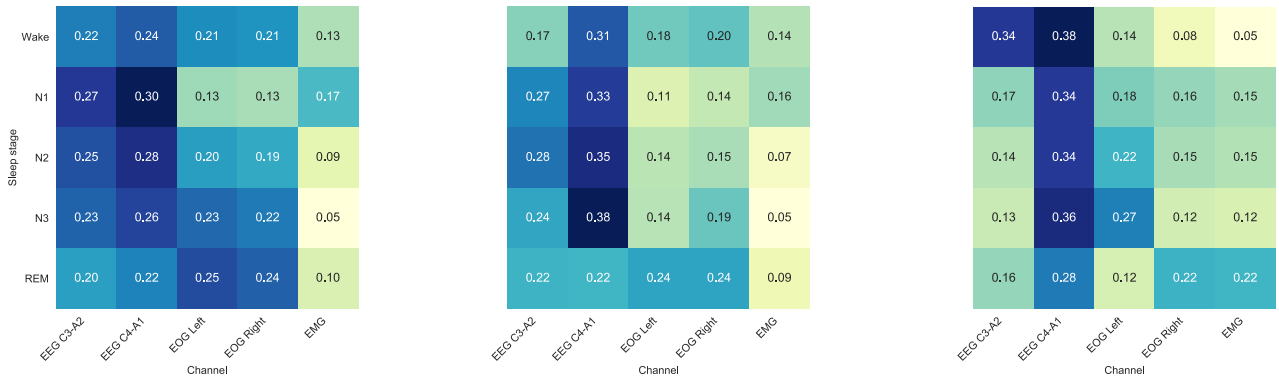
The proposed channel contribution experiments were based on 60 randomly selected subjects[6] from the training set of SHHS-1. For the LRP, we computed channel importance for both models, including and excluding the spatial learning component. Comparing Fig. 4b and Fig. 4c, we observe that both the LRP and the eCAN attributed information usage to all channels, which suggests that deep sleep scoring models try to utilize all accessible information. However, the EEG modality achieved much higher importance scores than the EOG and EMG modalities, complying with the AASM manual [2]. More specifically, both methods identified channel C4-A1 as the most important EEG channel, which is also recommended in [2] with channel C3-A2 being a backup for channel C4-A1. However, the two methods gave vague views on the most important EOG channel, matching the fact that the sleep signals of the 2 EOG channels in SHHS-1 are

[6]We used 60 subjects only, for computational reasons.

TABLE V

PERFORMANCE COMPARISON OF OUR MULTI-CHANNEL MULTI-MODAL MODEL AND THE STATE-OF-THE-ARTS ON SLEEPEDF-13 AND SHHS-1. 'CNN' AND 'CNN-RNN' IN THE '**MODEL**' COLUMN CORRESPOND TO THE MODELS OBTAINED IN THE TWO TRAINING STEPS INTRODUCED IN SECTION VI-A; 'WLF' AND 'OS' REFER TO THE TWO DATA BALANCING TECHNIQUES DISCUSSED IN SECTION IV-C. BEST VALUES AMONG THE MODELS BASED ON CLASSIC CNN-RNN ARCHITECTURES ARE MARKED IN BOLD. THE MODEL WHOSE METRICS ARE MARKED IN ITALICS USED ADVANCED ARCHITECTURES AND OUTPERFORMED OUR MODEL. "–" DENOTES THAT THE VALUE IS NOT AVAILABLE IN THE RESPECTIVE PUBLICATION

