Performance and utility trade-off in interpretable sleep staging

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent advances in deep learning have led to the development of models approaching human level of accuracy. However, healthcare remains an area lacking in widespread adoption. The safety-critical nature of healthcare results in a natural reticence to put these black-box deep learning models into practice. In this paper, we explore interpretable methods for a clinical decision support system, sleep staging, based on physiological signals such as EEG, EOG, and EMG. A recent work has shown sleep staging using simple models and an exhaustive set of features can perform nearly as well as deep learning approaches but only for certain datasets. Moreover, the utility of these features from a clinical standpoint is unclear. On the other hand, the proposed framework, `NormIntSleep` shows that by representing deep learning embeddings using normalized features, great performance can be obtained across different datasets. `NormIntSleep` performs 4.5% better than the exhaustive feature-based approach and 1.5% better than other representation learning approaches. An empirical comparison between the utility of the interpretations of these models highlights the improved alignment with clinical expectations when performance is traded-off slightly.

1 Introduction

There has been a steady accumulation of digital records of patient health data due to the increasingly widespread adoption of electronic health records [1, 2, 3]. Breakthroughs in deep learning have leveraged this influx of clinical data to create increasingly complex and capable systems [4, 5, 6]. However, the lack of interpretability and explainability of deep learning models prevents most of them from being used in practice because clinicians need to understand the explanation behind each classification to avoid noise and bias [7, 8, 9, 10, 11]. Although challenging to design due to the manual effort required, linear models paired with a robust set of features can provide a degree of interpretability [12]. However, linear models paired with complex features may result in models whose interpretation does not have clinical relevance. In this paper, using sleep staging as a case study, we take a deeper dive into the trade-off between clinical relevance of explanations and performance, and propose a model attempting to tackle this issue.

Around 70 million US adults are affected by sleep disorders [13, 14] such as insomnia, narcolepsy, or sleep apnea. The most crucial step for the diagnosis of sleep disorders is sleep staging [15]. The gold standard for sleep staging remains manual annotation of Polysomnogram (PSG) by clinicians [16]. During this process, clinicians inspect the PSG signals from multiple channels, and annotate each 30s segment with one of five sleep stages, i.e. wake, rapid eye movement (REM), and the non-REM stages N1, N2, and N3, by following guidelines stated in the American Academy of Sleep Medicine (AASM) Manual for the Scoring of Sleep and Associated Events [17]. This manual annotation

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Table 1: Datasets

	Number of Subjects	Sampling Frequency (Hz)	Channel Names	Annotation Schema
ISRUC [33]	100	200	F3-A2, C3-A2, F4-A1, C4-A1, O1-A2, O2-A1, ROC-A1, LOC-A2, Chin-EMG	AASM [17]
PhysioNet-EDFX [34, 35]	197	100	EEG Fpz-Cz, EEG Pz-Oz, EOG horizontal, EMG submental	R&K [36, 37]

scheme is time-consuming and expensive because a clinician needs several hours to annotate a patient's recordings from a single night [18].

There has been considerable research in automating sleep staging to overcome this. These approaches have mostly remained in the realm of deep learning that lack interpretability [19], for example convolutional neural networks (CNN) [20, 21, 22, 23], recurrent neural networks (RNN) [24, 25], recurrent convolutional neural networks (RCNN) [26, 27], graph convolutional networks (GCN) [28, 29], and attention [30, 31, 28]. On the other hand, the AASM sleep scoring manual guidelines [17], are interpretable for clinicians without providing the clear definitions needed to design a robust computational model [32].

A recent study proposed a feature-based linear model [12] that performed as well as deep neural network models. However, the features used in this study were not designed considering clinical guidelines. In order to generate clinically relevant explanations, we propose `NormIntSleep`, a representation learning framework that projects deep neural network embeddings into an interpretable normalized feature space. Thus, by choosing an appropriate feature space, `NormIntSleep` can unite clinically relevant explanations with the high accuracy of a deep learning model.

2 Method

Two publicly available datasets are used for evaluation. These datasets are summarized in Table 1.

