STQS

Pathak, Shreyasi; Lu, Changqing; Nagaraj, Sunil Belur; van Putten, Michel; Seifert, Christin

Published in:
Artificial Intelligence in Medicine

DOI:
10.1016/j.artmed.2021.102038

Publication date:
2021

Artificial Intelligence In Medicine 114 (2021) 102038

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Position Paper

STQS: Interpretable multi-modal *S*patial-*T*emporal-se*Q*uential model for automatic *S*leep scoring

Shreyasi Pathak [a,*], Changqing Lu [a], Sunil Belur Nagaraj [c], Michel van Putten [a,b], Christin Seifert [a,d]

[a] *University of Twente, Netherlands*
[b] *Medisch Spectrum Twente, Netherlands*
[c] *University Medical Center Groningen, Netherlands*
[d] *University of Duisburg-Essen, Germany*

ARTICLE INFO

Keywords:
Sleep scoring
Sleep stage annotation
Deep learning
EEG, EOG, EMG signals
Post-hoc interpretability
Explainable AI

ABSTRACT

Sleep scoring is an important step for the detection of sleep disorders and usually performed by visual analysis. Since manual sleep scoring is time consuming, machine-learning based approaches have been proposed. Though efficient, these algorithms are black-box in nature and difficult to interpret by clinicians. In this paper, we propose a deep learning architecture for multi-modal sleep scoring, investigate the model's decision making process, and compare the model's reasoning with the annotation guidelines in the AASM manual. Our architecture, called STQS, uses convolutional neural networks (CNN) to automatically extract spatio-temporal features from 3 modalities (EEG, EOG and EMG), a bidirectional long short-term memory (Bi-LSTM) to extract sequential information, and residual connections to combine spatio-temporal and sequential features. We evaluated our model on two large datasets, obtaining an accuracy of 85% and 77% and a macro F1 score of 79% and 73% on SHHS and an in-house dataset, respectively. We further quantify the contribution of various architectural components and conclude that adding LSTM layers improves performance over a spatio-temporal CNN, while adding residual connections does not. Our interpretability results show that the output of the model is well aligned with AASM guidelines, and therefore, the model's decisions correspond to domain knowledge. We also compare multi-modal models and single-channel models and suggest that future research should focus on improving multi-modal models.

1. Introduction

Polysomnograms (PSGs) are recordings of body activities collected during sleep to aid the diagnosis of sleep disorders. A PSG typically encompasses ≈8 h of different signals, e.g., electroencephalograms (EEGs), electrooculograms (EOGs), electromyograms (EMGs). PSGs are analyzed and annotated by sleep technologists[1] [1] based on sleep annotation guidelines [2,3]. Sleep technologists annotate each 30 s interval separately, resulting in ≈960 manual annotations (approx. 2–3 h annotation time) per PSG. Automatic approaches aim to make the annotation process more efficient, and can be coarsely divided into traditional machine learning (e.g., [4–6]) and deep learning approaches (e.g., [7,8]). The latter have the advantage of automatically extracting features and have been shown to outperform traditional approaches [9].

Deep learning approaches for sleep scoring apply architectures and techniques from deep learning on general domains, e.g., by directly applying convolutional neural network (CNN)-based architectures [10], incorporating sequential information via long short-term memory (LSTM) [11] networks, or adding residual connections (RC) [11]. However, the individual contribution of these architectural components have not yet been investigated. Pioneering deep learning approaches for sleep scoring base their prediction on single modality inputs, i.e., only EEG [11]. Subsequent work provides evidence that incorporating multiple modalities (e.g., EOG and EMG) improves performance [12]. Though very efficient in performance, a general downside of deep learning is its black-box nature, which hinders its adoption in clinical settings. Research in eXplainable Artificial Intelligence (XAI) [13] aims to make these black-boxes more transparent, e.g., by using post-hoc

* Corresponding author.
E-mail address: s.pathak@utwente.nl (S. Pathak).

[1] Sleep technologists are professionals performing polysomnography, including sleep stage scoring.

https://doi.org/10.1016/j.artmed.2021.102038
Received 16 July 2020; Received in revised form 27 January 2021; Accepted 16 February 2021
Available online 27 February 2021

interpretability methods to explain the outcome of such models [14].

In this work, we propose a deep learning architecture, *Spatio-Temporal-seQuential-Sleep-scoring (STQS)*, for multi-modal, multi-channel input data, designed to account for spatial, temporal and sequential information in the signals and evaluate the contribution of various architectural components. We apply post-hoc interpretability methods to investigate the alignment of the model with scoring guidelines [2], and evaluate the contribution of different modalities for the prediction. Specifically, our contributions are:

1. We show how to leverage spatio-temporal and sequential information from multi-modal, multi-channel input signals in a deep neural network and evaluate the effect of adding sequential information, information transfer via residual connections and various class-imbalance techniques (results in Section 7.1).
2. We evaluate our model on a public benchmark dataset (SHHS, 5793 subjects) and an in-house dataset (1418 subjects) and compare it to multiple baselines (experiments in Section 7.1). We investigate the importance of multiple modalities using post-hoc interpretability methods (results in Section 7.2).
3. We show the model's alignment with AASM sleep scoring guidelines [2] by applying 3 different methods of post-hoc interpretability: frequency-domain occlusion, time-domain occlusion, and pattern visualization of temporal filters in the CNN (results in Section 8).

The remainder of this paper is organized as follows. Sections 2 and 3 introduce the state-of-the-art and datasets. Section 4 explains STQS, Section 5 discusses our post-hoc interpretability approach and Section 6 reports the experimental setup. Prediction results of our model are reported in Section 7, while Section 8 investigates the model's reasoning. We discuss implications of our results in Section 9 and conclude in Section 10.

2. Related work

In this section, we describe manual sleep scoring in detail, review traditional and deep-learning based approaches to automatic sleep scoring and review explainable AI methods.

2.1. PSGs and manual sleep scoring

A PSG is a sleep study, for which signals like EEG, EOG, EMG, electrocardiograms and leg movement are recorded from a patient. Humans experience 5 stages during sleep: Wake (W), Rapid Eye Movement (REM), Non-REM stage 1 (N1), Non-REM stage 2 (N2) and Non-REM stage 3 (N3). The analysis of sleep stages is crucial for the detection of sleep disorders, e.g., the periodic leg movement syndrome. The signals are collected for an 8 h period, i.e., a whole night of sleep and divided into 30 s epochs. Each epoch is annotated with a sleep stage according to the American Academy of Sleep Medicine (AASM) [2] or Rechtschaffen and Kales (R&K) [3] sleep manual. This annotation process is called sleep scoring and is usually based on EEG, EOG and EMG signals only. The sleep stages are annotated based on distinctive characteristics of the signals (cf. Table 1 for the AASM manual characteristics). The EEG signals per stage vary in amplitude, frequency and exhibit distinctive patterns, e.g., K-complex, or vertex waves. The EOG and EMG signals per stage mainly vary in amplitude. Currently, sleep stages are annotated based on visual inspection by sleep technologists. Annotating one PSG takes about 2–3 h.

2.2. Machine learning for sleep scoring

Traditional Machine Learning approaches rely on expert-defined features capturing temporal, frequency and non-linear properties of the data. Li et al. [5] combined random forests and rules developed from the R&K sleep manual achieving an accuracy of 0.86 and Cohen's kappa (κ) of 0.805 on a dataset with 198 subjects from Cleveland Sleep Study [24,25]. Koley et al. [6] and Lajnef et al. [8] used similar expert-defined features to train a support vector machine. [6] achieved κ of 0.86 on 28 subjects from Center of Sleep Disorder Diagnosis, India and [8] achieved an accuracy of 0.88 on 15 subjects from DyCog Lab, France respectively. Hassan et al. [26] exploited bootstrap aggregation to classify the sleep stages based on statistical moments extracted by tunable-Q wavelet transform, achieving an accuracy of 0.937 on Sleep-EDF13 dataset. Alickovic et al. [27] used discrete wavelet transform to extract features from the EEG channel and then trained an ensemble classifier called rotational support vector machine for sleep stage scoring achieving an accuracy of 0.91 on Sleep-EDF dataset. Experiments by Khalighi et al. [28] indicate that the best performance (accuracy of 0.92) for sleep scoring is obtained using 9 multi-modal EEG, EOG and EMG channels as input.

Table 1
Characteristic features of sleep stages according to the AASM manual. Frequency band information are δ (delta, 0.16–3.99 Hz), θ (theta, 4–7.99 Hz), α (alpha, 8–11.99 Hz), σ (sigma, 12–15.99 Hz) and β (beta, 16–30 Hz).

Stages	EEG		EOG	EMG
	Time domain	Frequency domain		
W	None	α, σ, β	Movement (0.5–2 Hz)	Variable amplitude but is usually higher than sleep stages
N1	vertex sharp waves	θ, α	Slow movement	Lower amplitude than W
N2	k-complex &sleep spindles	θ, σ	No movement; slow movement may persist in some	Lower than W; may be as low as REM
N3	high amplitude (>75 μV)	δ	Usually no movement	Lower than N2; sometimes as low as REM
REM	sawtooth waves	θ, α	Rapid eye movement	Lowest amplitude of all stages

The usage of expert-defined features requires expert and/or domain knowledge. Additionally, the above approaches were evaluated on small datasets only. Malafeev et al. [9] compared traditional machine learning (a combination of random forest and Hidden Markov model), to deep neural networks (combination of convolutional neural networks and long short-term memory networks). They conclude that deep neural networks are superior in their generalization ability. Thus, we decided to focus on deep learning approaches for sleep scoring.

2.3. Deep learning for sleep scoring

Deep learning approaches can be distinguished based on their usage of input modalities. In the following section, we describe approaches on single modalities (usually EEG) and multi-modal approaches (usually EEG, EOG and EMG). Table 2 provides a comprehensive overview of the approaches, grouped by the type of modalities used.

Vilamala et al. [10] used spectrogram images of one EEG channel as input to a pre-trained VGGNet. Supratak et al. [11] developed a single channel EEG model using a CNN, a bidirectional LSTM (Bi-LSTM) and residual connections. Their CNN uses two different filter sizes for capturing both, frequency information and temporal patterns. Mousavi et al. [19] also used single EEG channel input and a similar architecture, with attention mechanism to learn the parts of the sequence to focus on, and a novel loss function to address class imbalance. Biswal et al. [20] used raw input and spectrogram input of multiple EEG channels on an architecture with CNN, Recurrent Neural Network (RNN) and residual connections.

Phan et al. [23] used time-frequency images from each modality (EEG, EOG, EMG) as input to a multi-task CNN model. Their experiments

Table 2
State-of-the-art deep learning approaches in automatic sleep scoring.