Dataset	Model	Channels	Input	Evaluation	Acc	MF1	κ	Wake	N1	N2	N3	REM
SleepEDF-13	Supratak et al. [6]	1EEG	raw	20-fold CV	82.0	76.9	0.76	84.7	46.6	85.9	84.8	82.4
	Mousavi et al. [7]	1EEG	raw	20-fold CV	84.3	79.7	0.79	89.2	52.2	86.8	85.1	85.0
	Seo et al. [17]	1EEG	raw	20-fold CV	83.9	77.6	0.78	87.7	43.4	87.7	86.7	82.5
	Supratak et al. [16]	1EEG	raw	20-fold CV	85.4	80.5	0.80	90.1	51.4	88.5	88.3	84.3
	Eldele et al. [19]	1EEG	raw	20-fold CV	85.6	80.9	0.80	90.3	47.9	**89.8**	**89.0**	85.0
	Paisarnsrisomsuk et al. [21]	2EEGs+1EOG	raw	4-fold CV	81.2	72.3	0.73	57.3	45.8	87.3	87.2	83.8
	Jia et al. [26]	1EEG+1EOG	raw	20-fold CV	*87.5*	*83.0*	-	*92.3*	*56.2*	*89.9*	*87.2*	*89.2*
	own (CNN,WLF)	2EEGs+1EOG+1EMG	raw	20-fold nested CV	81.0	75.7	0.74	88.4	44.4	84.4	80.0	81.2
	own (CNN,OS)	2EEGs+1EOG+1EMG	raw	20-fold nested CV	74.4	70.8	0.67	83.5	40.4	79.7	74.0	76.2
	own (CNN-RNN,WLF)	2EEGs+1EOG+1EMG	raw	20-fold nested CV	**87.2**	**82.1**	**0.82**	92.2	**54.1**	89.0	84.6	**90.6**
	own (CNN-RNN,OS)	2EEGs+1EOG+1EMG	raw	20-fold nested CV	87.1	81.9	**0.82**	**92.4**	53.0	88.9	84.5	90.5
SHHS-1	Sors et al. [15]	1EEG	raw	50-20-30	86.8	78.5	0.81	91.0	42.7	87.9	85.0	85.4
	Seo et al. [17]	1EEG	raw	50-20-30	86.7	79.8	0.81	90.1	48.1	88.4	**85.2**	87.2
	Pathak et al. [24]	2EEGs+2EOGs+1EMG	raw	81-9-10	85.0	76.6	0.79	92.1	41.3	84.8	76.3	88.7
	own (CNN,WLF)	2EEGs+2EOGs+1EMG	raw	81-9-10	81.6	73.0	0.74	89.8	35.0	82.2	76.6	81.2
	own (CNN,OS)	2EEGs+2EOGs+1EMG	raw	81-9-10	77.1	68.8	0.68	89.4	32.0	80.5	68.6	73.5
	own (CNN-RNN,WLF)	2EEGs+2EOGs+1EMG	raw	81-9-10	**89.1**	**81.4**	**0.85**	**94.2**	48.1	**89.4**	82.2	**92.9**
	own (CNN-RNN,OS)	2EEGs+2EOGs+1EMG	raw	81-9-10	**89.1**	**81.4**	**0.85**	**94.2**	**48.6**	**89.4**	82.0	**92.9**
	own (CNN,WLF)	1EEG+1EOG+1EMG	raw	81-9-10	81.4	71.9	0.74	89.6	32.4	82.3	75.6	79.8
	own (CNN-RNN,WLF)	1EEG+1EOG+1EMG	raw	81-9-10	88.9	80.9	0.84	93.9	46.6	89.2	82.1	92.6