`NormIntSleep` uses the PSG recordings, $\boldsymbol{X}$, to generate an interpretable representation for deep neural network embeddings using the following steps:

1. A CNN-LSTM network [38] is trained end-to-end on sleep staging. The CNN is composed of 3 convolutional layers where each layer is followed by batch normalization, ReLU activation, and max pooling. The kernel sizes of the three layers are 201, 11, and 11, and the output channels are 256, 128, and 64. The CNN output is used as input for a layer of bi-directional Long Short-Term Memory (LSTM) cells with 256 hidden states. The resulting 512 hidden states represent the *embedding space*, $\boldsymbol{E}(\boldsymbol{X})$. During model training, the LSTM output is connected to a fully connected layer with 5 outputs for the 5 sleep stages, and cross-entropy loss is used.
2. Features, $\boldsymbol{F}(\boldsymbol{X})$, are extracted from the dataset. There are two possible sets of features:
   - **FeatLong**: an exhaustive list of features inspired by the recent work of Van Der Donckt et al. [12]. The features are not designed considering clinical guidelines. It results in 2488 features for the ISRUC dataset and 1048 features for the Physionet dataset. 10% of the most significant features are retained for the next steps using ANOVA, resulting in 249 features for the ISRUC dataset and 105 for the Physionet dataset.
   - **FeatShort**: a smaller set of clinically interpretable features inspired by the recent work of Al-Hussaini et al. [38]. The features are designed according to AASM [39]. 87 features are extracted for the ISRUC dataset and 38 for the Physionet dataset. 90% of the most significant features are retained for the next steps using ANOVA, resulting in 78 features for the ISRUC dataset and 34 for the Physionet dataset.
3. A linear transformation, $\boldsymbol{T}$, is learned from the embedding space to the feature space, $\boldsymbol{E}(\boldsymbol{X}) \xrightarrow{\boldsymbol{T}} \boldsymbol{F}(\boldsymbol{X})$, using linear least squares regression with L_2 regularization. $\boldsymbol{E}(\boldsymbol{X})\boldsymbol{T}$ is then used to obtain a representation of the embedding in the interpretable feature space called $\boldsymbol{R}$. These representations of the embeddings are normalized before being used as input to classifiers.

$$\boldsymbol{R'} = \frac{\boldsymbol{R} - \mu_R}{\sigma_R}$$

where $\mu_R = \frac{\sum R}{N}$ and $\sigma_R = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(R_i - \mu_R)^2}$.

These normalized interpretable representations of the embeddings are used as inputs to simple classifiers such as logistic regression.

3 Experiments

3.1 Experimental setup

The CNN-LSTM was trained using PyTorch 1.0 [40] using a batch size of 1000 samples from 1 PSG. The training was continued for 20 epochs with a learning rate of 10^{-4} using ADAM [41] as the optimization method. The simple models were trained using scikit-learn [42, 43], XGBoost [44], and CatBoost [45, 46]. The features were extracted using scikit-learn [42, 43], yasa [47], and tsflex [48]. 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. The training set is used for each dataset 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.

3.2 Baselines

- CNN-LSTM: the model used to generate the embeddings
- FeatLong and FeatShort: used as inputs to simple classifiers - CatBoost [45, 46], XGBoost [44], Logistic Regression, Gradient Boosted Trees
- SLEEPER [32]: a prototype based interpretable sleep staging algorithm
- SERF [38]: an interpretable sleep staging algorithm based on embeddings, rules, and features
- U-Time [21]: state-of-the-art deep learning model for sleep staging