Paper	Year	Dataset	PSGs	Channels	Method	Evaluation	Accuracy	Parameters
Tsinalis et al. [17]	2016	Sleep-EDF13	20	1EEG	CNN	20-fold	0.75	140M[1]
Vilamala et al. [10]	2017	Sleep-EDF13	39	1EEG	CNN	75-20-5	0.86	138M[3]
Sors et al. [18]	2018	SHHS-1	5728	1EEG	CNN	50-20-30	0.87	199M[2]
Supratak et al. [11]	2017	Sleep-EDF13/MASS	39/62	1EEG	CNN-BiLSTM-RC	20-fold/31-fold	0.82/0.86	20M[1]
Mousavi et al. [19]	2019	Sleep-EDF13/Sleep-EDF18	39/61	1EEG	CNN-BiLSTM-Attention	20-fold/10-fold	0.84	1.6M[4]
Biswal et al. [20]	2018	SHHS-1/MGH	5804/10,000	2EEG/6EEG	CNN-RNN-RC	90–10	0.78/0.88	n.a.[5]
Fernandaz-Blanco et al. [15]	2020	SHHS-1	5804	2EEG	Depthwise Separable CNN	70-10-20	0.85	16K
Paisarnsrisomsuk et al. [21]	2018	Sleep-EDF13	39	2EEG+1EOG	CNN	4-fold	0.81	4M[4]
Yildrim et al. [22]	2019	Sleep-EDF02/Sleep-EDF18	8/61	1EEG+1EOG	CNN	70-15-15	0.91/0.91*	796K
Chambon et al. [12]	2018	MASS	61	20EEG+2EOG+3EMG	CNN	5-fold	0.83	100K
Phan et al. [23]	2019	Sleep-EDF13/MASS	39/200	1EEG+1EOG+1EMG	CNN	20-fold	0.82/0.84	510K[4]
Malafeev et al. [9]	2018	Univ. Zurich (Healthy)/ Inst. Warsaw (Patients)	43/54	1EEG+2EOG+1EMG	CNN-BiLSTM-RC	70-15-15	0.85/n.a	934K[4]
This work	2020	SHHS-1/MST	5793/1418	2EEG+2EOG+1EMG/ 8EEG+3EOG+1EMG	CNN-BiLSTM-RC	81-9-10	0.85/0.77	98K

* indicates that Wake was the majority class in the dataset and may lead to higher accuracy, as Wake is easier to predict than other sleep stages. Number of parameters are in thousands (K) or million (M); values are taken from the original paper, unless otherwise noted; n.a. numbers are not available in their or other papers; source indicators: [1]from [12], [2]from [15], [3]from [16], [4]calculated (approx. values) from the architectural details given in their paper, [5]architecture details (i.e., dimensions of some layers) not available in the paper.

showed an increase of 4.1% in accuracy when combining EEG and EOG and an additional increase of 1% when adding EMG. Paisarnsrisomsuk et al. [21] used both, EEG and EOG and found that EOG increased the accuracy by 1%. A similar observation was made by Yildrim et al. [22]. Chambon et al. [12] also observed increasing performance when adding multiple modalities. They applied temporal and spatial filtering via CNN to extract features from multi-modal inputs (EEG, EOG and EMG) and also encoded temporal context from neighbouring epochs into their input to the CNN. Their extensive experiments showed that using additional modalities (EOG and EMG) with ≈6 EEG channels improved performance over only using EEG channels. However, there was no improvement in performance on further increasing the number of input EEG channels. Further sleep scoring approaches include an architecture based on depthwise separable convolutional layers [15] achieving an accuracy of 0.85 and a model to address database variability [29]. The latter paper reported the model's cross dataset performance on a private and several public sleep datasets. Their model achieved a kappa of 0.78 on SHHS visit 2 dataset.

While spiking neural networks (SNNs), i.e., third generation neural networks, have been shown to learn prediction based on EEG signals [30], they have not been applied to sleep stage detection yet. We chose to focus on 2nd generation neural networks instead to make use of the large body of work on explainable AI for these types of networks. Since previous work found that multi-modal approaches outperform approaches based on single input modalities, we focus our work on multi-modal approaches. However, most of the previous multi-modal approaches rely on vanilla CNN architectures. Therefore, we investigate architectural components proposed for single-modality networks in a multi-modal settings. More specifically, we investigate (i) spatio-temporal feature extraction with CNNs [12], (ii) using Bi-LSTM to encode sequential information [11] and (iii) using residual connections to explicitly forward information from earlier to later layers in the network [11].

2.4. Explainable AI (XAI)

The purpose of XAI in our work is to justify the decisions of the black-box models by comparing it to existing sleep scoring guidelines [2], such that they can be trusted by the clinicians and adopted in clinical workflows [31]. In this paper, we use the term interpretability instead of explainability for XAI methods [31]. XAI methods can be distinguished into methods that create intrinsically interpretable models, and post-hoc interpretability methods [31]. Examples for the former are bayesian rule lists [32] or generalized additive models [33]. However, such models usually provide less accurate predictions. Post-hoc interpretability methods aim at explaining black-box models (such as Deep Neural Networks), examples are sensitivity analysis [34], layer-wise relevance propagation (LRP) [35], and occlusion techniques [14]. These methods determine the importance of specific input features for a prediction. Occlusion techniques observe the sensitivity of the output on perturbing some features of the input, whereas LRP calculates the relevance score of the input features through backward propagation. Vilamala et al. [10] used sensitivity maps to highlight the input features deemed important by their model for the prediction of a sleep stage. In this work, we apply occlusion and LRP to explain our deep learning models.

3. Datasets and preprocessing

We evaluate our models on two large datasets: the SHHS-1 benchmark dataset (5793 PSGs, ≈6 million epochs) and a dataset collected at Medisch Spectrum Twente (MST), Enschede, Netherlands (1418 PSGs, ≈1.4 million epochs). Table 3 provides an overview of the datasets.

From the Sleep Heart Health Study (**SHHS**) [36,37] we use the data from the first visit (SHHS-1). SHHS-1 contains 5793[2] PSG records (subjects' age $\geq$ 40). The PSGs consist of signals from 2 EEG sensors (C3-A2 and C4-A1) sampled at 125 Hz, 2 EOG sensors (left and right) sampled at 50 Hz and 1 EMG sensor sampled at 125 Hz. Sleep stages are annotated based on the R&K manual [3]: W, N1, N2, N3, N4, REM, Movement and Unscored. We unified the annotations of this data set to comply to AASM annotations [2] by combining N3 and N4 into a single stage N3 and removed the epochs annotated as Movement and Unscored.

The **MST** dataset was collected from Medisch Spectrum Twente,

[2] The dataset from the official website www.sleepdata.org contains 5793 PSGs, while 5804 are mentioned in the study. Data downloaded in Nov. 2018, last accessed Oct. 2020.

Table 3
Dataset overview. Showing number of epochs and class frequencies.

	W	N1	N2	N3	REM	Sum
SHHS	1,691,288	217,583	2,397,460	739,403	817,473	5,863,207
	28.8%	3.7%	40.9%	12.6%	13.9%	
MST	249,614	175,122	613,118	207,438	188,003	1,433,295
	17.4%	12.2%	42.8%	14.5%	13.1%	

Enschede, Netherlands. It contains 1418 PSGs from subjects (30–40% patients and 60–70% healthy), with almost equal number of home and at-hospital tests, including both male and female, aged between 25–70 (with mean around 50). Data was fully anonymized prior to this study. The dataset contains 8 EEG channels (F3-M2, F4-M1, C3-M2, C4-M1, O1-M2, O2-M1, C3-O1, C4-O2), 3 EOG channels (E2-M2, E1-M2, EOG horizontal) and 1 EMG channel, where M1 and M2 are reference electrodes on the earlobes. All channels were sampled at 250 Hz. The sleep stages were annotated by 9 sleep technologists (each PSG was annotated by only one of them) as W, N1, N2, N3 and REM, according to standard AASM guidelines [2].

The datasets were **pre-processed** as follows: All EEG, EOG and EMG channels of both datasets were low-pass filtered at 30 Hz [12], EEG and EOG channels were high-pass filtered at 0.16 Hz and EMG at 10 Hz [9].

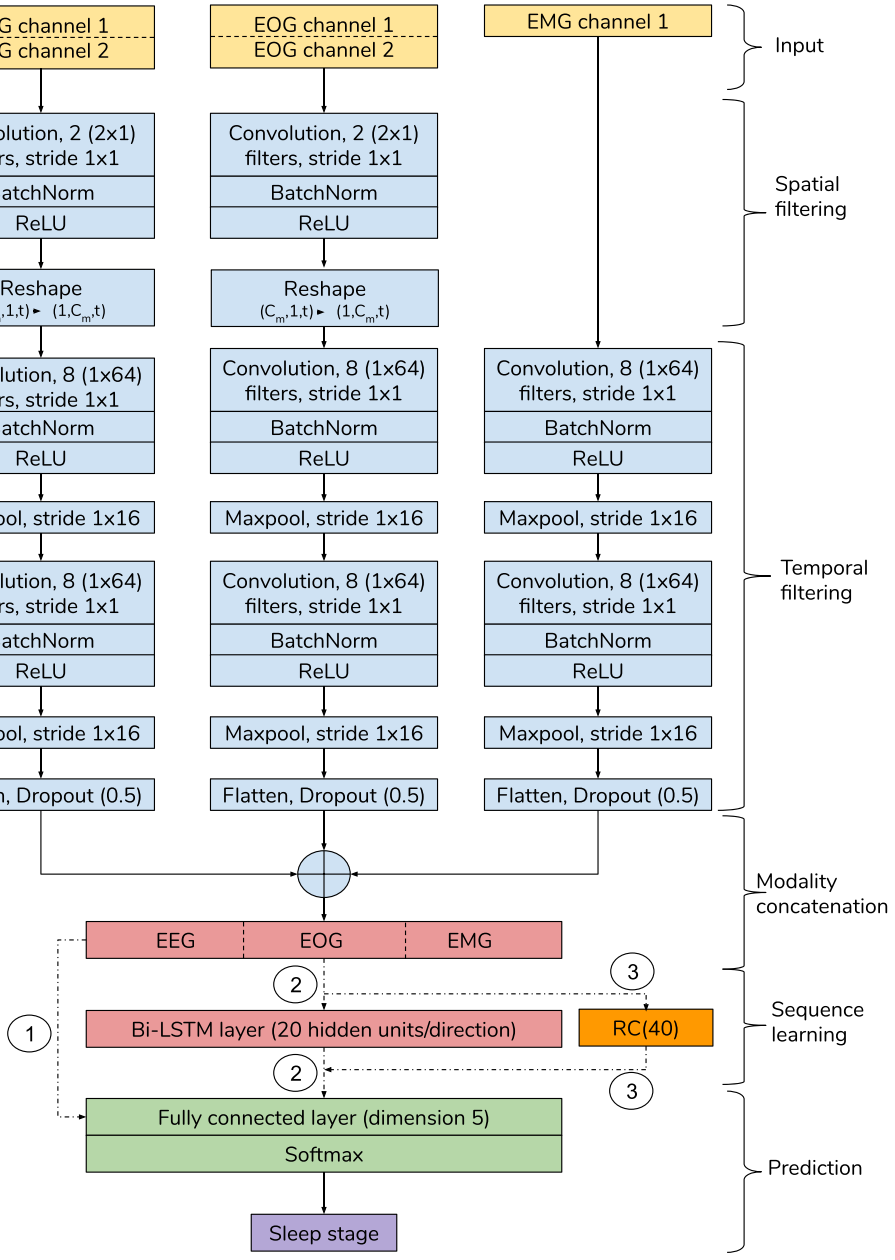


Fig. 1. STQS architecture: Input (yellow, shown for SHHS), CNN (blue), Bi-LSTM (red), RC (orange) and prediction layers (green). Parts connected with solid lines are common in all model variants. Dashed lines indicate connections for various model variants (cf. Section 6). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

All EEG, EOG and EMG channels of both datasets were further resampled to 125 Hz. In SHHS, EEG and EMG were already recorded at 125 Hz, therefore, only EOG channels needed to be upsampled from 50 Hz to 125 Hz using interpolation. In MST, all channels were downsampled from 250 Hz to 125 Hz using decimation. Resampling and filtering the signal are common preprocessing steps followed in existing sleep scoring literature [8,9,11,12]. Each PSG was divided into 30 s epochs (of length 3750 for 125 Hz sampling frequencies). The signal was cropped at the last timesteps, such that the signal length corresponds to multiples of 30 s. Each channel in each epoch was standardized to mean 0 and standard deviation 1.

For training and testing our models, we randomly assign 81% of PSGs for training, 9% for validation and 10% for testing in both the SHHS and MST dataset.

4. Approach for sleep stage prediction

We begin this section with a formal definition of the automatic sleep scoring task, and then describe our model architecture in detail.

