SERF: Interpretable Sleep Staging using Embeddings, Rules, and Features

Irfan Al-Hussaini
Georgia Institute of Technology
Atlanta, Georgia, USA
alhussaini.irfan@gatech.edu

Cassie S. Mitchell
Georgia Institute of Technology
Atlanta, Georgia, USA
cassie.mitchell@bme.gatech.edu

ABSTRACT

The accuracy of recent deep learning based clinical decision support systems is promising. However, lack of model interpretability remains an obstacle to widespread adoption of artificial intelligence in healthcare. Using sleep as a case study, we propose a generalizable method to combine clinical interpretability with high accuracy derived from black-box deep learning.

Clinician-determined sleep stages from polysomnogram (PSG) remain the gold standard for evaluating sleep quality. However, PSG manual annotation by experts is expensive and time-prohibitive. We propose SERF, *interpretable Sleep staging using Embeddings, Rules, and Features* to read PSG. SERF provides interpretation of classified sleep stages through meaningful features derived from the AASM Manual for the Scoring of Sleep and Associated Events.

In SERF, the embeddings obtained from a hybrid of convolutional and recurrent neural networks are transposed to the interpretable feature space. These representative interpretable features are used to train simple models like a shallow decision tree for classification. Model results are validated on two publicly available datasets. SERF surpasses the current state-of-the-art for interpretable sleep staging by 2%. Using Gradient Boosted Trees as the classifier, SERF obtains 0.766 κ and 0.870 AUC-ROC, within 2% of the current state-of-the-art black-box models.

CCS CONCEPTS

• **Applied computing** → **Health informatics**; • **Computing methodologies** → **Supervised learning**; **Machine learning algorithms**; **Knowledge representation and reasoning**.

KEYWORDS

sleep stage classification, interpretable, representation learning, embedding, cnn, lstm, eeg, rule learning

ACM Reference Format:
Irfan Al-Hussaini and Cassie S. Mitchell. 2022. SERF: Interpretable Sleep Staging using Embeddings, Rules, and Features. In *Proceedings of the 31st ACM Int'l Conference on Information and Knowledge Management (CIKM '22), Oct. 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3511808.3557700

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10.
https://doi.org/10.1145/3511808.3557700

1 INTRODUCTION

The prevalence of electronic health records has led to an abundance of patient health data [1, 16]. Meanwhile, recent advances in deep learning have shown great promise in utilizing these data for accurate clinical decision support systems. However, the lack of interpretability remains an obstacle to widespread adoption for a high stakes application like healthcare [6, 11, 15, 35, 40]. On the other hand, simple linear models are not accurate enough to be used by clinicians for decision-making [23]. Can a clinical decision support system be developed that is interpretable and accurate?

In this paper, we focus on sleep staging. Sleep stages annotated by clinicians from polysomnograms (PSG) remain the gold standard for evaluating sleep quality and diagnosing sleep disorders.

Sleep disorders affect 50–70 million US adults and 150 million in developing countries worldwide [13]. Sleep staging is the most important precursor to sleep disorder diagnoses such as insomnia, narcolepsy, or sleep apnea [34]. However, the clinician-determined sleep staging, which is the gold standard, is labor-intensive and expensive [14]. Neurologists visually analyze multi-channel PSG signals and provide empirical scores of sleep stages, including wake, rapid eye movement (REM), and the non-REM stages N1, N2, and N3, following guidelines stated in the American Academy of Sleep Medicine (AASM) Manual for the Scoring of Sleep and Associated Events [4]. Such a visual task is cumbersome and takes several hours for one sleep expert to annotate a patient's PSG signals from a single night [42].

Automated algorithms for sleep staging alleviate these limitations. Deep learning methods have successfully automated annotation of sleep stages by using convolutional neural networks (CNN) [7, 28, 33, 41], recurrent neural networks [10, 29], recurrent convolutional neural networks [5, 36], deep belief nets [24], autoencoders [37], attention [21, 30, 31], and graph convolutional neural networks [17, 21]. Although deep learning models can produce accurate sleep staging classification, they are often treated as black-box models that lack interpretability [22]. Lack of interpretability limits the adoption of the deep learning models in practice because clinicians must understand the reason behind each classification to avoid data noise and unexpected bias [2, 35]. Furthermore, current clinical practice at sleep labs relies on the AASM sleep scoring manual [4], which is interpretable for clinical experts but lacks precise definitions needed for a robust computational model [3].