3.3 Results

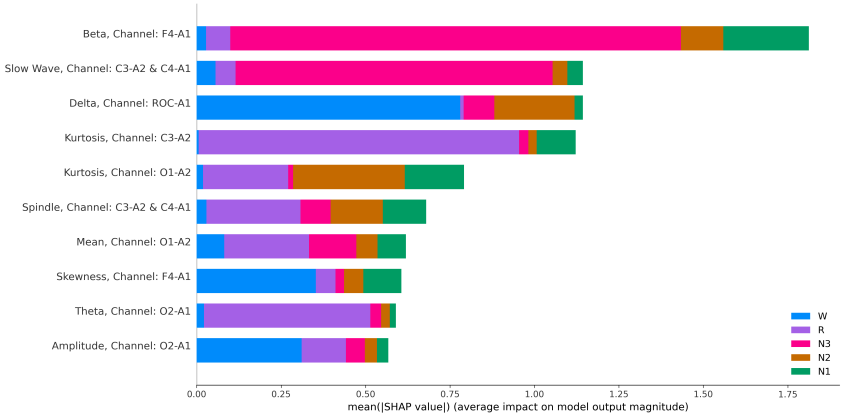
Table 2: Model Evaluation

Model	Accuracy		Cohen's κ		F1 Score (Macro)		Average Performance
	Physionet	ISRUC	Physionet	ISRUC	Physionet	ISRUC	
`NormIntSleep`-FeatLong-XGBoost	0.847	0.810	0.798	0.785	0.789	0.754	0.797
FeatLong-XGBoost [12]	0.836	0.750	0.777	0.712	0.773	0.675	0.754
`NormIntSleep`-FeatLong-CatBoost	0.846	0.815	0.793	0.788	0.787	0.760	0.798
FeatLong-CatBoost [12]	0.834	0.743	0.766	0.701	0.768	0.666	0.746
`NormIntSleep`-FeatLong-Logistic Regression	0.855	0.807	0.801	0.783	0.800	0.748	0.799
FeatLong-Logistic Regression [12]	0.809	0.760	0.736	0.717	0.733	0.689	0.741
`NormIntSleep`-FeatShort-XGBoost	0.832	0.809	0.772	0.781	0.768	0.753	0.786
FeatShort-XGBoost	0.825	0.790	0.767	0.747	0.759	0.726	0.769
`NormIntSleep`-FeatShort-CatBoost	0.831	0.809	0.770	0.777	0.766	0.752	0.784
FeatShort-CatBoost	0.818	0.788	0.750	0.741	0.747	0.724	0.761
`NormIntSleep`-FeatShort-Logistic Regression	0.855	0.797	0.799	0.774	0.800	0.735	0.793
SERF-XGBoost [38]	0.823	0.819	0.753	0.789	0.753	0.766	0.784
SERF-Logistic Regression [38]	0.829	0.795	0.759	0.773	0.762	0.733	0.775
SLEEPER-Gradient Boosted Trees [32]	0.807	0.797	0.721	0.756	0.729	0.736	0.758
U-Time [21]	0.862	0.840	0.811	0.816	0.810	0.793	**0.822**
CNN-LSTM [38]	0.864	0.831	0.815	0.819	0.813	0.783	0.821

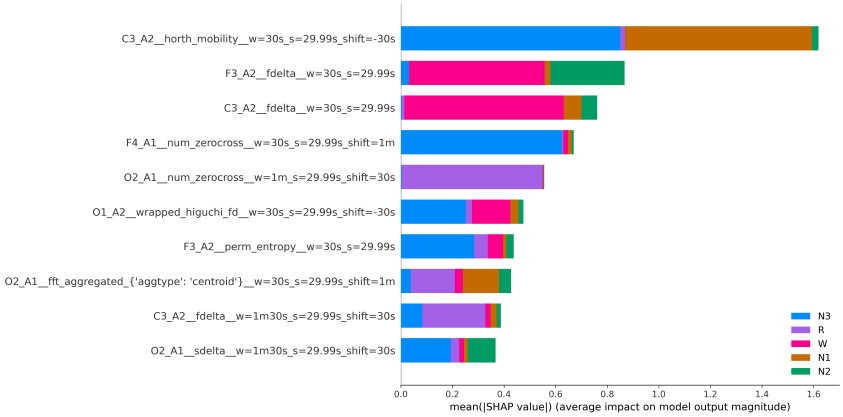
Accuracy, Macro F1-Score, and Cohen's κ are used to evaluate the models. The results in Table 2 were produced in the same experimental setup expanded upon in Section 3.1. Since the performance of models varies a lot between the two datasets, an aggregated metric, *Average Performance* is calculated based on the average value of each model across the two datasets and three metrics. The best interpretable methods using FeatShort and FeatLong are underlined, and the best overall method is highlighted in bold. It shows that `NormIntSleep` surpasses all other interpretable methods even when using the smaller set of interpretable features, FeatShort. The benefits of using this over the exhaustive set of features, FeatLong, are discussed in Section 3.4.