4.1. Problem definition

Let $X = (X_1, \ldots, X_t, \ldots, X_T)$ be a multi-modal PSG signal, with corresponding sleep stages $Y = (y_1, \ldots, y_t, \ldots, y_T)$. T is the total number of epochs and X_t is the t^{th} 30 s epoch annotated with sleep stage y_t. X_t contains C_m channels from each modality $m \in \{EEG, EOG, EMG\}$. The task is a multi-class classification problem: predict the sleep stage $\in S = \{W, N1, N2, N3, REM\}$ given X_t. Thus, we aim to learn a prediction function f, such that $\hat{y}_t = \arg\max f(X_t)$. Let y_t be the true class and $\hat{y}_t$ be the prediction. y_t is a 1-hot encoded vector of length $|S|$, with entry 1 for the true class and 0 otherwise. The prediction function f returns a probability distribution over the classes. Our model minimizes the categorical cross entropy loss between $f(X)$ and Y, defined as $L(f(X), Y) = -\sum_{j=1}^{|S|} Y_j log f_j(X)$.

4.2. STQS architecture

For processing multi-modal PSG data, we first apply spatial and temporal filters on all modalities separately, and then combine all modalities. A sequential learning component is added to incorporate longer temporal contexts, i.e., the previous and subsequent epoch. The architecture of STQS is shown in Fig. 1 and described in more detail in the remainder of this section.

4.2.1. Input layer

Our model has a separate feature extraction pipeline for each input modality m. Each 30 s epoch has a shape of $1 \times C_m \times t$ for each modality.[3]

4.2.2. CNNs for spatial filtering

Spatial filtering transforms the C_m raw channels into C_m spatially combined representations of all channels. The spatial filtering component motivated from [12] consists of a convolutional block, with C_m convolutional 2D filters each of shape $C_m \times 1$ that learns the spatial relationship among the channels of a modality. This is followed by batch normalization [38] and rectified linear unit activation ($ReLU(x) = max(0, x)$). Single-channel modalities (EMG, in Fig. 1) are directly passed to the temporal filtering stage. The generated feature vector for each modality after spatial filtering is of shape $C_m \times 1 \times t$ containing C_m spatial representations and is reshaped into $1 \times C_m \times t$ before passing to the temporal filtering component.

[3] We use the notation depth × height × width throughout this paper. This is the standard shape notation for tensors in pytorch.

4.2.3. CNNs for temporal filtering

The feature vector from spatial filtering is further processed by two consecutive convolutional-max pooling blocks to extract the temporal features. The convolutional filters are of size 1×64 (i.e., ≈0.5 s). Each convolutional block has 8 filters, followed by a max pooling of stride 1×16 to reduce the width and retain only the most important features. The filter size was set such that it can capture patterns like the k-complex, which lasts for at least 0.5 s. The generated feature vector is of size $8 \times H_m \times W_m$ for each modality with H_m, W_m being the height and width of the last temporal filtering layer, respectively. This feature vector from each modality is further flattened and dropout [39] is applied with probability 0.5 to prevent overfitting. The resulting feature vectors are horizontally concatenated into a single vector of length $\sum_{m=1}^{3}(8 * H_m * W_m)$.

4.2.4. Bi-LSTMs for sequential learning

Sleep technologists use information from previous and subsequent epochs to annotate one epoch. An example of such an annotation rule is "Epochs without k-complex or spindles following an N2 stage, will be annotated with N2 if the previous epoch contained a k-complex without arousal or a sleep spindle" [2]. In order to capture such rules, we add a sequential learning component. We use a bidirectional LSTM (Bi-LSTM) [40] layer for learning sequential information from both the forward and backward direction of the sequence (containing eight 30 s epochs) with 20 hidden neurons per direction. Let $LSTM_f$, shape 8×20 ($seq\_len \times features$) and $LSTM_b$, shape 8×20 be the feature vector of 8 epochs in the sequence in the forward direction and the backward direction, respectively. The output of the Bi-LSTM is a side-by-side concatenation of $LSTM_f$ and $LSTM_b$ (cf. Fig. 2), generating a feature vector of shape 8×40. The hidden and cell states of the Bi-LSTM layer are initialized with the hidden and cell state values from the last element of the previous sequence for each subject. This initialization, depicted in Fig. 2 with the dashed arrow, incorporates the global state of the signal into the prediction for each sequence. The hidden and cell states are initialized to zero for the first sequence of a new subject.

4.2.5. Residual connections

We use a residual connection block (RC) to add spatio-temporal and sequential features to the final prediction layer. The motivation for using RC is to improve predictions for stages for which temporal features are more important than sequential features. RCs add the multi-modal CNN features (spatio-temporal) to the Bi-LSTM (sequential) features element-wise. To add the feature vectors, the dimension of the CNN feature vector is reduced to the dimension of the Bi-LSTM feature vector using a fully connected layer of dimension 40, followed by batch normalization and ReLU activation.

4.2.6. Final prediction

A fully connected layer with a softmax activation function outputs the final prediction. Since there are 5 sleep stages, the output is of size 5.

4.3. Addressing class imbalance

The class imbalance in the dataset (cf. Table 3) is likely to decrease prediction performance for minority classes [41], i.e., stages N1, N3 and REM. We use two weighted cost functions $w_1(s) = 1 - \frac{N_s}{N}$ and $w_2(s) = \frac{1}{N_s} \cdot \frac{N}{|S|}$, where $w(s)$ is the weight of class $s \in S = \{W, N1, N2, N3, REM\}$, N_s is the number of instances in s and N is the total number of instances. The weighted cost functions result in a higher error for misclassifications on rare classes. For oversampling, we randomly duplicated the instances of all but the majority class, such that the dataset becomes balanced. We apply class imbalance techniques only to the CNN component of our model, to not lose the sequential arrangement of the epochs in a PSG.

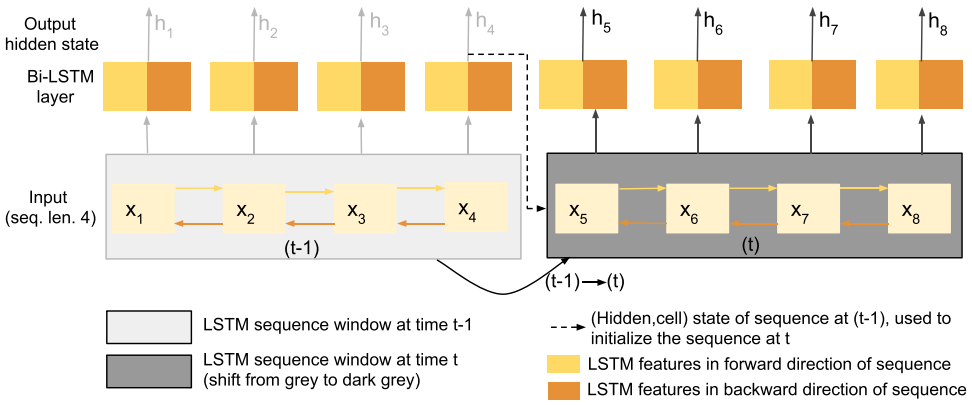


Fig. 2. Bi-LSTM sequence (shown for sequence length of 4). x_i is a 30 s epoch and h_i the corresponding hidden state.

5. Approach for post-hoc interpretability

We applied post-hoc interpretability methods to understand how the black-box models perform their prediction. Specifically, we are interested in the following 4 questions: (i) To what extent do different modalities (EEG, EOG, EMG) contribute to the prediction? (ii) What are the prediction-relevant patterns in the EEG signal? (iii) Which frequency bands of an EEG signal are most important? (iv) How do different temporal filters in the CNN contribute to the final decision? Are there stage-specific filters? Additionally, we are interested, whether our findings align with the AASM guidelines (cf. Table 1).

5.1. Modality importance

We occlude different combinations of modalities *(modality occlusion)* by setting the amplitudes of all channels of those modalities to 0 while keeping the other modalities unchanged [31]. The occluded epoch in the test set is sent to the trained model for prediction. The influence of the modalities is analyzed by comparing the results with occlusion to the results without occlusion on the same epochs.

5.2. Predictive patterns in EEG

To identify the most important EEG patterns, we occlude parts of the 30 s EEG epoch in the time-domain *(time-domain occlusion)*. We use a 5 s sliding window with a 1 s shift along the 30 s epoch. The signal within this occlusion window is set to 0. As we standardized the original epoch with mean 0, this choice represents an inactive signal without changing statistical properties. For each location of the occlusion window, we record the prediction of the model epoch and compare it with the prediction on the original epoch. The motivation is that the prediction of the epoch is likely to change if the occlusion window "hides" a pattern which is important for the prediction.

The idea is sketched in Fig. 3. The occlusion window w_k, $k \in \{1, \ldots, 26\}$ is shifted over the epoch. To consider all consecutive 5 s, we use overlapping occlusion windows. However, this resulted in any two consecutive occlusion windows (e.g., w_1, w_2) having an overlap of 4s. This overlap makes it hard to uniquely quantify the contribution of the occlusion window towards prediction. Therefore, we divide each epoch into non-overlapping 1 s *patterns*, x_p, and calculate the pattern-based importance PI_p, $p \in \{1, \ldots, 30\}$: $\mathrm{PI}_p = \frac{N_p^{pc}}{N_p}$, where N_p is the number of times x_p was occluded while shifting the window and N_p^{pc} is the number of times the prediction $\hat{y}_k^o$ changed from prediction $\hat{y}$ when occluding $x_p \in w_k$. $PI_p \in [0, 1]$, higher values indicate higher importance.

Occluding only one channel might not change the prediction, even though the pattern is important for a sleep stage. Hence, we use time-domain occlusion on both EEG channels in the SHHS dataset.

5.3. Predictive frequency bands in EEG

According to the AASM guidelines, some frequency bands of EEG signals are highly indicative for certain stages (cf. Table 1). To investigate if our model corresponds to this domain knowledge, we occlude the EEG signal in the frequency-domain *(frequency-domain occlusion)* [31] and investigate whether the prediction changes on occlusion. Removing the most prominent frequency bands should result in maximum misclassification for that stage and keeping only those frequency bands will lead to maximum misclassification for other stages. More concretely, we occluded δ (delta), θ (theta), α (alpha), σ (sigma) and β (beta) (cf. Table 1) frequency bands one-at-a-time. To occlude a frequency band, we (i) kept frequencies only in the specific range and removed the rest (band-pass filter in the frequency band), or (ii) removed all frequencies in a specific range (band-stop filter in the frequency band) from the signal. We performed our experiments on both EEG channels in SHHS.

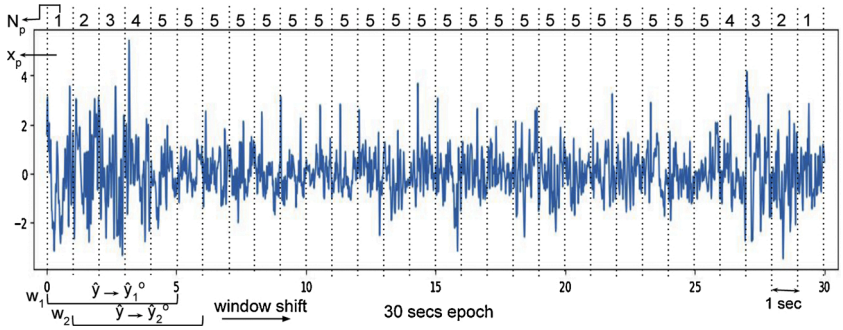


Fig. 3. Time-domain occlusion on EEG epoch with predicted class $\hat{y}$. Window w_1's prediction changed to $\hat{y}_1^o$ on occlusion.

5.4. Filter importance and visualization

We would like to know how different temporal filters contribute to the final decision and whether specific filters are learnt for certain sleep stages. To this end, we applied LRP [35], consisting of a forward activation and a specific backward calculation, to calculate an importance score of a filter per test instance. We focused on the filters in the first convolutional layer in temporal filtering (i.e., the 5th layer in the model), because feature complexity will increase in deeper layers [42] and patterns in the AASM guidelines are also basic features. Filter relevancy is averaged per sleep stage, resulting in an importance score per filter per stage. We selected 20 patients randomly from SHHS[4] to calculate the importance scores. To identify the frequency patterns learned by a specific filter, we performed power spectrum analysis on the activations of a filter. The dominant frequency components of both – the raw data and the activations of a filter, are compared. We also generated white noise with a uniform distribution of all effective frequencies of input signals to test the filter reactions to all frequency bands. Because EEG channels contain important frequency information, while EOGs and EMGs are mainly distinguished by amplitude (cf. Table 1), we analysed the 8 EEG filters on SHHS.

