Building neural networks' latent space to extract instance-based explanations for sleep staging

Guido Gagliardi*,[1,2,3], Antonio Luca Alfeo[1,4],
Mario G.C.A. Cimino[1,4], Gaetano Valenza[1,4] and Maarten De Vos[2]

Abstract— **Sleep disorders and their diagnosis are a significant public health concern. Automated sleep stage classification using deep learning models has shown promising results, but these models often lack transparency and interpretability. In this study, we propose an eXplainable Artificial Intelligence (XAI) approach to enhance the interpretability of cutting-edge deep learning sleep stage classification models. The proposed approach consists of a three-steps framework: (i) employing contrastive learning to order a neural network latent space based on input similarity; (ii) mining meaningful instances from that space; and (iii) explaining those instances by a customized XAI methodology. By doing this we are capable of extracting human-comprehensible insights about the model decision-making process, enhancing the applicability of the proposed approach in real-world clinical scenarios. The explanations provided point out high and low-representative sleep epochs of each sleep phase. These sleep epochs are analyzed considering both the single sleep epoch and the sequence of adjacent sleep epochs for the sleep phase classification.**

Our approach proved to maintain the original model performances, improve the model interpretability, and confirm that the network decision-making process is valid even from the perspective of a physician.

I. Introduction

Sleep deprivation and disorders are widespread issues, impacting millions of individuals globally and presenting significant public health concerns [1]. As a result, there is a growing need for precise evaluation, diagnosis, and continuous monitoring of sleep patterns [2]. To tackle these challenges, automated sleep scoring (e.g. via polysomnography) is crucial since providing a manual scoring can be very time-consuming and becomes impractical for handling extensive sleep data.

Polysomnography (PSG) allows analyzing the sleep stages and cycle by monitoring subjects' brain activity, eye movements, muscle activity, heart rate, and respiration. This results in the collection of several physiological signals such as electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), and electrocardiography (ECG). These signals need to be preprocessed and visually inspected to identify patterns and characteristics that are peculiar to specific sleep stages. For example, REM sleep is associated with desynchronized EEG patterns and rapid eye movements. These characteristics are collected and organized in manuals [3] that outline the guidelines for evaluating various parameters, such as EEG waveforms, eye movements, muscle tone, and respiratory patterns, to differentiate between different sleep stages and identify sleep-related events.

Overall, it can take approximately two hours for a sleep expert to provide a scoring for overnight polysomnography (PSG) recording [4]. In contrast to manual examinations, the approaches based on deep learning can complete the sleep stage recognition in a few seconds.

Over time, well-known deep learning architectures like Auto-encoders [5], Convolutional Neural Networks (CNNs) [6], and Recurrent Neural Networks (RNNs) [7] have been outperformed by more complex and task-specific architectures. Examples of these newer architectures include DNN+RNN [8], CNN+RNN [9], and hierarchical RNN [10], [11].

Despite the great recognition performance, complex deep neural networks work as black boxes, and this prevents domain experts from validating and trusting their outcomes [12]. This is especially important in real-world scenarios [13]. For instance, in medical contexts, only if the physician can validate and trust the outcome of the algorithm, this outcome can be employed in medical decision-making processes [14], [15].

To address this issue, eXplainable Artificial Intelligence (XAI) approaches can be employed. XAI approaches compound AI recognition capability with some explanation for the AI model's decision-making process [16]. There are different approaches of explaining an AI model and making it accessible to a particular audience. According to the specific application and its end users, a choice has to be made between comprehensible and faithful explanations. According to [17], *comprehensible* explanations can be presented in simple and not ambiguous form, whereas *faithful* explanations can describes the AI model comprehensively and correctly. Comprehensible explanations are more suitable for applications in which the end users are domain experts and decision-makers with no AI background, such as in the medical context, where medical practitioners often don't have AI background [**?**].

[1] Department of Information Engineering, University of Pisa, Pisa, Italy.

[2] Dept. of Electrical Engineering, KU Leuven, Belgium

[3] Dept. of Information Engineering, University of Florence, Italy

[4] Bioengineering & Robotics Research Center E. Piaggio, School of Engineering, University of Pisa, Pisa, Italy.