3.4 Interpretation and Clinical Relevance

Figure 1 highlights the top 10 important dimensions of the interpretable representation of the `NormIntSleep` embeddings and their influence on the classification of the 5 sleep stages. N1 is a



(a) `NormIntSleep`-XGBoost-FeatShort



(b) `NormIntSleep`-XGBoost-FeatLong

Figure 1: SHAP embedding representation importance using `NormIntSleep` on ISRUC Dataset

challenging sleep stage to classify, with around 50% agreement among human annotators, so its classification is not discussed in detail. The differences between Figure 1a and Figure 1b shows the importance of having clinically relevant representations. Looking at the top 3 representations in `NormIntSleep`-XGBoost-FeatShort (Figure 1a): (1) Beta waves [38] are defined as waves with frequency between 8 Hz and 20 Hz. This frequency range is present in Wake [49], REM [49], and N2 (through sigma band in Spindles [50]). So, it is the perfect attribute to differentiate N3 from the other stages. (2) Slow Waves are used to annotate N3 by clinicians [49, 39]. (3) Deflections in EOG indicate blinks during wakefulness, so delta waves in ROC-A1 (an EOG channel) are appropriate for classifying wake. On the other hand, 6 of the top 10 representations in `NormIntSleep`-XGBoost-FeatLong (Figure 1b) are not clinically relevant and and the rest are derived from the same frequency band. Thus, the clinical relevance of `NormIntSleep`-XGBoost-FeatLong's interpretations is dubious.

The feature space of the two datasets using FeatLong and FeatShort are shown in Appendix A and highlights the efficacy of FeatLong in segregating the Physionet dataset into the 5 classes. FeatLong and FeatShort feature importances are shown in Appendix B.

4 Conclusion

Interpretability is crucial for the adoption of clinical decision support systems. Complex features paired with a simple model can provide interpretation. However, those explanations might not be helpful if the features are not clinically meaningful. `NormIntSleep` provides a generic framework to leverage the most clinically significant features for classification with higher accuracy than complex feature-based models and nearly as accurate as black-box deep learning. As a result, the interpretations are clinically relevant. Thus `NormIntSleep` takes forward strides towards adoption.

References

[1] Chu Jianxun, Vincent Ekow Arkorful, and Zhao Shuliang. Electronic health records adoption: Do institutional pressures and organizational culture matter? *Technology in Society*, 65:101531, 2021.

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

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

[4] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

[5] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

[6] Beau Norgeot, Benjamin S Glicksberg, and Atul J Butte. A call for deep-learning healthcare. *Nature medicine*, 25(1):14–15, 2019.

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

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

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

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

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

[12] Jeroen Van Der Donckt, Jonas Van Der Donckt, Emiel Deprost, Michael Rademaker, Gilles Vandewiele, and Sofie Van Hoecke. Do not sleep on linear models: Simple and interpretable techniques outperform deep learning for sleep scoring. *arXiv preprint arXiv:2207.07753*, 2022.

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

[14] Sarah Holder and Navjot S Narula. Common sleep disorders in adults: Diagnosis and management. *American Family Physician*, 105(4):397–405, 2022.

[15] 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. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature Communications*, 9(1):5229–5229, 2018.

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

[17] 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. 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):597–619, 2012.

[18] Hanrui Zhang, Xueqing Wang, Hongyang Li, Soham Mehendale, and Yuanfang Guan. Auto-annotating sleep stages based on polysomnographic data. *Patterns*, 3(1):100371, 2022.

[19] Zachary Chase Lipton. The mythos of model interpretability. *CoRR*, abs/1606.03490, 2016.

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

[21] Mathias Perslev, Michael Jensen, Sune Darkner, Poul Jø rgen Jennum, and Christian Igel. U-time: A fully convolutional network for time series segmentation applied to sleep staging. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 4415–4426. Curran Associates, Inc., 2019.