6. Experimental setup

In this section, we describe the model variants, the training process and evaluation metrics.

6.1. STQS architectures and baseline

We tested various combinations of the architectural components outlined in Section 4.2. The **ST** model performs spatio-temporal filtering and combines the modalities for the final prediction as described in Section 4.2.3 and 4.2.6. In Fig. 1, **ST** corresponds to the top part (input, spatial, and temporal filtering, modality concatenation), the dashed line ① and the prediction layer. **Q** denotes the sequential learning component described in Section 4.2.4 corresponding to dashed line ②. **RC** denotes the residual connection block described in Section 4.2.5, corresponding to the dashed line ③ in Fig. 1. The three different imbalance techniques introduced in Section 4.3 are denoted with superscript (oversampling ^{O}, and ^{W1} and ^{W2} are the two weighted cost functions). We further compared our STQS models to a baseline model, MLP (Multi-Layer Perceptron), made up of Input-3 FC blocks-Output (18,750, 10,000, 5000, 1000, 5 neurons respectively). Input is a concatenation of features of all channels, each FC block consists of a FC layer followed by BatchNorm and ReLU and the output layer consists of a FC layer with a softmax activation function. We evaluate the MLP only on SHHS due to the huge size of the input layer in the MST dataset.[5]

We tested the following model variants ST, ST^{W1}, ST^{W2}, ST^{O}, ST-Q, ST-Q^{W1}, ST-Q^{W2}, ST-Q^{O}, and ST-Q-RC^{O} using the same **hyperparameters** for all. We trained with the Adam optimizer [43] with learning rate $\lambda = 10^{-4}$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$.[6] We did not use weight decay for regularization, as it resulted in worse performance than the class imbalance techniques. Weights for batch normalization, convolutional and fully connected layer were initialized following a normal distribution $\mathcal{N}(0, 0.02)$.[7] For the Bi-LSTM layer, we used orthogonal weight initialization [44] and a sequence length of 8 epochs (4 min) as suggested by previous work [9].[8] Our batch size of 192 epochs corresponds to 96 min, i.e., approx. one sleep cycle. Thus, on average each batch contains most sleep stages. For the ST-Q variants, a batch (of size 192 epochs) is reshaped into 24×8 before passing it to Bi-LSTM layer. This means the Bi-LSTM layer is trained with a batch size of 24 sequences of length 8. Before passing the Bi-LSTM output to the prediction layer, it is reshaped back to the batch size of 192 to consider all the epochs in the sequence for prediction. The last input epoch of each PSG was filled to sequence length 8 by copying the epochs from the beginning of that PSG. Thus, both the first and the last sequence from a PSG contains the first few epochs. For evaluation, we only consider the Bi-LSTM output from the first sequence for these common epochs. We applied the **interpretability** techniques on the ST^{O} model, since ST-Q^{O} learns sequential information from the neighbouring stages for prediction, and we cannot explain the predictions of this model solely based on stage-specific rules.

6.2. Training and testing

We train the spatio-temporal filters and the sequential filters successively. In *Stage 1* we train the spatio-temporal filters. We shuffle the input data, and if oversampling is used, we additionally augment the data accordingly. Then, we train only the ST part of the architecture. In *Stage 2* we train the sequential parts of the model. Each input PSG is divided into 8 non-overlapping sequences of 30 s epochs (cf. Fig. 2). Models are initialized with the weights learned in Stage 1 (except the prediction layer) and all layers are trained. No class imbalance technique is used in this training step.

We used early stopping on validation loss with a patience of 7, i.e., if the validation loss did not decrease for 7 training iterations, the training was stopped. The PyTorch implementation and trained models are available online.[9]

6.3. Evaluation metrics

We calculated the overall accuracy (a) at 95% confidence interval (CI), the balanced accuracy (a_b) to account for class imbalance, the macro-averaged F1 score ($F1^M$) and Cohen's kappa (κ). All values are reported in percentage for better readability in Section 7.

$$a = \frac{\sum_{s=1}^{|S|} \mathrm{TP}_s}{N}, \quad \mathrm{CI} = 1.96\sqrt{\frac{a(1-a)}{N}}$$

$$a_b = \frac{1}{|S|}\sum_{s=1}^{|S|}\frac{\mathrm{TP}_s}{N_s}, \quad F1^M = \frac{\sum_{s=1}^{|S|} F1_s}{|S|}$$

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \quad \text{with } p_e = \sum_{s=1}^{|S|}\frac{N_s^{\mathrm{tr}}}{N}\cdot\frac{N_s^{\mathrm{pr}}}{N}$$

where TP_s, $F1_s$ and N_s are the number of true positives, F1 score and number of epochs of class s respectively. $|S|$ denotes the number of classes and N is the total number of epochs in the test set. p_o is the relative agreement among ground truth and prediction, p_e is the hypothetical probability of a chance agreement, N_s^{tr} is the number of epochs in the true class s and N_s^{pr} is the number of epochs in the predicted class s. We also report predicted hypnograms for two test PSGs, with highest and lowest distance (among the test set) from their ground truth, calculated using hypnogram distance, *HD*. *HD* is based on the assumption that stages further away during sleep, should have more influence on the distance. Therefore, we encode the sleep stages as numbers (W as 0, N1 as 1, N2 as 2, N3 as 3, and REM as 4). The difference between N1-N2 (or N2-N1) and W-REM (or REM-W) is then 1 unit and 4 units respectively.

[4] We only used 20 patients due to limited RAM.
[5] 12 channels × 3750 features.
[6] $\lambda = 10^{-3}$ decreased performance.
[7] $\mathcal{N}(0, 1)$ took longer to converge.
[8] [9] did not test sequence length <8. We tested with a value of 5, i.e., 1 min before and after the to-be-classified epoch, as suggested by sleep technologists, but this did not improve performance.
[9] Code is available on Github: https://github.com/ShreyasiPathak/STQS.

$$\mathrm{HD}_{\mathrm{PSG}} = \frac{\sum_{n=1}^{N_{\mathrm{PSG}}} |y_n - \widehat{y_n}|}{N_{\mathrm{PSG}}} \in [0, 4]$$

$$\mathrm{sim}_{\mathrm{PSG}} = 1 - \frac{\mathrm{HD}}{4} \in [0, 1],$$

where $y_n, \widehat{y_n}$ are the true and predicted sleep stage of an epoch n. N_{PSG} denotes the total number of epochs in a PSG. The factor $\frac{1}{4}$ normalises similarity (sim_{PSG}) to the interval [0, 1].

7. Sleep stage prediction results

In this section, we first report predictive performance of various architectural choices. Then we show the importance of modalities for sleep stage prediction and its conformance to AASM guidelines.

7.1. Model performance

An *overall comparison of model variants* (cf. Section 6) on both data sets (cf. Section 3) can be found in Table 4. We report accuracy values at 95% CI (cf. Appendix Table 3 & 4 for results table with all confidence interval values). ST-Q^O is overall the best performing model with a and a_b of 85.0% and 75.9% respectively on SHHS and a and a_b of 77.2% and 72.7% on MST. It only shows slightly worse results than ST-Q-RC^O for a_b on SHHS (75.9% vs. 76.0%). All our STQS models outperform our baseline model, MLP, showing the importance of spatio-temporal feature learning by ST over simple fully connected feature learning. Overall, all sequential models (denoted with -Q-) outperform Biswal et al. [20] in terms of a and κ, while the amount of improvement depends on the class balancing technique. However, the overall performance of our best model is ≈2% lower than Sors et al. [18], except for F1 score of W and REM stage, which is 1.5% and 3.7% higher respectively. This may be due to the fact that Sors et al. is a single channel model whereas STQS is a multi-modal, multi-channel model. Therefore, we investigate the importance of various modalities for sleep stage prediction in Section 7.2.

Comparing the ST model with its *class imbalance* counterparts ST^O, ST^{W1} and ST^{W2}, we see that class imbalance techniques improve balanced accuracy, but not necessarily overall accuracy. This shows that class imbalance techniques improve the performance on rare classes, but possibly at the expense of misclassifying more epochs in total. There is, however, no clear picture for the choice of *class imbalance* technique of non-sequential models (ST^O, ST^{W1}, ST^{W2}). Taking balanced accuracy for comparison, oversampling outperforms weighted cost functions on SHHS, whereas the opposite is true for MST. The sequential model ST-Q (without class imbalance) performs worse (1.5% decrease in accuracy on SHHS) over its class imbalance counterparts (ST-Q^O, ST-Q^{W1}, ST-Q^{W2}), among which oversampling shows better results than weighted cost functions. ST-Q-RC^O and ST-Q^O show similar performance: all performance metrics are equal or differ by a maximum of 1%, indicating that combining information with *residual connections*, while adding more trainable parameters, does not improve the overall performance. In summary, we observe performance improvements by adding sequential information and accounting for class imbalance. Adding residual connections to a spatio-temporal model does not show any further improvement.

The *F1 scores of all stages* for both datasets as predicted by ST-Q^O are above or near to 0.80 except for N1 and N3. These two classes are also among the classes with least number of instances. To report and compare the major class misclassifications as predicted by our model variants, we show the row-wise normalized confusion matrices of ST^O, ST-Q^O and ST-Q-RC^O for SHHS dataset in Fig. 5. On comparing the confusion matrices, we see that all the 3 models make misclassification between N1-N2 and N2-N3 and ST^O (Fig. 5a) additionally misclassifies between REM-N1 and REM-N2. ST^O has the highest true positive rate (TPR) among all confusion matrices for N1 and N3, ST-Q^O (Fig. 5b) has

Table 4
Performance comparison of different STQS model variants. Reporting aggregated values and F1 scores per class. Best values among our model are marked in bold and best values when compared with related work are marked in bold and italics,"–" indicates that the values are not available in the respective publication. We show the model of Sors et al. [18] for completeness, however, the results are not directly comparable ([18] excludes subjects from the dataset (cf. Section 9.5).

Models	SHHS				SHHS					MST				MST				
	a	a_b	$F1^M$	κ	W	N1	N2	N3	REM	a	a_b	$F1^M$	κ	W	N1	N2	N3	REM
MLP	69.8 ± 0.11	55.9	55.9	57.2	75.8	5.2	72.5	65.8	60.1	–	–	–	–	–	–	–	–	–
ST	75.4 ± 0.11	63.3	61.9	65.7	83.2	8.6	75.3	71.5	70.7	75.1 ± 0.22	68.9	70.2	65.2	81.3	41.3	79.4	73.4	75.6
ST^{W1}	75.5 ± 0.11	61.9	61.2	65.6	84.3	9.5	76.9	72.6	62.4	73.7 ± 0.23	71.6	71.1	64.6	81.2	46.8	77.2	75.0	75.2
ST^{W2}	71.4 ± 0.12	70.8	64.5	62.3	87.1	29.4	67.5	67.2	71.2	72.8 ± 0.23	71.6	70.5	63.7	80.4	47.1	76.2	75.1	73.5
ST^O	73.6 ± 0.11	72.1	66.4	64.8	90.2	32.4	72.5	67.7	69.1	73.5 ± 0.23	70.7	70.2	64.2	80.0	45.0	77.4	74.3	74.3
ST-Q	83.5 ± 0.10	74.5	74.6	77.1	90.7	36.6	83.1	74.8	87.9	**77.5 ± 0.21**	72.6	73.6	**68.9**	**83.1**	48.5	**81.1**	75.7	79.5
ST-Q^{W1}	80.3 ± 0.10	69.2	68.8	72.4	90.4	21.8	79.3	71.5	81.0	76.5 ± 0.22	72.5	72.8	67.8	81.0	47.9	80.2	**75.8**	79.0
ST-Q^{W2}	83.0 ± 0.10	71.9	72.8	76.0	90.4	30.5	82.9	74.8	85.3	77.0 ± 0.22	**72.7**	73.5	68.3	82.7	49.6	80.5	75.2	79.5
ST-Q^O	**85.0±0.09**	75.9	**76.6**	**79.0**	92.1	**41.3**	**84.8**	**76.3**	88.7	77.2±0.22	**72.7**	**73.8**	68.5	82.7	**50.9**	80.7	75.0	**79.7**
ST-Q-RC^O	84.9 ± 0.09	**76.0**	76.5	**79.0**	**92.5**	40.3	84.4	76.0	**89.1**	77.1 ± 0.22	72.5	73.5	68.4	82.5	49.8	80.8	74.9	**79.7**
Biswal et al. [20]	77.9	–	–	73.4	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Sors et al. [18]	***87.0***	–	***78.0***	***81.0***	91.0	***42.7***	***87.9***	***85.0***	85.4	–	–	–	–	–	–	–	–	–

the highest TPR for N2 and ST-Q-RC^O (Fig. 5c) has the highest TPR for REM. Though ST^O has higher TPR for N1 and N3, the false positives for the same classes are higher as well. N1 and N3 being predicted more often suggests to be an effect of oversampling the dataset in order to account for rare classes.