(a) LRP, including spatial learning (b) LRP, excluding spatial learning (c) eCAN

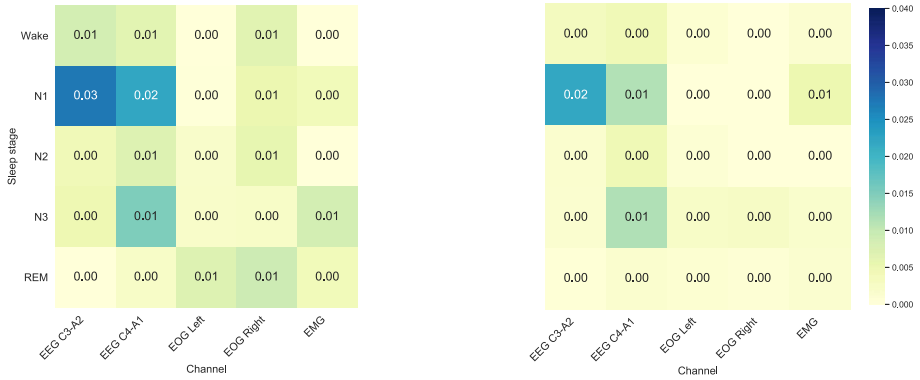
Fig. 4. Channel contribution obtained by the LRP applied to the CNNs of our multi-channel multi-modal model including (a) and excluding (b) the spatial learning component, and channel contribution obtained by the eCAN (c).

collected from symmetric sensors and thus contain similar information [2], [32].

Observations on the most important EEG and EOG channels suggest that deep multi-channel multi-modal sleep scoring models may select one channel per modality as their main feature sources and use other channels to complement the information. Moreover, compared to the model without the spatial learning component, the one with that tried to give more even attention to different channels in a modality (cf., Fig. 4a and Fig. 4b), which reflects the intention of spatial learning for feature attributions, i.e. aggregating information from multiple channels. Furthermore, the LRP and the eCAN are different feature attribution methods (i.e., post-hoc vs. intrinsic). The specific importance scores of channels to particular sleep stages identified by them varied slightly. For instance, the LRP worked more naturally in assigning larger importance scores to the EOG channels for stages Wake and REM, as they contain more eye movements [2]. Similar patterns can also be found in the relations between the EMGs and stages Wake and N1. Despite this, the two methods both showed a general pattern: multi-channel multi-modal models mainly rely on a single important channel per modality, which suggests that *incorporating all channels may not be necessary to obtain acceptable prediction performance*.

C. Hypothesis Verification

Fig. 5 presents the performance decreases when excluding single channels of different modalities. Overall, the results verify the hypothesis above: mostly, one channel per modality is relevant for model predictions. Specifically, excluding EEG channels, especially the EEG C4-A1 channel, resulted in a larger performance decrease than excluding others. The EOG left and right channels led to almost identical performance decreases. However, we observe that the exact difference of the decreases caused by excluding different channels of a modality



(a) CNN (b) CNN-RNN

Fig. 5. Decreases in pF1's obtained by channel exclusion experiments applied to (a) CNNs and (b) CNNs & RNNs & RC. The spatial learning component is excluded.

is rather small. In addition, comparing Fig. 5a and Fig. 5b, it is interesting to find that CNNs & RNNs & RC achieved less performance decreases when excluding single channels than CNNs. This indicates that CNNs & RNNs & RC recovered from excluding channels and points toward that the addition of sequential information compensates for the information loss of the excluded channel. We thus conclude adding transition information is beneficial especially when only a few channels are available.

For the sake of completeness, we additionally tested whether one channel per modality is really sufficient for acceptable prediction performance. We trained our multi-channel multi-modal model on SHHS-1 only with the identified important channel in each modality (the EEG C4-A1 channel, the EOG left channel and the EMG channel). Results are shown in the last two rows in Table V. Compared to the original multi-channel multi-modal model (i.e., rows above), the best κ only dropped from 0.85 to 0.84. This indicates that incorporating multiple channels per modality, while increasing the number of parameters to train, does not improve the performance much. The most predictive information is contained in single important channels of different modalities.