[22] Stanislas Chambon, Mathieu N. Galtier, Pierrick J. Arnal, Gilles Wainrib, and Alexandre Gramfort. 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):758–769, 2018.

[23] Bufang Yang, Xilin Zhu, Yitian Liu, and Hongxing Liu. 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*, 68:102581, 2021.

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

[25] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. 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):400–410, 2019.

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

[27] A. Supratak, H. Dong, C. Wu, and Y. Guo. Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017.

[28] Menglei Li, Hongbo Chen, and Zixue Cheng. An attention-guided spatiotemporal graph convolutional network for sleep stage classification. *Life*, 12(5):622, 2022.

[29] Ziyu Jia, Youfang Lin, Jing Wang, Ronghao Zhou, Xiaojun Ning, Yuanlai He, and Yaoshuai Zhao. Graphsleepnet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification. In *IJCAI*, pages 1324–1330, 2020.

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

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

[32] Irfan Al-Hussaini, Cao Xiao, M Brandon Westover, and Jimeng Sun. Sleeper: interpretable sleep staging via prototypes from expert rules. In *Machine Learning for Healthcare Conference*, pages 721–739. PMLR, 2019.

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

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

[35] 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. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[36] SLEEP COMPUTING COMMITTEE OF THE JAPANESE SOCIETY OF SLEEP RESEARCH SOCIETY (JSSR):, Tadao Hori, Yoshio Sugita, Einosuke Koga, Shuichiro Shirakawa, Katuhiro Inoue, Sunao Uchida, Hiroo Kuwahara, Masako Kousaka, Toshinori Kobayashi, et al. Proposed supplements and amendments to 'a manual of standardized terminology, techniques and scoring system for sleep stages of human subjects', the rechtschaffen & kales (1968) standard. *Psychiatry and clinical neurosciences*, 55(3):305–310, 2001.

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

[38] Irfan Al-Hussaini and Cassie S Mitchell. Serf: Interpretable sleep staging using embeddings, rules, and features. *arXiv preprint arXiv:2209.11174*, 2022.

[39] 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. 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):597–619, 2012.

[40] 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. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[41] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[42] 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. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[43] Sebastian Raschka, Joshua Patterson, and Corey Nolet. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *arXiv preprint arXiv:2002.04803*, 2020.

[44] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[45] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018.

[46] Anna Veronika Dorogush, Andrey Gulin, Gleb Gusev, Nikita Kazeev, Liudmila Ostroumova Prokhorenkova, and Aleksandr Vorobev. Fighting biases with dynamic boosting. *arXiv preprint arXiv:1706.09516*, 2017.

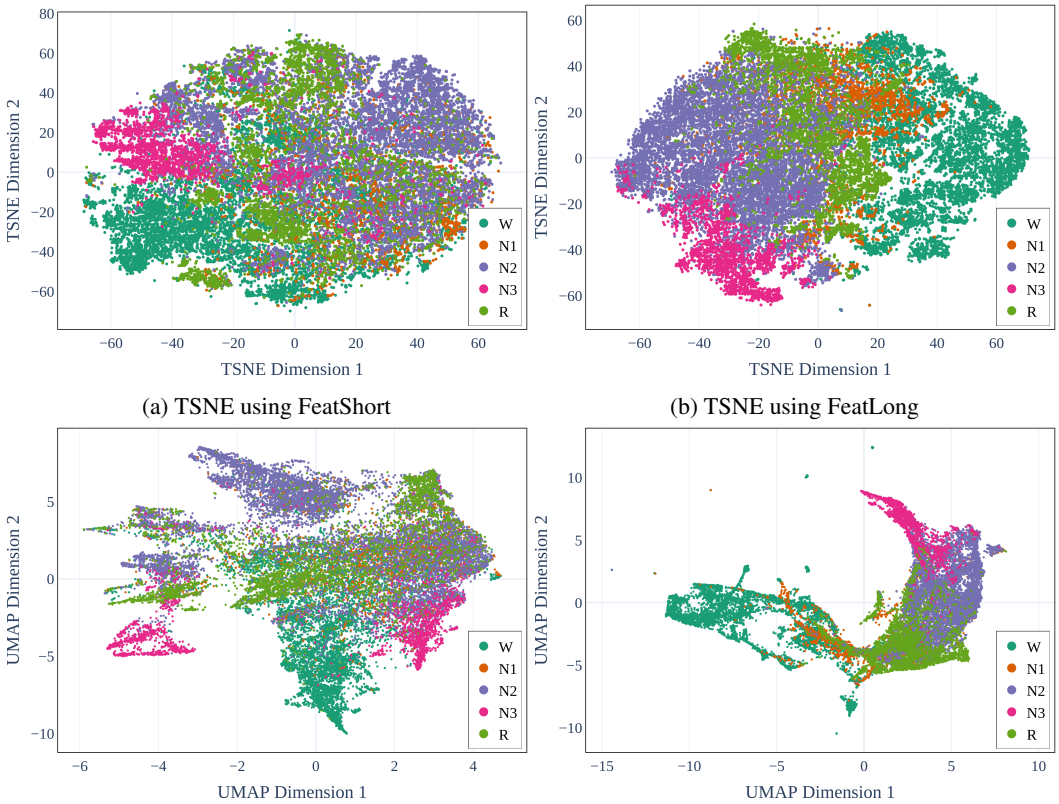
[47] Raphael Vallat and Matthew P Walker. An open-source, high-performance tool for automated sleep staging. *eLife*, 10:e70092, oct 2021.