The results reported so far show the performance per epoch. In order to investigate the *misclassifications on the PSG level*, we identified the two PSGs in SHHS for which the predictions are most similar and least similar to the ground truth (cf. Section 6.3). Fig. 4 shows the true (top) and predicted hypnogram (bottom) of those PSGs. It can be seen that the most similar hypnogram has many W stages, leading to the high similarity score. The N2-REM, W-N2, N1-N2 transitions have been predicted correctly, whereas the predicted hypnogram could not identify all the N2-N3 transitions. The most dissimilar hypnogram has more stage transitions presumably making it harder to predict. Not all W-N2, N2-N3 and W-REM transitions could be correctly predicted.

7.2. Modality importance

We investigated the contribution of each modality using the confusion matrices of the model generated on occluding combinations of modalities (cf. Fig. 6) in SHHS dataset. For instance, the contribution of EEG can be inferred by either **occluding EEG or keeping only EEG** (while occluding EOG and EMG). EEG is very important for the model and results in high TPR for all stages except N1, which is misclassified as REM. **EOG alone** cannot classify stages correctly. However, **removing EOG** lowers the TPR of all stages, except for REM. In fact, without EOG, REM achieves the best TPR, along with other stages also getting misclassified with REM, especially N1. Moreover, we find that **occluding EMG** results in lowest TPR for REM, while TPR of other stages remains almost unchanged. Further, **EMG alone** can only identify REM and W. If **all channels are occluded**, the model predicts W, which is the non-sleep stage and the class with the highest number of training instances.

In conclusion, the results show that (i) EMG and EEG are sufficient to correctly predict REM, (ii) EEG and EOG are sufficient to correctly classify N1 and N3, (iii) EEG alone is sufficient to classify W and N2, and (iv) EOG is necessary to reduce the misclassification of other stages with REM. These observations can be justified by the following facts in AASM (cf. Table 1): (i) The lowest amplitude of EMG in REM may make EMG important to identify REM, (ii) W and N2 generally have a very characteristic EEG signal – high frequency signal in W and unique time-domain patterns in N2, making these 2 stages easily identifiable using EEG alone and (iii) The misclassification of other stages with REM, most prominently N1 (EEG signal of N1 and REM are quite similar leading to such high misclassification), is reduced by adding EOG due to the rapid eye movement patterns in REM.

8. Model interpretability results

In this section, we analyse important patterns in the time and frequency domain of EEG signals for prediction and investigate activation patterns of the temporal filters in the CNN.

8.1. Predictive patterns in EEG

To analyse which EEG patterns are most relevant for the prediction, we occluded the EEG signal in the time domain. Four examples of correctly classified epochs from the SHHS test set are shown in Fig. 7, lighter colors denote more important patterns. This means, if a yellow 1 s pattern is occluded by the 5 s occlusion window, the prediction changed more often than for patterns in darker colours. Purple means no change of prediction on occlusion. The annotations at the top of each figure indicate the following: The bottom number shows the PI_p of a pattern p and the annotation above it shows the corresponding prediction on occlusion, $\hat{y}_p^o$. The example of **stage N1** shows multiple consecutive 1 s patterns with vertex waves as most important for the prediction. The example of **stage N2** highlights the k-complex in green, showing that it is among the important patterns for prediction, but not the most important. This suggests that the epoch contains other important information which leads to predicting N2, even if the k-complex is occluded. For **stage N3** high amplitude patterns are important, while the prediction for **stage REM** is based on saw-tooth waves and high-amplitude patterns. These findings conform to the characteristic patterns according to AASM guidelines (cf. Table 1).

While the previous examples show the importance of patterns for example epochs, we also investigated the change of prediction on an aggregated level. The Sankey diagram in Fig. 8 shows how many epochs per stage change prediction on occlusion. The figure is based on 10 randomly selected PSGs. We show the number of epochs in the ground truth s_T, the prediction s_P, and the prediction on occlusion s_{PO}. For instance, 94,380 epochs in the ground truth are N2. From those, 77,142 are predicted as N2. On occlusion, 94,797 epochs are classified as N2. Nearly all epochs predicted as N2 are also predicted as N2 on occlusion. This means, N2 mostly does not contain 5 s patterns that solely identify N2. If N3 epochs are occluded, a considerable amount is misclassified as N2, indicating 5 s windows in the epochs that hold the information for distinguishing N2 from N3. Further, many true N2, which are misclassified as N3, are correctly predicted as N2 on occlusion. This shows some similar patterns between N2 and N3 which results in these misclassifications and on occluding these patterns, the epoch is correctly predicted.

8.2. Predictive frequency bands in EEG

To identify the most relevant EEG frequency bands for the

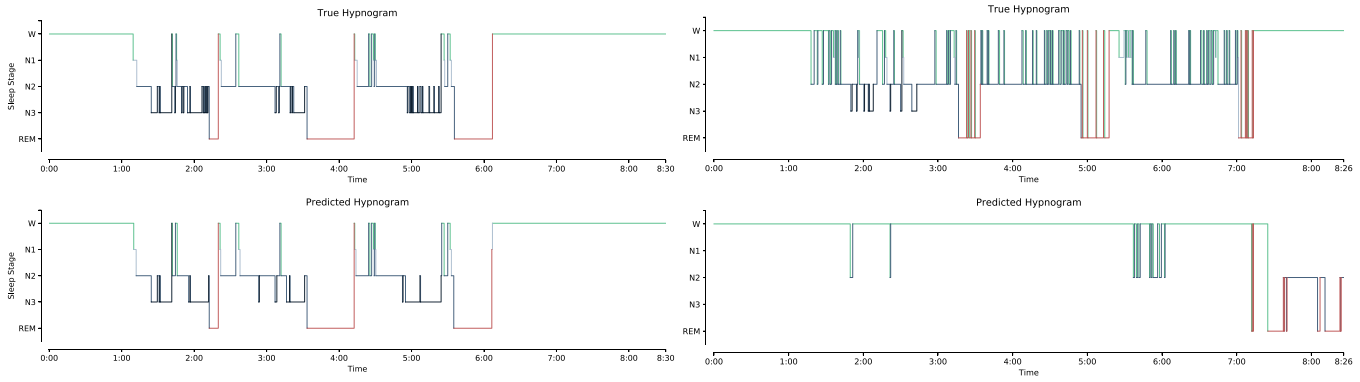
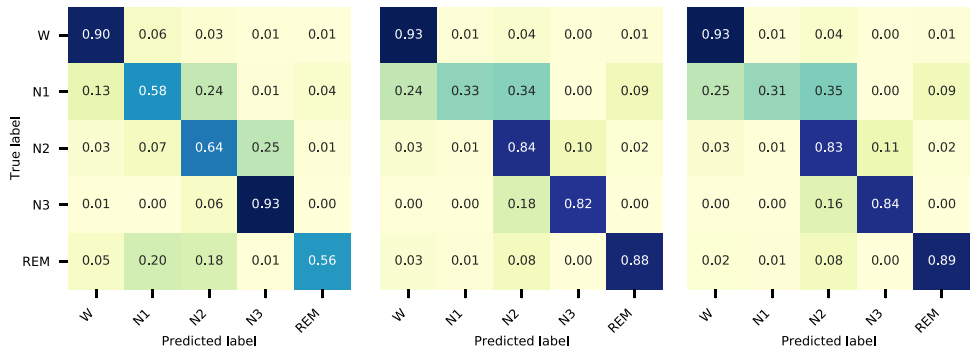


Fig. 4. Example hypnograms from SHHS: Ground truth (top), prediction of ST-Q^O (bottom). Hypnograms with highest similarity ($HD = 0.06$, $sim = 0.99$; left) and lowest similarity ($HD = 1.58$, $sim = 0.61$; right) between prediction and truth.



(a) SHHS ST^O (b) SHHS ST-Q^O (c) SHHS ST-Q-RC^O

Fig. 5. Row-wise normalized confusion matrices for our model variants on SHHS.

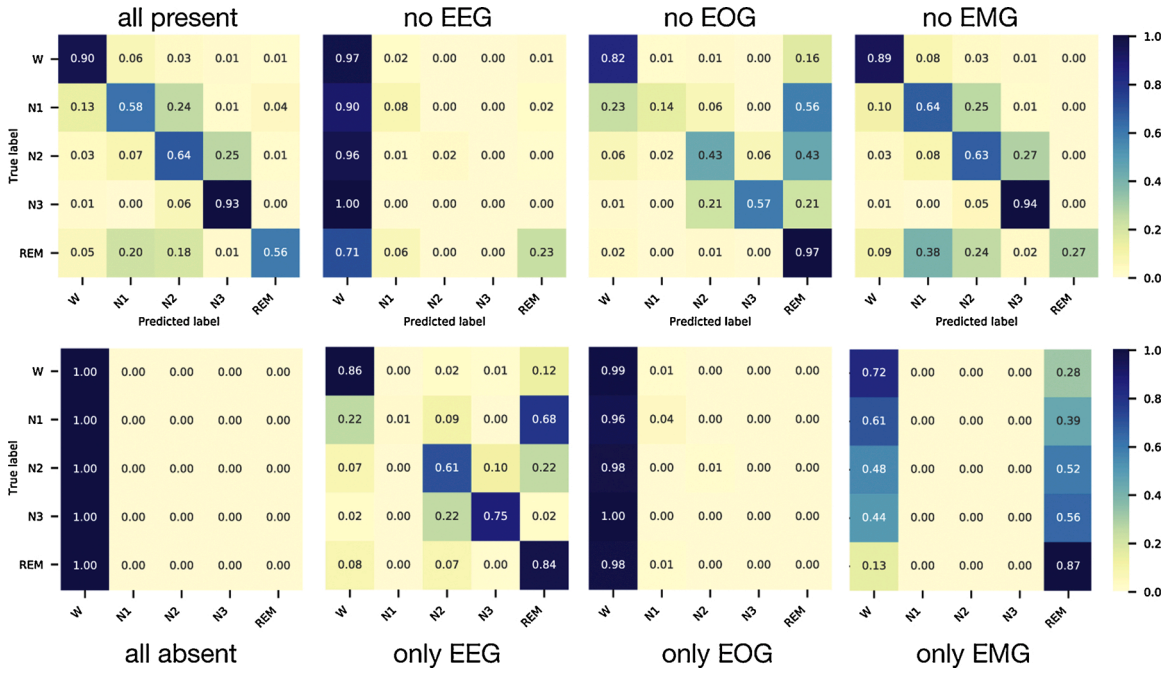


Fig. 6. Modality occlusion for ST^O model on SHHS. Row-wise normalized confusion matrices if one modality is removed (top row) or only one modality is kept (bottom row).