Thus, an automated model for sleep staging should ideally be as clinically interpretable as the sleep scoring manual and as accurate as the black-box neural network models. To this end, we propose SERF, *interpretable Sleep staging using Embeddings, Rules, and Features*, which combines clinical interpretability with the accuracy derived from a deep learning model. It provides clinically

Table 1: Datasets

	Number of Subjects	Sampling Frequency (Hz)	Number of Channels	Annotation Schema
ISRUC [20]	100	200	9	AASM [4]
PhysioNet-EDFX [12, 19]	197	100	4	R&K [32, 38]

meaningful explanations, derived from the AASM manual [4], with each sleep stage prediction. SERF surpasses the interpretable sleep staging baseline [3] and performs within 2% of state-of-the-art black-box deep learning models [28].

2 DATA

The following two publicly available datasets, summarized in Table 1, are used to evaluate the performance of SERF:

- PhysioNet EDFX [12, 19]: The PhysioNet EDFX database contains 197 whole-night PSG sleep recordings. The corresponding hypnograms were scored by technicians according to the Rechtschaffen and Kales guidelines [32]. Sleep Stages N3 and N4 were combined to adhere to AASM standards. 153 subjects were healthy Caucasians without any sleep-related medication. 44 subjects had mild difficulty falling asleep. It contains two Electroencephalography (EEG) channels, one Electrooculography (EOG) channel, and one Electromyography (EMG) channel sampled at 100 Hz.
- ISRUC [20]: The ISRUC dataset contains PSG recordings of 100 subjects with evidence of having sleep disorders. The data was collected from 55 male and 45 female subjects, whose ages range from 20 to 85 years old, with an average age of 51. The corresponding hypnograms were manually scored by technicians according to the AASM manual [4]. It includes six EEG channels (F3, F4, C3, C4, O1, and O2), two EOG channels (E1 and E2), and a single EMG channel sampled at 200 Hz.

3 METHOD

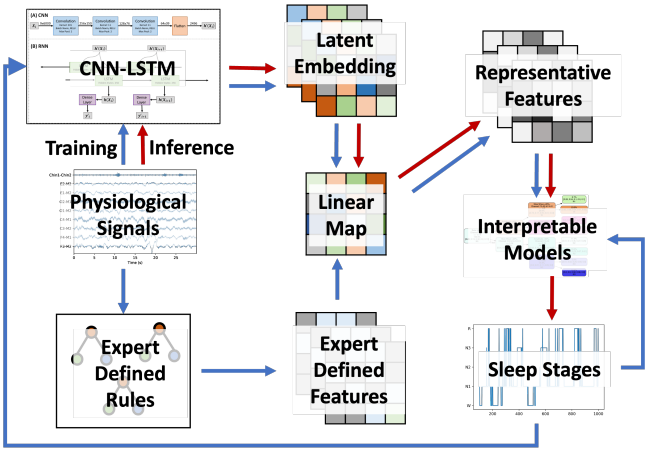

Figure 1: SERF Framework

SERF predicts sleep stages using PSG data through an interpretable model derived from expert-defined features and embeddings from a deep neural network. The method is explained using the ISRUC dataset. Table 1 and Section 2 can be used to find the corresponding metrics for PhysioNet-EDFX.

The input data are multi-channel PSG signals. It consists of multivariate continuous time-series data, $\mathcal{X}$, partitioned into N segments called epochs, denoted as $\mathcal{X} = \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_N\}$. Each epoch $\boldsymbol{X}_n \in \mathbb{R}^{9\times 6{,}000}$ is 30 seconds long and contains 9 physiological signals from 9 channels. The sampling frequency is 200Hz. Each epoch, $\boldsymbol{X}_n$, also has an associated sleep stage label $y_n \in \{\text{Wake, REM, N1, N2, N3}\}$ provided by a clinical expert. The goal is to predict a sequence of sleep stages, $\boldsymbol{S} = \{s_1, \ \ldots \ s_N\}$ based on $\mathcal{X}$ so that they mimic the labels provided by human experts, $Y = \{y_1, \ \ldots \ y_N\}$. In SERF, these predictions also contain meaningful explanations derived from expert-defined features. The SERF framework shown in Figure 1 comprises the following steps:

3.1 Latent Embedding

The multivariate PSG signals are embedded using CNN-LSTM to capture translation invariant and complex patterns. The CNN is composed of 3 convolutional layers. Each convolutional layer is followed by batch normalization, ReLU activation, and max pooling. Using a kernel size of 201, the convolutions in the first layer extract features based on 1-second segments. Subsequent layers have a kernel size of 11. The output channels of the three convolution layers are 256, 128, and 64. The output of the final convolutional layer is flattened and fed into a single layer of bi-directional Long Short-Term Memory (LSTM) cells with 256 hidden states to capture temporal relations between epochs. This results in 512 hidden states for each epoch, $\boldsymbol{X}_i$ and represents the latent embedding used in subsequent steps:

$$\boldsymbol{h}(\boldsymbol{X}_i) \in \mathbb{R}^{512}$$

A single fully connected layer with softmax activation is then used to predict the five sleep stages:

$$\boldsymbol{z}_i = \boldsymbol{W}^T h(\boldsymbol{X}_i) + \boldsymbol{b}$$

$$\boldsymbol{s}_i = softmax(\boldsymbol{z}_i)$$

where $\mathbf{W} \in \mathbb{R}^{512\times 5}$ is the weight matrix, $\mathbf{b} \in \mathbb{R}^5$ is the bias vector, and $\boldsymbol{s}_i$ is the estimated probabilities of all 5 sleep stages at epoch i. Cross-entropy loss is used to train the model:

$$L(\boldsymbol{y}_i, \boldsymbol{s}_i) = -\sum_{j}^{5} \boldsymbol{y}_i[j] log(\boldsymbol{s}_i[j])$$

where $L(\boldsymbol{y}_i, \ \boldsymbol{s}_i)$ is the estimated cross-entropy loss for epoch i between human labels $\boldsymbol{y}_i$ and the predicted probabilities $\boldsymbol{s}_i$. After training on sleep stage prediction, the latent embedding $\boldsymbol{h}(\mathcal{X}) \in \mathbb{R}^{N\times 512}$ is obtained from the hidden states of the LSTM.

3.2 Expert Defined Features

Concurrently, each epoch is encoded into a feature vector. Expert suggestions are incorporated to supplement the technical guidelines in the AASM manual [4]. Using this rule augmentation procedure, a set of M' features are extracted, $\boldsymbol{F'}(\boldsymbol{X_n}) = \left[f'_1(\boldsymbol{X}_n), \ldots f'_{M'}(\boldsymbol{X}_n)\right]$, where element $f'_j(\boldsymbol{X}_n)$ is the function generating feature j for epoch $\boldsymbol{X}_n$.

These meaningful features are described below:

- Sleep spindles are bursts of oscillatory signals originating from the thalamus and depicted in EEG [9]. It is a discriminatory feature of N2.
- Slow-wave sleep is prominent in N3 and is marked by low-frequency and high-amplitude delta activity [39].
- Delta (0.5–4Hz) waves feature in N3, Theta (4–8 Hz) in N1, Alpha (8–12 Hz), and Beta (>12 Hz) help to differentiate between Wake and N1 [18]. EMG bands determine the muscle tone used to distinguish between REM and Wake [25]. Each band's Power Spectral Density (PSD) is calculated using a multi-taper spectrogram.
- Amplitude is vital in finding K Complexes, Chin EMG amplitude, and Low Amplitude Mixed Frequency and thus used to differentiate Wake, REM, N1, and N2.
- Mean, Variance, Kurtosis, and Skew are used to capture the distribution of values in the channel. It can help detect outlier values that highlight signatures such as K-Complexes.

The importance of expert-defined features is analyzed using the ANOVA test to select the top 90% of the most discriminative features, $\boldsymbol{F}$. It reduces the number of features from M' to M, where $\boldsymbol{F} \subset \boldsymbol{F'}$. These M features from N epochs leads to our feature matrix $\boldsymbol{F}(\mathcal{X}) \in \mathbb{R}^{N \times M}$, which forms the basis of the interpretation module of our framework where an element $f_j(\boldsymbol{X}_i)$ is the value of feature j at the epoch $\boldsymbol{X}_i$. Number of features in ISRUC dataset: $M' = 87$ and $M = 78$, and in Physionet dataset: $M' = 38$ and $M = 34$ due to fewer channels.