VIII. Conclusion and Future Work

In this paper, we investigated to which extent multi-channel multi-modal sleep scoring models utilize information from different channels of multiple modalities. To obtain a state-of-the-art multi-channel multi-modal model, we first tested the prospective impacts of particular model features from high-performing single channel EEG models on the performance in the multi-channel multi-modal setting. We found that all four model features presented in Table IV improve the performance. Second, we employed two explainability methods, the LRP and the eCAN, to extract channel importance in multi-channel multi-modal sleep scoring. We found that deep learning based multi-channel multi-modal models incorporate information from all accessible channels but tend to focus on one important channel per modality and use the remainders to complement information. We verified this hypothesis in a reverse ablation study, where we retrained the multi-channel multi-modal model by excluding single channels of different modalities. Overall, the performance difference between single-channel EEG and multi-channel multi-modal approaches is still rather small, indicating that while additional channels contain useful information, current multi-channel multi-modal models under classic CNN-RNN architectures may not be able to reliably use the predictive information from additional channels but may also get distracted by the confusing information or the noise from other channels.

The first direction in the future would be to evaluate the channel contribution results on a sleep data set with many channels per modality (e.g., Montreal Archive of Sleep Studies [47]). Moreover, based on our obtained empirical results, the second direction would be to analyze channel contribution deeply, combining deep learning based predictions and actual sleep mechanisms. The validated hypothesis can then be utilized for efficient sleep scoring in small sleep study laboratories: i) collecting and using only important channels of the modalities and ii) training small deep learning models on a limited amount of research subjects. Additionally, considering that sleep is not a global homogeneous event in the brain, another interesting direction is to design multi-channel multi-modal sleep scoring models that, while learning from the channels that contain the most important predictive information, can incorporate additional predictive information from other channels but do not additionally learn the distractors.

References

[1] A. Rechtschaffen and A. Kales, "A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects," *Arch. Gen. Psychiatry*, vol. 20, no. 2, pp. 246–247, 1968.

[2] C. Iber, S. Ancoli-Israel, A. L. Chesson, and S. F. Quan, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. Darien, IL, USA: American Academy of Sleep Medicine, 2007.

[3] A. R. Hassan and M. I. H. Bhuiyan, "A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features," *J. Neurosci. Methods*, vol. 271, pp. 107–118, Sep. 2016.

[4] T. L. T. da Silveira, A. J. Kozakevicius, and C. R. Rodrigues, "Single-channel EEG sleep stage classification based on a streamlined set of statistical features in wavelet domain," *Med. Biol. Eng. Comput.*, vol. 55, no. 2, pp. 343–352, Feb. 2017.

[5] M. Sharma, D. Goyal, P. V. Achuth, and U. R. Acharya, "An accurate sleep stages classification system using a new class of optimally time-frequency localized three-band wavelet filter bank," *Comput. Biol. Med.*, vol. 98, pp. 58–75, Jul. 2018.

[6] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 1998–2008, Nov. 2017.

[7] S. Mousavi, F. Afghah, and U. R. Acharya, "SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PLoS ONE*, vol. 14, no. 5, May 2019, Art. no. e0216456.

[8] W. Wang, P. Liao, Y. Sun, G. Su, S. Ye, and Y. Liu, "Automatic sleep stage classification using marginal Hilbert spectrum features and a convolutional neural network," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 625–628.

[9] L. Fiorillo et al., "Automated sleep scoring: A review of the latest approaches," *Sleep Med. Rev.*, vol. 48, Dec. 2019, Art. no. 101204.

[10] S. Khalighi, T. Sousa, G. Pires, and U. Nunes, "Automatic sleep staging: A computer assisted approach for optimal combination of features and polysomnographic channels," *Exp. Syst. Appl.*, vol. 40, no. 17, pp. 7046–7059, Dec. 2013.

[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 Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 4, pp. 758–769, Apr. 2018.

[12] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140.

[13] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[14] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11534–11542.

[15] A. Sors, S. Bonnet, S. Mirek, L. Vercueil, and J.-F. Payen, "A convolutional neural network for sleep stage scoring from raw single-channel EEG," *Biomed. Signal Process. Control*, vol. 42, pp. 107–114, Apr. 2018.

[16] A. Supratak and Y. Guo, "TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 641–644.