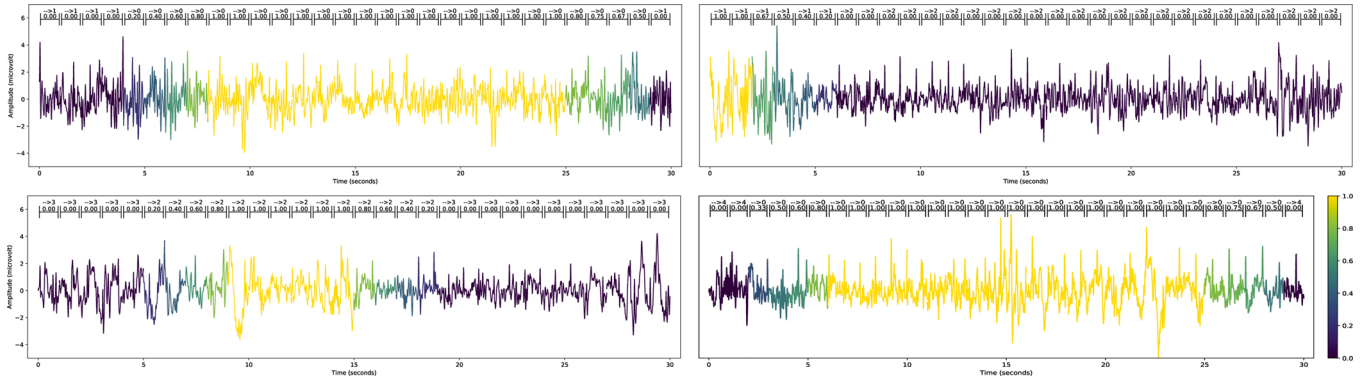


Fig. 7. Time-domain occlusion for EEG (C3-A2) on SHHS. Top left: N1, top right: N2, bottom left: N3, bottom right: REM. The annotation at the top of each plot shows $\hat{y}^o$ represented by numbers (W: 0, N1: 1, N2: 2, N3: 3, REM: 4). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

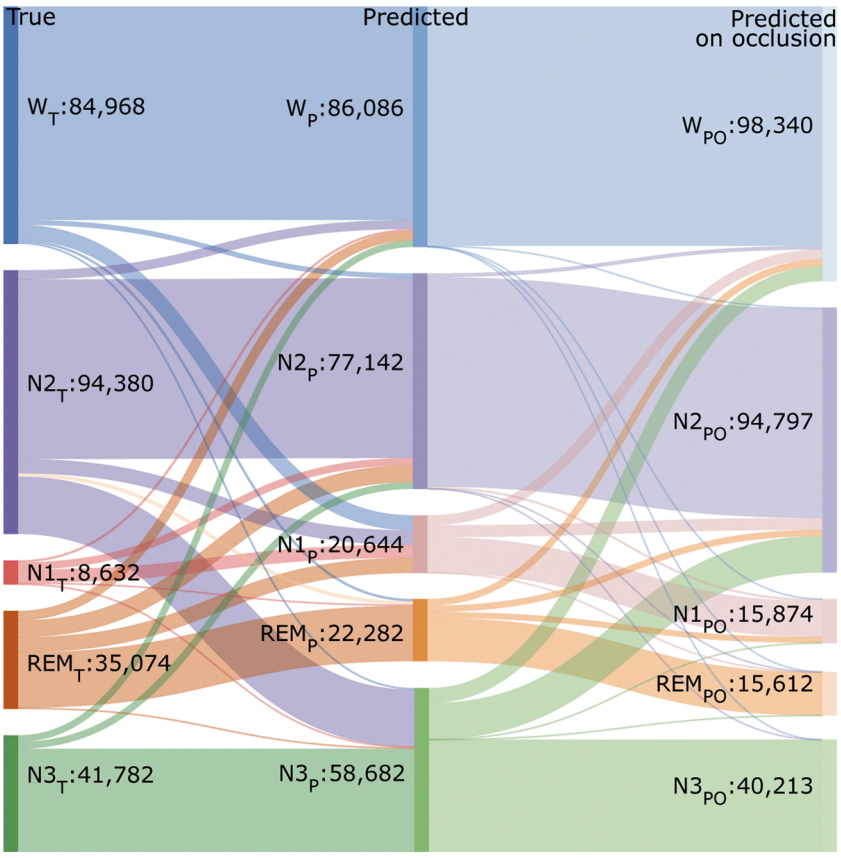


Fig. 8. Sleep stage changes visualized as flow for SHHS. Showing the number of epochs in the ground truth s_T (left), predictions of ST^O, s_P (middle) and prediction on time-domain occlusion s_{PO} (right) for each stage s.

prediction, we occluded the EEG signal in the frequency domain. Fig. 9 shows the confusion matrix on occluding different frequency bands. With only $\boldsymbol{\delta}$ **frequencies** present, most epochs are (correctly and incorrectly) predicted as N3, while in absence of those frequencies N3 is nearly never predicted. This shows that δ is the characteristic frequency band of N3, which also conforms with the characteristic frequency of N3 according to AASM (cf. Table 1). If only $\boldsymbol{\theta}$ **frequencies** are present, mostly N1 and W are (correctly and incorrectly) predicted, while omitting θ, decreases the TPR of N1, but increases the TPR of W. This shows that θ is a characteristic frequency of N1 (conforms with the AASM guidelines). N2 and REM also show considerable TPR when only θ is present, however, on removing θ, they still show a considerable TPR. This shows that θ is one of the characteristic frequency bands of N2 and REM, also in accordance with the AASM guidelines (cf. Table 1). If only $\boldsymbol{\alpha}$ **frequencies** are present, stages are classified as W, whereas its absence results in increase of TPR in all other stages. This shows that α is the characteristic frequency band for W. A similar argument holds for $\boldsymbol{\beta}$ **frequencies**, suggesting that β is a characteristic frequency band for W,

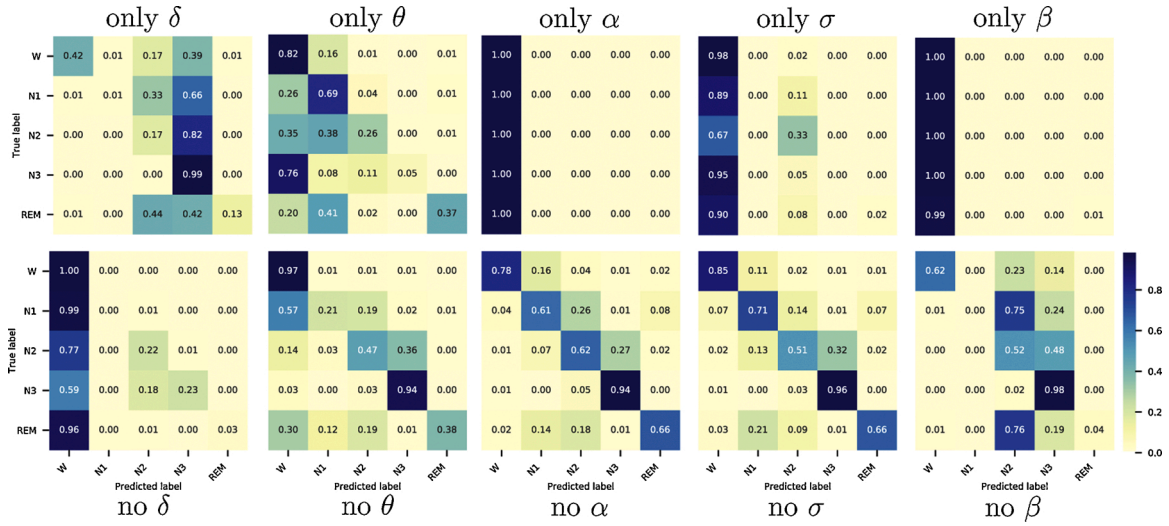


Fig. 9. Frequency-domain occlusion for ST^O on SHHS. Row-wise normalized confusion matrix with only specific frequency bands (top) or on omitting specific bands (bottom).

which aligns with the AASM guidelines (cf. Table 1). If only **σ frequencies** are kept, all stages are classified as W; except N2. The absence of σ frequencies results in higher TPR for all stages, but W. A considerable number of TP in N2 for the case when only σ is present, is attributable to the sleep spindles in N2, whose frequency lies in the σ band. This shows σ is the characteristic band of W and also present in N2, again confirming the domain knowledge from the AASM guidelines (cf. Table 1).

8.3. Filter importance and visualization

The LRP [35] importance scores for the 8 CNN's temporal filtering layers are shown in Fig. 10. Most filters have relatively small absolute importance scores, i.e., single patterns are not very discriminative for a sleep stage. Single filters with high importance are found for W, N2 and N3. Few filters react significantly to N1 and REM, illustrating the hardness of predicting these stages. For identifying and visualising the frequency patterns learnt by EEG filters, we selected the filters with high absolute importance scores and analysed the dominant frequency components of their activations. Fig. 11 shows Power spectral densities of the raw signal and filter activations from typical epochs of filters 1, 2, 7 for W, filter 2, 3, 4 for N2 and filter 2, 3, 6 for REM. Filter reactions to the white noise of 0.5–30 Hz (effective frequency components of EEG inputs) are plotted in Fig. 12 to visualise the frequency patterns learnt by all 8 EEG filters. If we compare the filter reactions of the white noise to corresponding filter reactions in Fig. 11, we can see that the same filter always extracts the same frequency patterns independent of the input, which verifies that every EEG filter has its invariant unique function in feature extraction. Moreover, given all frequencies in the white noise have the same amplitude, the actual contributing value of a filter to a frequency band can be conferred via the corresponding amplitude in Fig. 12.

In addition, frequency patterns learnt by the EEG filters can be specifically explained when compared to the AASM guidelines and the importance scores. Fig. 11a shows that filter 1 and 7 extract the frequency components around 13 Hz, 0–2 Hz and 8–10 Hz from the EEGs of W, and filter 2 extracts 0–6 Hz. In the AASM guidelines, EEGs of W mainly contain α and σ frequencies (8–16 Hz). Therefore, filters 1 and 7 react positively, and filter 2 reacts negatively when predicting W. Similar observations can be made for Non-Wake stages. For N2 (cf. Fig. 11b), filters 2 and 3 react positively as they mainly recognize the frequencies between 0–14 Hz (θ frequencies, k-complex and sleep spindles), and filter 4 reacts negatively as it detects the frequencies between 15–25 Hz. For REM (cf. Fig. 11c), filters 2 and 3 extract the frequency components in 0–12 Hz and filter 6 mainly extracts 20–30 Hz. Filters 2 and 3 are highly important to REM, as the main components of REM are θ and α frequencies (4–12 Hz). Additionally, if we compare the importance scores of the same filter in the prediction of different stages, the quantitative importance scores can exactly show the contribution of a filter in predicting a particular sleep stage. For example, the frequency pattern learnt by filter 2 is 0–6 Hz which matches N2 better than REM, so filter 2 has a higher importance score in N2 than REM.

9. Discussion

In this section, we discuss the contribution of the architectural components, reasons for misclassifications and put our work in context to related work.

9.1. Model analysis

We developed a multi-modal sleep scoring model which can learn from EEG, EOG and EMG. This was motivated from the fact that the AASM manual recommends the use of all three modalities for sleep scoring, as the three modalities together have unique characteristics to differentiate various stages. Sleep technologists also usually consider all 3 modalities for scoring. We show that automatic sleep scoring on raw multi-modal input signals can be performed with 85% accuracy.

The confusion matrix from the ST^O model (cf. Fig. 5a) shows a higher TPR for N1 and N3 than ST-Q^O (cf. Fig. 5b). Adding residual connections to combine spatio-temporal with sequential features (ST-Q-RC^O model), however, did not increase the overall performance, but increased the TPR of N3 (cf. Fig. 5c). This indicates that N3 can be predicted quite well solely based on temporal features, whereas sequential information seems to decrease performance of N3. This observation is in line with the AASM manual, which does not mention sequential rules for N3. The confusion matrix of ST-Q^O (cf. Fig. 5b) shows that misclassifications generally occur between contiguous sleep stages (W-N1, N1-N2 and N2-N3). The confusion matrix of ST^O (cf. Fig. 5a) also shows misclassifications based on feature similarities, e.g., REM-N1 and REM-N2 misclassifications. REM and N1 lie in similar frequency bands for EEG and EOG signals: EEG lies in θ, α (cf. Table 1) and EOG lies in 0.1–0.4 Hz [45]. Similarly, REM and N2 have EEG frequencies in θ and the amplitude of EMG in N2 can be as low as the amplitude in REM (cf. Table 1). The AASM manual also indicates that a stage is scored as REM when the majority of the epoch has REM characteristics, even though it contains a k-complex suggesting it to be N2, which can make REM and N2 hard to distinguish.