3.3 Linear Map

A linear map, $\boldsymbol{T}$, combines the inputs encoded by features and latent embeddings. After the feature matrix, $\boldsymbol{F}(\mathcal{X})$, and latent CNN-LSTM embedding matrix, $\boldsymbol{h}(\mathcal{X})$, are generated, a linear transformation $\boldsymbol{T}$ is learned, which maps the features into the latent space defined by the embeddings. The linear transformation matrix, $\boldsymbol{T}$, is learned using ridge regression:

$$\min_{T} ||\boldsymbol{h}(\mathcal{X}) - \boldsymbol{F}(\mathcal{X})\,T||_2^2 + ||T||_2^2$$

3.4 Representative Features

The linear map, $\boldsymbol{T}$, is used to generate a representative feature, $\boldsymbol{s}_j$, for each epoch using the embedding vectors $\boldsymbol{h}(\boldsymbol{X}_n)$. These representative features collectively form the similarity matrix $\boldsymbol{S}$ and is used to train an interpretable classifier like a shallow decision tree.

Given an epoch, $\boldsymbol{X}_j$, the CNN-LSTM embedding module is used to obtain the embedding, $\boldsymbol{h}(\boldsymbol{X}_j)$. A representative feature similarity matrix, $\boldsymbol{S}$, is then generated using the linear map, T:

$$S = \boldsymbol{h}(X_j)T^T$$

This representative feature similarity matrix is used as input to simple classifiers such as a shallow decision tree. When a new PSG is provided, $\boldsymbol{X}$, the embedding vector $\boldsymbol{h}(\boldsymbol{X})$ is first generated using the CNN-LSTM network followed by the representative feature similarity matrix $\boldsymbol{S} = \boldsymbol{h}(\boldsymbol{X}_i)\boldsymbol{T}^T$. These representative features are used as input to simple classifiers and form the basis for model interpretability.

Table 2: Model Evaluation[a]

Model	Accuracy (%)		ROC-AUC (%)		Cohen's κ		Macro F1	
	EDFx	ISRUC	EDFx	ISRUC	EDFx	ISRUC	EDFx	ISRUC
SERF-DT	81.2	80.5	82.5	85.8	0.735	0.747	0.719	0.768
SERF-XG	82.3	81.9	84.4	87.0	0.753	0.766	0.753	0.789
SERF-GB	82.2	81.7	84.8	87.0	0.753	0.763	0.758	0.789
SERF-LR	82.9	79.5	85.0	85.3	0.762	0.733	0.759	0.773
Features-XG	81.0	77.4	83.2	83.0	0.734	0.704	0.732	0.722
Features-DT	68.7	71.6	74.3	79.4	0.555	0.629	0.583	0.665
SLEEPER-DT [3]	78.8	78.0	81.5	83.4	0.704	0.712	0.696	0.730
SLEEPER-GB [3]	80.7	79.7	82.8	85.1	0.729	0.736	0.721	0.756
U-Time [28]	86.2	84	88.3	88.8	0.810	0.793	0.811	0.816
CNN-LSTM	86.4	83.1	88.6	88.8	0.813	0.783	0.815	0.819
1D-CNN [3]	84.4	82.5	86.6	87.2	0.784	0.773	0784	0.789

[a]XG: DART Gradient Boosted Trees, DT: Decision Tree, LR: Logistic Regression, GB: Gradient Boosted Trees

4 EXPERIMENTS

Implementation Details: SERF was built using PyTorch 1.0 [26], scikit-learn [27], and XGBoost [8]. A batch size of 1000 samples from 1 PSG is used. Each model is trained for 20 epochs with a learning rate of 10^{-4} using ADAM as the optimization method. The data is randomly split by subjects into a training and test set in a 9:1 ratio with the same seed for each experiment. For each dataset, the training set is used to fix model parameters, and the test set is used to obtain performance metrics. The same model hyperparameters and feature extraction schema are used to prevent overfitting and ensure consistent performance across different datasets.