[17] H. Seo, S. Back, S. Lee, D. Park, T. Kim, and K. Lee, "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*, vol. 61, Aug. 2020, Art. no. 102037.

[18] C. Sun, C. Chen, W. Li, J. Fan, and W. Chen, "A hierarchical neural network for sleep stage classification based on comprehensive feature learning and multi-flow sequence learning," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 5, pp. 1351–1366, May 2020.

[19] E. Eldele et al., "An attention-based deep learning approach for sleep stage classification with single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 809–818, 2021.

[20] L. Fiorillo, P. Favaro, and F. D. Faraci, "DeepSleepNet-lite: A simplified automatic sleep stage scoring model with uncertainty estimates," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2076–2085, 2021.

[21] S. Paisarnsrisomsuk, M. Sokolovsky, F. Guerrero, C. Ruiz, and S. A. Alvarez, "Deep Sleep: Convolutional neural networks for predictive modeling of human sleep time-signals," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, London, U.K., Aug. 2018.

[22] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Joint classification and prediction CNN framework for automatic sleep stage classification," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1285–1296, May 2019.

[23] Z. Jia et al., "GraphSleepNet: Adaptive spatial–temporal graph convolutional networks for sleep stage classification," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 1324–1330.

[24] S. Pathak, C. Lu, S. B. Nagaraj, M. van Putten, and C. Seifert, "STQS: Interpretable multi-modal Spatial-Temporal-seQuential model for automatic Sleep scoring," *Artif. Intell. Med.*, vol. 114, Apr. 2021, Art. no. 102038.

[25] H. Phan, O. Y. Chén, M. C. Tran, P. Koch, A. Mertins, and M. De Vos, "XSleepNet: Multi-view sequential model for automatic sleep staging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5903–5915, Sep. 2022.

[26] Z. Jia, Y. Lin, J. Wang, X. Wang, P. Xie, and Y. Zhang, "SalientSleepNet: Multimodal salient wave detection network for sleep staging," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, 2021, pp. 1–10.

[27] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," 2015, *arXiv:1506.06579*.

[28] M. Böhle, F. Eitel, M. Weygandt, and K. Ritter, "Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification," *Frontiers Aging Neurosci.*, vol. 11, p. 194, Jul. 2019.

[29] A. A. Bastidas and H. Tang, "Channel attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 881–888.

[30] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000.

[31] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 9, pp. 1185–1194, Sep. 2000.

[32] S. F. Quan et al., "The sleep heart health study: Design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–1085, 1997.

[33] A. Malafeev et al., "Automatic human sleep stage scoring using deep neural networks," *Frontiers Neurosci.*, vol. 12, p. 781, Nov. 2018.

[34] S. P. Patil, "What every clinician should know about polysomnography," *Respiratory Care*, vol. 55, no. 9, pp. 1179–1195, 2010.

[35] C. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[36] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, Jul. 2015, pp. 448–456.

[37] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[38] Y. Li, X. Tang, Z. Xu, W. Liu, and J. Li, "Temporal correlation between two channels EEG of bipolar lead in the head midline is associated with sleep-wake stages," *Australas. Phys. Eng. Sci. Med.*, vol. 39, no. 1, pp. 147–155, Mar. 2016.

[39] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*.

[40] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*.

[41] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[42] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[43] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, "A survey on addressing high-class imbalance in big data," *J. Big Data*, vol. 5, no. 1, p. 42, Dec. 2018.

[44] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "The dangers of post-hoc interpretability: Unjustified counterfactual explanations," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 2801–2807.

[45] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process.*, vol. 73, pp. 1–15, Feb. 2018.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[47] C. O'Reilly, N. Gosselin, J. Carrier, and T. Nielsen, "Montreal archive of sleep studies: An open-access resource for instrument benchmarking and exploratory research," *J. Sleep Res.*, vol. 23, no. 6, pp. 628–635, 2014.