[48] Jonas Van Der Donckt, Jeroen Van Der Donckt, Emiel Deprost, and Sofie Van Hoecke. tsflex: flexible time series processing & feature extraction. *SoftwareX*, 2021.

[49] Aakash K Patel, Vamsi Reddy, and John F Araujo. Physiology, sleep stages. In *StatPearls [Internet]*. StatPearls Publishing, 2021.

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

A Appendix A: Feature Space Visualization



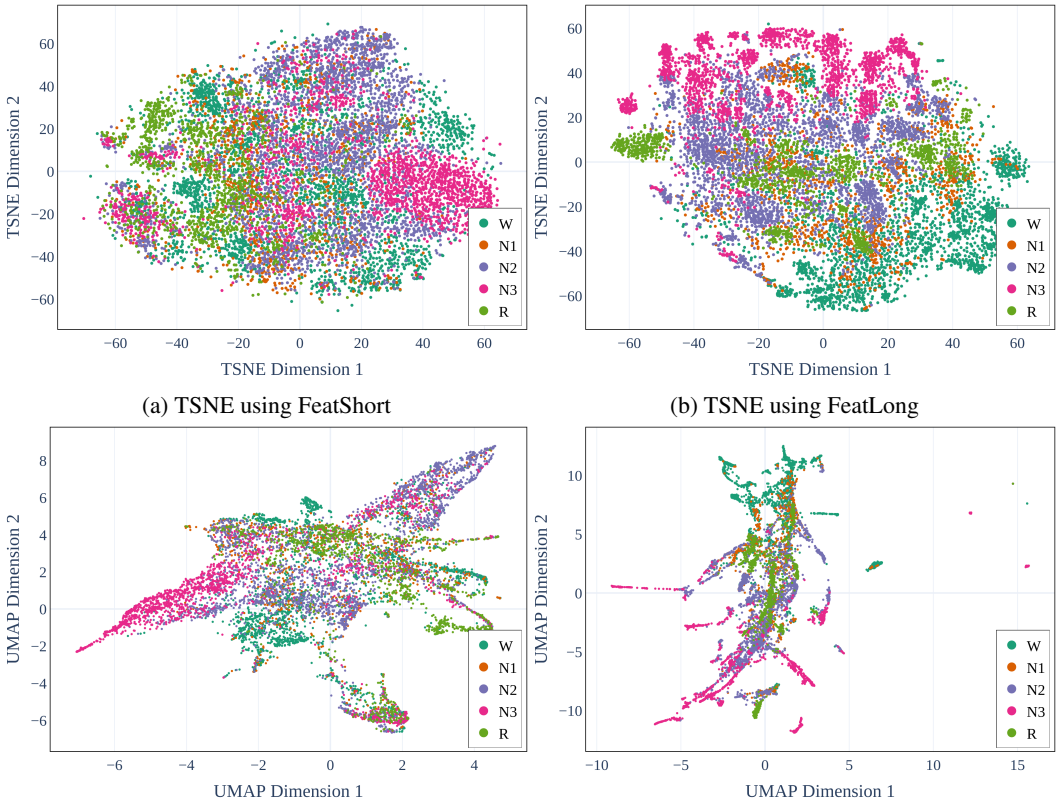
(a) TSNE using FeatShort

(b) TSNE using FeatLong

(c) UMAP using FeatShort