Baselines:

- Convolutional Neural Network with a stacked bi-directional LSTM layer (CNN-LSTM): the black-box model used in obtaining the signal embeddings.
- 1D-Convolutional Neural Network (1D-CNN): a black-box model proposed in [3].
- Expert Feature Matrix, $\boldsymbol{F}(\mathcal{X})$, as input to simple classifiers.
- SLEEPER [3]: an interpretable sleep staging algorithm based on prototypes.
- U-Time [28]: state-of-the-art black-box deep learning model which adapts the U-Net architecture for sleep staging.

Metrics:

- Accuracy = $\frac{|\mathcal{Y} \cap \mathcal{Y}'|}{N}$
- Sensitivity, $S^{(k)} = \frac{|\mathcal{Y}^{(k)} \cap \mathcal{Y}'^{(k)}|}{|\mathcal{Y}'^{(k)}|}$
- Precision, $P^{(k)} = \frac{|\mathcal{Y}^{(k)} \cap \mathcal{Y}'^{(k)}|}{|\mathcal{Y}^{(k)}|}$
- F1 score = $\frac{2 * P * S}{P + S}$
- Cohen's $\kappa = \frac{Acc - p_e}{1 - p_e}$, where $p_e = \frac{1}{N^2} \sum_k^5 \left|\mathcal{Y}^{(k)}\right| \left|\mathcal{Y}'^{(k)}\right|$

Given expert annotations $\mathcal{Y}'$ and predicted stages $\mathcal{Y}$ of size N, $k = \{W, N1, N2, N3, R\}$ indicating the sleep stage, and $\left|\mathcal{Y}'^{(k)}\right|$ ($\left|\mathcal{Y}^{(k)}\right|$) is the number of human (algorithm) labels from sleep stage k.

Results: The results from experiments are compared in Table 2 and 3. SERF performs within 2% of state-of-the-art black box models such as U-Time [28] and far exceeds the performance of expert feature-based models. Table 3 shows that for all sleep stages other than Wake, SERF surpassed SLEEPER [3] and was comparable to black-box models, U-Time [28] and CNN-LSTM. N1 is a challenging sleep stage to identify where SERF was comparable to black-box

Table 3: Sensitivity across Sleep Stages

Model	Wake		N1		N2		N3		R	
	EDFx	ISRUC	EDFx	ISRUC	EDFx	ISRUC	EDFx	ISRUC	EDFx	ISRUC
SERF-DT	0.897	0.889	0.283	0.440	0.849	0.798	0.803	0.855	0.763	0.859
SERF-XG	0.892	0.894	0.386	0.481	0.863	0.809	0.823	0.889	0.802	0.870
SERF-GB	0.892	0.888	0.410	0.495	0.863	0.811	0.824	0.888	0.802	0.865
SERF-LR	0.904	0.895	0.398	0.535	0.865	0.768	0.832	0.847	0.799	0.819
Features-XG	0.891	0.873	0.340	0.324	0.842	0.768	0.828	0.859	0.761	0.788
Features-DT	0.757	0.821	0.110	0.250	0.742	0.717	0.692	0.789	0.612	0.749
SLEEPER-DT [3]	0.902	0.893	0.256	0.329	0.840	0.764	0.819	0.865	0.661	0.799
SLEEPER-GB [3]	0.911	0.902	0.313	0.397	0.848	0.787	0.835	0.876	0.700	0.816
U-Time [28]	0.941	0.890	0.552	0.504	0.897	0.839	0.749	0.898	0.883	0.951
CNN-LSTM	0.936	0.871	0.507	0.548	0.905	0.781	0.838	0.937	0.859	0.956
1D-CNN [3]	0.926	0.914	0.397	0.398	0.910	0.840	0.817	0.886	0.819	0.907

models. These results show that SERF can generalize better than other interpretable methods [3] and is comparable to black-box models [28] in identifying all stages.

SERF has the following improvements over the interpretable method, SLEEPER [3]: (1) SERF surpasses SLEEPER by 2-3% in all evaluation metrics. (2) SERF has a lower representative feature dimension by utilizing raw feature values instead of binary rules resulting in smaller matrices. (3) SERF learns a linear map using ridge regression with a lower dimension than the prototype matrix learned in SLEEPER using cosine similarity, thus resulting in a smaller model size. (4) The smaller feature size of SERF results in faster training and faster inference from the simple classifiers. (5) Meaningful feature value cutoffs are obtained at nodes of the decision tree, as seen in Figure 3, instead of just similarity indices.