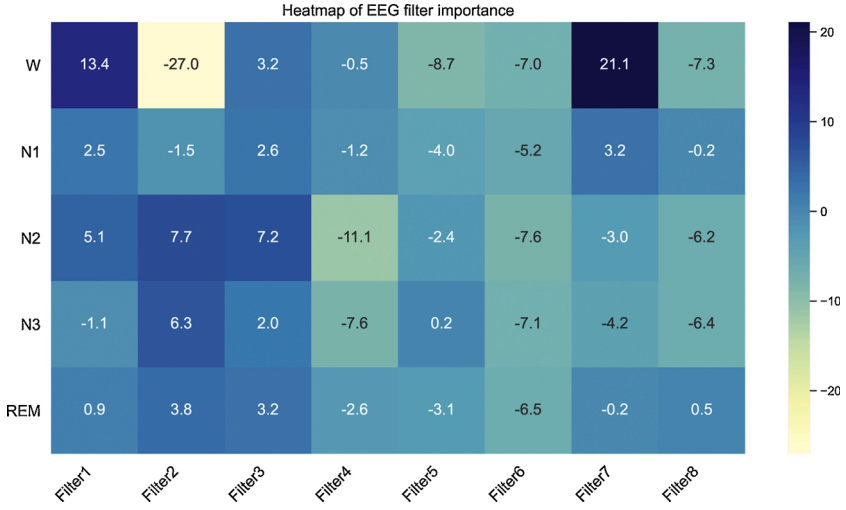
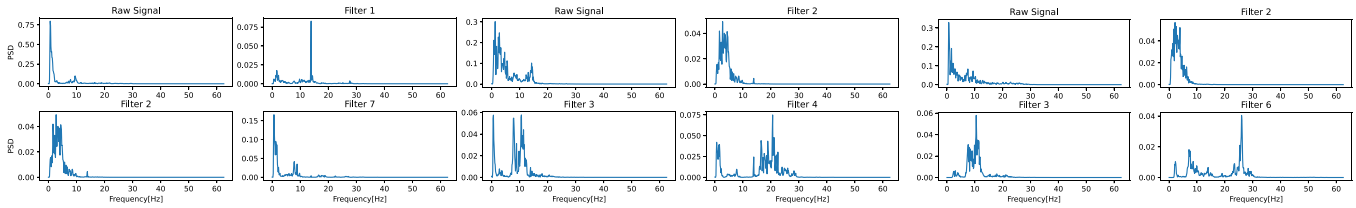


Fig. 10. Filter importance of the 8 EEG filters at the first temporal filtering layer (layer 5 in Fig. 1) for SHHS.



(a) W (b) N2 (c) REM

Fig. 11. Frequency spectrum of the raw data and corresponding EEG activation patterns for W, N2 and REM for SHHS.

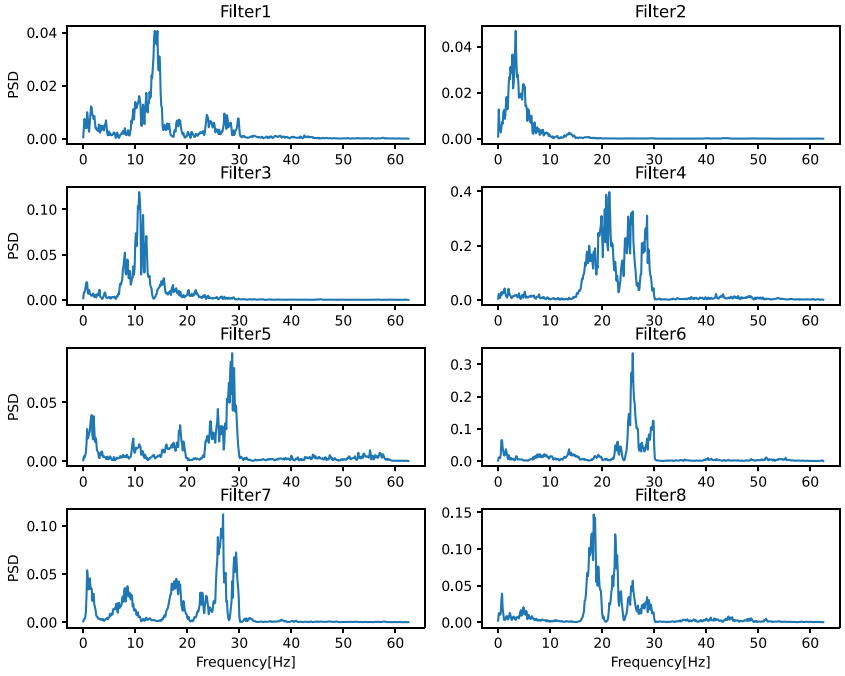


Fig. 12. Frequency spectrum of the activations of 8 temporal EEG filters to white noise for SHHS.

9.2. Additionally tested experimental setup

We performed some additional experiments to decide on the design of our Bi-LSTM component, STQS architecture and training process.

Bi-LSTM Component: We compared initialization of hidden and cell state of a sequence with information from the previous sequence to initialization of the same with zeros. We found better performance for the former and therefore only report those results. More specifically, the performance of REM and N2 increased, as for those stages, more information from contiguous epochs is necessary to decide the sleep stage of the current epoch [2] and some of these contiguous epochs can be present in neighbouring sequences. We also experimented with passing overlapping sequences to the Bi-LSTM such that the sequence window would be shifted by 1 input epoch for a new sequence, instead of shifting by the sequence length. The output taken into consideration for each sequence was the hidden state of the middle 30 s epoch in the sequence. The motivation for overlapping sequences was to provide more context for learning i.e., providing an equal amount of past and future epochs for the prediction of an input epoch. This resulted in more training time and similar results when compared to non-overlapping sequences. The similarity in results was mainly due to the fact that in non-overlapping sequences, hidden and cell state value of a sequence are passed on to the next sequence, generating a similar effect as overlapping sequences.

STQS Architecture: We compared the addition of (i) a common Bi-LSTM after concatenating the CNN features from all modalities against (ii) separate Bi-LSTM layers after the CNN layer in each modality pipeline, concatenating these Bi-LSTM feature maps from all modalities and then passing it for prediction. Both versions resulted in similar overall accuracy (85%), while the separate Bi-LSTM performed slightly better in predicting N1 (44.1% vs. 41.3% F1 on SHHS), but also required more training time due to an increased number of parameters. We therefore chose the model with a common Bi-LSTM layer.

Training Process: We explored 3 different 2-step training processes – (i) training the **ST** architecture, then retraining the **ST-Q** architecture with the same parameters, (ii) using different learning rates for both training steps [11], and (iii) training the **ST** architecture, then freezing the ST weights and using ST to calculate the feature maps and training only the **Q**. Process (i) seemed to result in the best performance and so, we used that for training.

9.3. Effect of dataset and class imbalance techniques

Our model was trained and tested on 2 different datasets – SHHS and MST. Our model performed better on SHHS (a : 0.85) as compared to MST (a : 0.77), which might be due to higher number of training data in SHHS. However, we think that though all modalities are important for sleep scoring, the amount of channels in each modality may contribute negatively to the performance. MST dataset has 12 channels and SHHS has 5. Having to combine information from so many channels in each modality pipeline may confuse the model (information within the channels of a modality may not support each other) rather than supporting it with additional information. More evidence on this can be

found in Section 9.5.

We experimented with 3 class imbalance techniques and found class imbalance handling to be helpful in general, however, there seems to be an interaction with the dataset. Oversampling was found to be the best for SHHS whereas a weighted cost function was best for MST. We think this may be related to the amount of channels in the dataset. As the MST dataset contains a lot of channels for each modality, duplicating data may have added to the confusion and made it hard for the model to learn from this data. On the other hand, the weighted cost function increased the error for rare classes making the model update its weight more for rare classes. We used 2 weighted cost function techniques and found contradicting observation for both dataset, so we cannot conclude which one is the best weighted cost function.

9.4. Post-hoc interpretability

Our post-hoc interpretability results showed that our model's prediction conforms to the AASM guidelines by giving importance to the unique characteristics mentioned for each stage.

However, we found that the time-domain occlusion is not a reliable interpretability method as for some epochs, it did not find any important pattern. This suggests the absence of 5 s patterns, which are solely responsible for the prediction. This can be explained from the fact that sleep stages can either contain multiple non-localized patterns throughout the epoch like vertex or saw-tooth waves in N1 and REM or single localized pattern like k-complex in N2. In the former case, the model can either find all the multiple patterns to be important (Fig. 7 (N1)) or no particular pattern to be specifically important due to multiple occurrences of the pattern. Further, localized patterns are not always the only reason for the decision (Fig. 7(N2)), again suggesting the possibility of no particular pattern being solely important.

We experimented with 5 s, 10 s and 15 s occlusion window sizes for time-domain occlusion. Results indicated that there is a trade-off between finding the most important patterns and occluding enough information for the prediction to change. We chose 5 s, as the larger window sizes found more patterns to be important, losing on our objective to find only the most important patterns. Limitations of our method are, for example, how much time-domain, frequency-domain or modality information our model uses for prediction. Also, our modality occlusion does not calculate the importance of the channels in each modality. We have developed methods for interpreting our ST models, but extensions for the sequential part of the architecture are left for future work.

9.5. Comparison with state-of-the-art

The performance of ST-Q^O is on par with the state-of-the-art **multi-modal deep learning approaches** on sleep scoring (cf. Table 2). Comparing to other papers is not straight-forward, due to differences in datasets and its preprocessing. We compare to other works that are closely related to our model and use similar data. Our CNN architecture was motivated by [12] and therefore we did not experiment with various filter sizes or other parameters of our CNN model.

Comparing to models evaluated on the same dataset (SHHS), we found that we outperform **single modal model**, Biswal et al. [20] and are at-par with **single modal model**, Fernandez-Blanco et al. [15] (however, their dataset has more PSGs than ours (cf. Table 2)), but we perform 2% lower than **single channel model**, Sors et al. [18]. Due to different data preprocessing, dataset splits and class imbalance techniques, results are not directly comparable. Our ST-Q model without any class imbalance technique has 1.5% lower accuracy than ST-Q^O on SHHS and can be used to compare to state-of-the-art models, which do not use any class imbalance technique, like [18]. Still, the models are not directly comparable because of the other variations. To make our model directly comparable to [18],[10] we trained and tested their model on our SHHS dataset split (81-9-10%, train-validation-test), including all PSGs and wake stages (unlike [18]) and did not filter the EEG signals (like [18]). To select the best model checkpoint, we used the same criterion as our model (loss value) instead of accuracy. This resulted in 85.9 (a), 78.0 ($F1^M$) and 80.2 (κ) on input of C4-A1 EEG channel (around 1% higher than ST-Q^O) and 85.1 (a), 76.2 ($F1^M$) and 78.8 (κ) on C3-A2 EEG channel (almost the same as ST-Q^O). This shows that [18] performs slightly better than our model using only one EEG channel. We replaced their training method (model validation after training on some batches) with our training method (model validation after iterating through the whole training set), but found no significant influence on the performance.