(d) UMAP using FeatLong

Figure 2: Dimensionality reduction on the physionet dataset: shows distinct clusters for classes using FeatLong



(a) TSNE using FeatShort

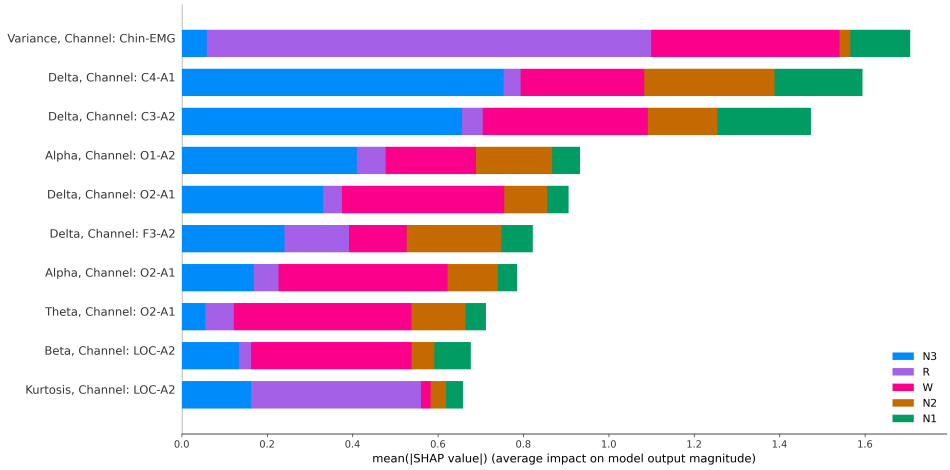
(b) TSNE using FeatLong

(c) UMAP using FeatShort

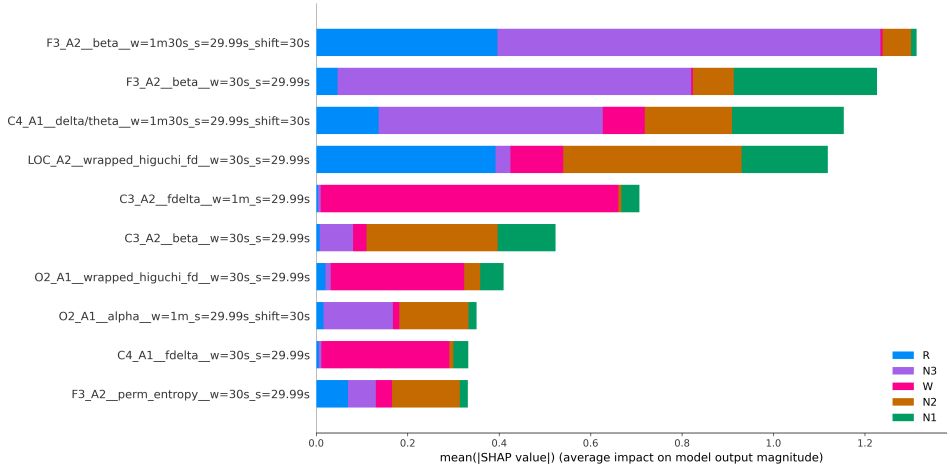
(d) UMAP using FeatLong

Figure 3: Dimensionality reduction on the ISRUC dataset

B Appendix B. SHAP Feature Importance



(a) XGBoost-FeatShort



(b) XGBoost-FeatLong

Figure 4: SHAP feature importance using feature-based models on ISRUC Dataset