The results also show the significance of different channels when building an interpretable model. The clinical sleep staging manual [4] utilizes all the 9 channels in the ISRUC dataset. Since the Physionet EDFx dataset only contains 4 channels there are some features that cannot be extracted. As a result, SERF, SLEEPER [3] and Feature models exhibit worse performance relative to black-box models for the EDFx dataset than ISRUC.

Figure 2 shows the significance of representative features for classifying each sleep stage based on SERF and gradient boosted trees. We focus on 2 of the top 7 features, which can be better attributed to guidelines in the sleep staging manual [4]. The rest of the top 7 features are general meaningful features rather than distinctive signal traits embedded in epochs. The 3rd feature, Spindle in the C3-A2 and C4-A1 contra-lateral channel pairs, is important in identifying REM stages. The manual states the absence of Spindles as a critical observation when annotating REM. The 7th most impactful feature, Slow Wave in the C3-A2 and C4-A1 contra-lateral channel pairs, contributes significantly to the distinction of N3 from N2. The most distinctive attribute clinicians look for in N3 is a slow wave.

5 INTERPRETATION

Figure 3 shows a decision tree of depth 4, based on SERF. The leftmost node denotes the root of the tree. The color indicates the most frequent sleep stage at a node, and the intensity is proportional to its purity. The five rows of each node contain the following: (1) the feature and the channels used, (2) the feature cutoff value, (3) the percentage of data passing through, (4) the ratio of each sleep stage in the following order: [Wake, N1, N2, N3, REM], (5) the most frequent sleep stage, in other words if classification is performed at that node, this label is assigned. Analyzing the resulting decision

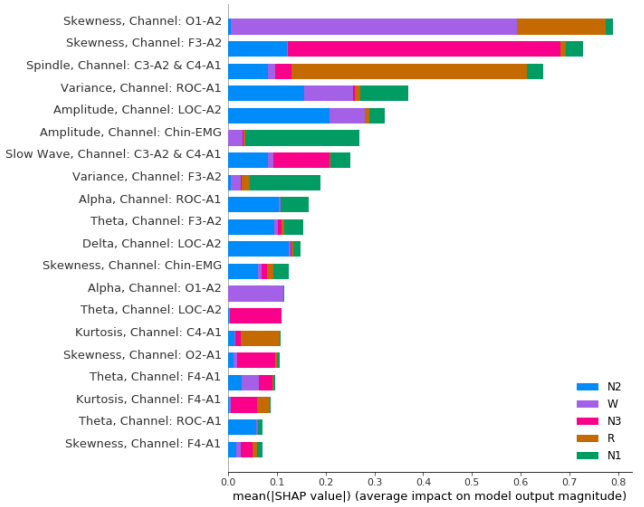

Figure 2: SERF-XG feature importance SHAP values (ISRUC)

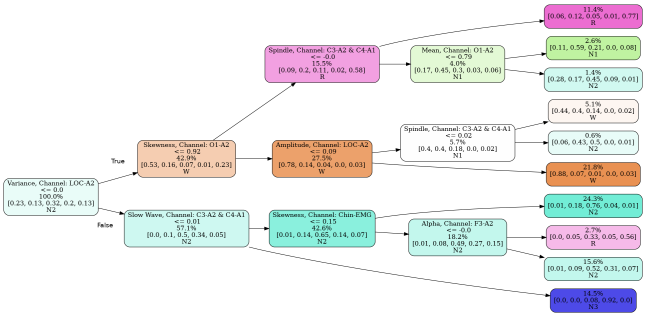

Figure 3: SERF-Decision Tree (ISRUC)

tree reveals some promising aspects of SERF. According to the sleep staging guidelines for human annotators [4], N3 is distinguished by the occurrence of slow waves, one of the underlying features of SERF. The bottom leaf node is partitioned using Slow Waves >= 0.01 in the C3-A2 & C4-A1 channel pair. 92% of this leaf node contains N3, while only 34% of the previous node contains it.