We hypothesize that the slight difference in performance may be due to our **multi-modal input versus their single channel input**. We can see that the performance on using multi-modal, multi-channel input is almost equal to single channel input with only 1% difference. This would mean that one channel has enough information to identify stages uniquely. However, our multi-modal XAI experiments (cf. Section 7.2) show that the model has learnt the stage-specific modality importance, which conforms to the AASM guidelines. We hypothesize that the content of the information condensed from all the channels may have led to better prediction for some of the epochs (wrongly classified by a single channel model) and at the same time, to more confusion for some other epochs (correctly classified by a single channel model), resulting in an overall similar accuracy. To verify this hypothesis, we calculated the number of prediction mismatches between Sors et al. [18] (model trained on our dataset split) and ST-Q^O on our test set (containing 580 PSGs). We found that 7.4% of the epochs were incorrectly classified by ST-Q^O, but correctly by Sors' model, 7.0% were incorrectly classified by Sors' model, but correctly by ST-Q^O, and for 85.6% of the epochs, both models agreed. This indicates, that model ensembling and/or multi-modal models with attention mechanisms could improve prediction.

9.6. Comparison to human performance

We compared the agreement between our predicted class and ground-truth with previously reported **human inter-annotator agreement scores** [47,46]. The agreement was reported for correctly classified instances and therefore, we compare their scores to the TPR from our confusion matrix of ST-Q^O (Fig. 5b) and report the comparison in Table 5. The results from Whitney et al. [46] are more comparable to our model than Rosenberg et al. [47] as the former reported their inter-annotator agreement for SHHS dataset. However, please note the difference in the number of epochs among the 3 studies (cf. Table 5). From the scores, we can conclude that our model's overall agreement with the ground truth is comparable to the agreement between humans. Our model has a better agreement than humans from both [47,46] for W and N3, comparable to [47] and better than [46] for N2 and REM and less than [47], but better than [46] for N1.

Table 5
Comparison of ST-Q^O model's performance to inter-annotator agreement in terms of TPR (%) rounded off to nearest integer.

Method	Overall	W	N1	N2	N3	REM	Epochs
ST-Q^O	85	**93**	33	84	**82**	88	586,168
Rosenberg et al. [47]	83	84	**63**	**85**	67	**91**	1800
Whitney et al. [46]*	**87**	89	23	79	69	78	29,507

* Values reported in [46] are averaged over results from 3 pairs of annotators.

[10] We chose [18] over [20], due to [18]'s code availability.

10. Conclusion and future work

Our model can predict sleep stages with an overall accuracy of 85.5% and stages W, N1, N2, N3 and REM with an F1 score of 92.5%, 41.3%, 84.8%, 76.3% and 89.1% respectively on SHHS. We created a multi-modal model which can learn from EEG, EOG and EMG inputs. We evaluated various architecture choices and found sequential learning (Bi-LSTM) improved predictive performance over spatio-temporal filtering (CNN), while residual connections did not. Through various post-hoc interpretability techniques, we found that our model conforms to the AASM guidelines. Thus, our model can be used to support sleep technologists in annotating sleep stages and explaining the reason for the automated annotation.

Through our multi-modal versus single channel experiments, we found that the single-channel model by Sors et al. [18] slightly outperforms our multi-modal model, while approx. 7% of the epochs are correctly classified by one, but not the other model. Moreover, through modality occlusion, we found that specific modalities are important for predicting specific stages. Therefore, we suggest that future work could investigate automatic channel selection for multi-modal sleep scoring models.

In a clinical setting, sleep scoring is used for diagnosing sleep disorders. Misclassifications that do not change this diagnosis, should be considered less serious. Evaluating our model based on predictive performance for diagnosing disorders is left for future work.

Conflict of interest

None of the authors have any conflict of interest.

Acknowledgment

This work was supported by *Pioneers in Healthcare Innovation Fund 2017 fortheproject "DeepSleep"awarded by Menzis Health Insurance and partially funded by Faculty of EEMCS, University of Twente*. The authors would like to thank Mike van Klooster, Thomas Oosterveld, Mirjam Stappenbelt-Groot Kormelink and Marleen Tjepkema-Cloostermans from Medisch Spectrum Twente for very helpful discussions on sleep scoring guidelines and data collection from the hospital; Xenia Hoppenbrouwer and Ainara Garde from University of Twente for very helpful discussions related to processing the signals and for guidance at the start of the project; and Jörg Schlötterer for improving the paper during the review process.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.artmed.2021.102038.

References

[1] AASM, Sleep (Polysomnographic) Technologist, American Academy of Sleep Medicine, https://aasm.org/technologist-description/ [accessed 15.10.20].
[2] Iber C, Ancoli-Israel S, Chesson AL, Quan SF, et al. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications, vol. 1. Westchester, IL: American Academy of Sleep Medicine; 2007.
[3] Hobson JA. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects: A. rechtschaffen and a. kales (editors). 1969 (public health service, us government printing office, washington, dc, 1968, 58 p., 4.00).
[4] Fraiwan L, Lweesy K, Khasawneh N, Wenz H, Dickhaus H. Automated sleep stage identification system based on time-frequency analysis of a single eeg channel and random forest classifier. Comput Methods Programs Biomed 2012;108(1):10–9.
[5] Li X, Cui L, Tao S, Chen J, Zhang X, Zhang G-Q. Hyclasss: a hybrid classifier for automatic sleep stage scoring. IEEE J Biomed Health Inform 2017;22(2):375–85.
[6] Koley B, Dey D. An ensemble system for automatic sleep stage classification using single channel eeg signal. Comput Biol Med 2012;42(12):1186–95.
[7] Ebrahimi F, Mikaeili M, Estrada E, Nazeran H. Automatic sleep stage classification based on eeg signals by using neural networks and wavelet packet coefficients. 2008 30th annual international conference of the IEEE engineering in medicine and biology society 2008:1151–4.
[8] Lajnef T, Chaibi S, Ruby P, Aguera P-E, Eichenlaub J-B, Samet M, et al. Learning machines and sleeping brains: automatic sleep stage classification using decision-tree multi-class support vector machines. J Neurosci Methods 2015;250:94–105.
[9] Malafeev A, Laptev D, Bauer S, Omlin X, Wierzbicka A, Wichniak A, et al. Automatic human sleep stage scoring using deep neural networks. Front Neurosci 2018;12:781.
[10] Vilamala A, Madsen KH, Hansen LK. Deep convolutional neural networks for interpretable analysis of eeg sleep stage scoring. 2017 IEEE 27th international workshop on machine learning for signal processing (MLSP) 2017:1–6.
[11] Supratak A, Dong H, Wu C, Guo Y. Deepsleepnet: a model for automatic sleep stage scoring based on raw single-channel eeg. IEEE Trans Neural Syst Rehabil Eng 2017;25(11):1998–2008.
[12] Chambon S, Galtier MN, Arnal PJ, Wainrib G, Gramfort A. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. IEEE Trans Neural Syst Rehabil Eng 2018;26(4):758–69.
[13] Gunning D. Explainable artificial intelligence (xai). Defense advanced research projects agency. 2018.
[14] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. European conference on computer vision 2014:818–33.
[15] Fernandez-Blanco E, Rivero D, Pazos A. Eeg signal processing with separable convolutional neural network for automatic scoring of sleeping stage. Neurocomputing 2020;410:220–8.
[16] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014 (arXiv preprint), arXiv:1409.1556.
[17] Tsinalis O, Matthews PM, Guo Y, Zafeiriou S. Automatic sleep stage scoring with single-channel eeg using convolutional neural networks. 2016 (arXiv preprint), arXiv:1610.01683.
[18] Sors A, Bonnet S, Mirek S, Vercueil L, Payen J-F. A convolutional neural network for sleep stage scoring from raw single-channel EEG. Biomed Signal Process Control 2018;42:107–14.
[19] Mousavi S, Afghah F, Acharya UR. Sleepeegnet: automated sleep stage scoring with sequence to sequence deep learning approach. PLOS ONE 2019;14(5).
[20] Biswal S, Sun H, Goparaju B, Westover MB, Sun J, Bianchi MT. Expert-level sleep scoring with deep neural networks. J Am Med Inform Assoc 2018;25(12):1643–50.
[21] Paisarnsrisomsuk S, Sokolovsky M, Guerrero F, Ruiz C, Alvarez SA. Deep sleep: convolutional neural networks for predictive modeling of human sleep time-signals. KDD Deep Learn Day 2018.
[22] Yildirim O, Baloglu UB, Acharya UR. A deep learning model for automated sleep stages classification using psg signals. Int J Environ Res Public Health 2019;16(4):599.
[23] Phan H, Andreotti F, Cooray N, Chén OY, De Vos M. Joint classification and prediction cnn framework for automatic sleep stage classification. IEEE Trans Biomed Eng 2018;66(5):1285–96.
[24] Cleveland children's sleep and health study. National Sleep Research Resource, Boston, MA, USA. Available from: https://sleepdata.org/datasets/ccshs [Online].
[25] The cleveland family study. National Sleep Research Resource, Boston, MA, USA. Available from: https://sleepdata.org/datasets/cfs [Online].
[26] Hassan AR, Subasi A. A decision support system for automated identification of sleep stages from single-channel eeg signals. Knowl-Based Syst 2017:115–24.
[27] Alickovic E, Subasi A. Ensemble svm method for automatic sleep stage classification. IEEE Trans Instrum Meas 2018;67(6):1258–65.
[28] Khalighi S, Sousa T, Pires G, Nunes U. Automatic sleep staging: a computer assisted approach for optimal combination of features and polysomnographic channels. Expert Syst Appl 2013;40(17):7046–59.
[29] Alvarez-Estevez D, Fernández-Varela I. Dealing with the database variability problem in learning from medical data: an ensemble-based approach using convolutional neural networks and a case of study applied to automatic sleep scoring. Comput Biol Med 2020:103697.
[30] Virgilio G CD, Sossa A JH, Antelis JM, Falcón LE. Spiking neural networks applied to the classification of motor tasks in eeg signals. Neural Netw 2020;122:130–43. https://doi.org/10.1016/j.neunet.2019.09.037. http://www.sciencedirect.com/science/article/pii/S0893608019303193.
[31] Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE Access 2018;6:52138–60.
[32] Letham B, Rudin C, McCormick TH, Madigan D, et al. Interpretable classifiers using rules and bayesian analysis: building a better stroke prediction model. Ann Appl Stat 2015;9(3):1350–71.
[33] Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining 2015:1721–30.
[34] Cortez P, Embrechts MJ. Opening black box data mining models using sensitivity analysis. 2011 IEEE symposium on computational intelligence and data mining (CIDM) 2011:341–8.
[35] Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE 2015;10(7):1–46. https://doi.org/10.1371/journal.pone.0130140.
[36] Quan SF, Howard BV, Iber C, Kiley JP, Nieto FJ, O'Connor GT, et al. The sleep heart health study: design, rationale, and methods. Sleep 1997;20(12):1077–85.
[37] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–20.

[38] Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 2015 (arXiv preprint), arXiv:1502.03167.
[39] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15 (1):1929–58.
[40] Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Trans Signal Process 1997;45(11):2673–81.
[41] Leevy JL, Khoshgoftaar TM, Bauder RA, Seliya N. A survey on addressing high-class imbalance in big data. J Big Data 2018;5(1):42.
[42] Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H. Understanding neural networks through deep visualization. 2015 (arXiv preprint), arXiv:1506.06579.
[43] Kingma DP, Ba J. Adam: A method for stochastic optimization. 2014 (arXiv preprint), arXiv:1412.6980.
[44] Vorontsov E, Trabelsi C, Kadoury S, Pal C. On orthogonality and learning recurrent networks with long term dependencies. Proceedings of the 34th international conference on machine learning-volume 70 2017:3570–8.
[45] Estrada E, Nazeran H, Barragan J, Burk J, Lucas E, Behbehani K. Eog and emg: two important switches in automatic sleep stage classification. 2006 international conference of the IEEE engineering in medicine and biology society 2006:2458–61.
[46] Whitney CW, Gottlieb DJ, Redline S, Norman RG, Dodge RR, Shahar E, et al. Reliability of scoring respiratory disturbance indices and sleep staging. Sleep 1998; 21(7):749–57.
[47] Rosenberg RS, Van Hout S. The American academy of sleep medicine inter-scorer reliability program: sleep stage scoring. J Clin Sleep Med 2013;9(1):81–7.