6 CONCLUSION

We propose a method, SERF, to provide accurate and interpretable clinical decision support and demonstrate it on automated sleep stage prediction. In order to achieve this goal, SERF combines embeddings from a deep neural network with clinically meaningful features. SERF achieves high performance metrics, comparable to state-of-the-art deep learning baselines. Moreover, the SERF expert feature module incorporates standard AASM guidelines to ensure the model enables transparent clinical interpretability, as illustrated using two qualitative case studies.

ACKNOWLEDGMENTS

This research was funded by NSF 1944247, NIH U19-AG056169, GT McCamish Award to C.M.

REFERENCES

[1] Julia Adler-Milstein and Ashish K Jha. 2017. HITECH Act drove large gains in hospital electronic health record adoption. *Health affairs* 36, 8 (2017), 1416–1422.

[2] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. 2018. Interpretable Machine Learning in Healthcare. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (Washington, DC, USA) *(BCB '18)*. Association for Computing Machinery, New York, NY, USA, 559–560. https://doi.org/10.1145/3233547.3233667

[3] Irfan Al-Hussaini, Cao Xiao, M Brandon Westover, and Jimeng Sun. 2019. SLEEPER: interpretable Sleep staging via Prototypes from Expert Rules. In *Machine Learning for Healthcare Conference*. PMLR, 721–739.

[4] Richard B. Berry, Rohit Budhiraja, Daniel J. Gottlieb, David Gozal, Conrad Iber, Vishesh K. Kapur, Carole L. Marcus, Reena Mehra, Sairam Parthasarathy, Stuart F. Quan, et al. 2012. Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events. *Journal of clinical sleep medicine* 8, 05 (2012), 597–619.

[5] Siddharth Biswal, Joshua Kulas, Haoqi Sun, Balaji Goparaju, M. Brandon Westover, Matt T. Bianchi, and Jimeng Sun. 2017. SLEEPNET: automated sleep staging system via deep learning. *arXiv preprint arXiv:1707.08262* (2017).

[6] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.

[7] Stanislas Chambon, Mathieu N. Galtier, Pierrick J. Arnal, Gilles Wainrib, and Alexandre Gramfort. 2018. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26, 4 (2018), 758–769.

[8] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[9] Luigi De Gennaro and Michele Ferrara. 2003. Sleep spindles: an overview. *Sleep medicine reviews* 7, 5 (2003), 423–440.

[10] Hao Dong, Akara Supratak, Wei Pan, Chao Wu, Paul M. Matthews, and Yike Guo. 2018. Mixed neural network approach for temporal sleep stage classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26, 2 (2018), 324–333.

[11] Radwa Elshawi, Mouaz H Al-Mallah, and Sherif Sakr. 2019. On the interpretability of machine learning-based model for predicting hypertension. *BMC medical informatics and decision making* 19, 1 (2019), 1–32.

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

[13] Antonio Guglietta. 2015. *Drug treatment of sleep disorders*. Springer.

[14] Antoine Guillot, Fabien Sauvet, Emmanuel H During, and Valentin Thorey. 2020. Dreem open datasets: Multi-scored sleep datasets to compare human and automated sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28, 9 (2020), 1955–1965.

[15] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9, 4 (2019), e1312.

[16] Peter B Jensen, Lars J Jensen, and Søren Brunak. 2012. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics* 13, 6 (2012), 395–405.

[17] Ziyu Jia, Youfang Lin, Jing Wang, Ronghao Zhou, Xiaojun Ning, Yuanlai He, and Yaoshuai Zhao. 2020. GraphSleepNet: Adaptive Spatial-Temporal Graph Convolutional Networks for Sleep Stage Classification.. In *IJCAI*. 1324–1330.

[18] Sharon Keenan and Max Hirshkowitz. 2011. Monitoring and staging human sleep. *Principles and practice of sleep medicine* 5 (2011), 1602–1609.

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

[20] Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. 2016. ISRUC-Sleep: a comprehensive public dataset for sleep researchers. *Computer methods and programs in biomedicine* 124 (2016), 180–192.

[21] Menglei Li, Hongbo Chen, and Zixue Cheng. 2022. An Attention-Guided Spatiotemporal Graph Convolutional Network for Sleep Stage Classification. *Life* 12, 5 (2022), 622.

[22] Zachary Chase Lipton. 2016. The Mythos of Model Interpretability. *CoRR* abs/1606.03490 (2016). http://arxiv.org/abs/1606.03490

[23] Alex John London. 2019. Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report* 49, 1 (2019), 15–21.

[24] Martin Längkvist, Lars Karlsson, and Amy Loutfi. 2012. Sleep Stage Classification Using Unsupervised Feature Learning. *Adv. Artif. Neu. Sys.* 2012 (2012), 5:5–5:5. http://dx.doi.org/10.1155/2012/107046

[25] Erik Naylor, Daniel V Aillon, Seth Gabbert, Hans Harmon, David A Johnson, George S Wilson, and Peter A Petillo. 2011. Simultaneous real-time measurement of EEG/EMG and L-glutamate in mice: a biosensor study of neuronal activity during sleep. *Journal of Electroanalytical Chemistry* 656, 1-2 (2011), 106–113.

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[28] Mathias Perslev, Michael Jensen, Sune Darkner, Poul Jø rgen Jennum, and Christian Igel. 2019. U-Time: A Fully Convolutional Network for Time Series Segmentation Applied to Sleep Staging. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 4415–4426. http://papers.nips.cc/paper/8692-u-time-a-fully-convolutional-network-for-time-series-segmentation-applied-to-sleep-staging.pdf

[29] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. 2019. SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27, 3 (2019), 400–410.

[30] Huy Phan, Kaare B Mikkelsen, Oliver Chen, Philipp Koch, Alfred Mertins, and Maarten De Vos. 2022. SleepTransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering* (2022).

[31] Wei Qu, Zhiyong Wang, Hong Hong, Zheru Chi, David Dagan Feng, Ron Grunstein, and Christopher Gordon. 2020. A residual based attention model for eeg based sleep staging. *IEEE journal of biomedical and health informatics* 24, 10 (2020), 2833–2843.

[32] Allan Rechtschaffen. 1968. A manual for standardized terminology, techniques and scoring system for sleep stages in human subjects. *Brain information service* (1968).

[33] Arnaud Sors, Stéphane Bonnet, Sébastien Mirek, Laurent Vercueil, and Jean-François Payen. 2018. A convolutional neural network for sleep stage scoring from raw single-channel EEG. *Biomedical Signal Processing and Control* 42 (2018), 107–114.

[34] Jens B. Stephansen, Alexander N. Olesen, Mads Olsen, Aditya Ambati, Eileen B. Leary, Hyatt E. Moore, Oscar Carrillo, Ling Lin, Fang Han, Han Yan, et al. 2018. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature Communications* 9, 1 (2018), 5229–5229.

[35] Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. 2020. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 5 (2020), e1379.

[36] A. Supratak, H. Dong, C. Wu, and Y. Guo. 2017. DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25, 11 (2017), 1998–2008. https://doi.org/10.1109/TNSRE.2017.2721116

[37] Orestis Tsinalis, Paul M. Matthews, and Yike Guo. 2016. Automatic Sleep Stage Scoring Using Time-Frequency Analysis and Stacked Sparse Autoencoders. *Annals of Biomedical Engineering* 44, 5 (2016), 1587–1597. https://doi.org/10.1007/s10439-015-1444-y

[38] B Van Sweden, B Kemp, HAC Kamphuisen, and EA Van der Velde. 1990. Alternative electrode placement in (automatic) sleep scoring (f pz-cz/p z-oz versus c4-at). *Sleep* 13, 3 (1990), 279–283.

[39] Nathan W Whitmore, Adrianna M Bassard, and Ken A Paller. 2022. Targeted memory reactivation of face-name learning depends on ample and undisturbed slow-wave sleep. *npj Science of Learning* 7, 1 (2022), 1–6.

[40] Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. 2019. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine* 25, 9 (2019), 1337–1340.

[41] Bufang Yang, Xilin Zhu, Yitian Liu, and Hongxing Liu. 2021. A single-channel EEG based automatic sleep stage classification method leveraging deep one-dimensional convolutional neural network and hidden Markov model. *Biomedical Signal Processing and Control* 68 (2021), 102581.

[42] Hanrui Zhang, Xueqing Wang, Hongyang Li, Soham Mehendale, and Yuanfang Guan. 2022. Auto-annotating sleep stages based on polysomnographic data. *Patterns* 3, 1 (2022), 100371. https://doi.org/10.1016/j.patter.2021.100371