∗ Correspondence to: `guido.gagliardi@phd.unipi.it`

Financial support was provided by the Research Foundation Flanders (FWO) via SBO mandate 1SH4Z24N. Work partially supported by (i) the EU Commission - Horizon 2020 Program under GA 101017727 of the project "EXPERIENCE", (ii) the Italian Ministry of Education and Research (MIUR) in the framework of the FoReLab project (Departments of Excellence), and (iii) PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme.

Instance-based explanation are among the XAI approaches that provide the most comprehensible explanations [18]. In brief, instance-based explanations are data instances that are *meaningful* according to the AI model, i.e. class prototypes and counterexamples [19]. By being real-word cases, these instances can be analyzed by using just the domain knowledge of the physician, who reasons about similar instances on a daily basis.

For instance, a domain expert knows which time series behaviors are peculiar to the NREM2 sleep phase, e.g., k-complexes and sleep spindles in the EEG data. Via instance-based explanations, the algorithm's reasoning can be validated by examining the time series most exemplary of the NREM2 phase (i.e. the prototypes) according to the algorithm. This allows understanding of the peculiar characteristics [3] of this sleep phase are considered by the algorithm since those are present in the prototype for that class. Moreover by inspecting instances on the boundary between two sleep phases (i.e. counterexamples) we can exploit which specific features of the time series imply the network output to change between two sleep phases and hence validate if the knowledge exploited by the AI model matches the medical expertise.

Meaningful instances [20] can be retrieved by exploring the latent space of a DNN [21], [22]. However, given the complexity of DNN, this search is far from trivial. For instance, the information about a sleep stage may be distributed among the nodes of the whole DNN rather than represented in a specific portion of it [23]. To address this issue, contrastive learning can be employed to constrain the network during training to force the organization of its latent space to represent sleep stage information [24], [25].

Via contrastive learning, similar instances (e.g. epochs characterized by the same sleep stage) are projected in the latent space as closer to each other and separated from different instances. By doing so, the model learns to capture meaningful patterns and similarities in the data, and such information is readily available and usable for finding meaningful instances for the model, without sacrificing recognition performance.

Our aim is to seamlessly introduce interpretability without altering the primary function of a state of the art neural network, which is accurate sleep stage classification; second, in combination with contrastive learning we employed categorical cross-entropy, to ensure that the network remains structured as a standard classification network. This structural compatibility facilitates the application of other explainability algorithms, such as Integrated Gradients, for the extraction of saliency maps. These maps are instrumental in elucidating the factors contributing to sleep stage classification decisions.

In this research work we propose a novel machine learning framework to explain current state-of-the-art models in sleep staging (e.g. sequence-to-sequence models such as XSleepNet). The framework consists of four steps (i) adding a normalization layer inside the neural network to evaluate the network latent space; (ii) order such latent space using a combination of a contrasting learning loss and a classification loss during training; (iii) mine meaningful instances from the ordered latent-space; and (iv) analyze such meaningful instances by using a XAI attribution methodology such as Integrated Gradients.

This framework proved to maintain the original model performances, while remarkably improving its interpretability. Moreover, validating the mined meaningful instances confirmed that the network decision-making strategy resemble the expert reasoning.

II. Materials and Methods

In this section, we present a comprehensive description of our method for providing interpretability to neural networks used in sleep phase analysis. Our approach leverages the XSleepNet architecture as a starting point and introduces a novel framework for mining its latent space organized via contrastive learning. The goal is to extract instance-based explanations to understand if the DNN decision-making process aligns with the reasoning of expert medical practitioners.

A. Experimental Data

In this study we used the publicly available[1] Sleep-EDF-20 dataset. This is the Sleep Cassette (SC) subset of the Sleep-EDF Expanded dataset (2013 version), consisting of 20 subjects (10 males and 10 females) aged 25-34 years. For each subject, two consecutive day-night PSG recordings were collected. Each 30-second PSG epoch was manually labeled into one of eight categories W, N1, N2, N3, N4, REM, MOVEMENT, UNKNOWN by sleep experts according to the R&K standard. Stages N3 and N4 were considered collectively as N3, and the categories MOVEMENT and UNKNOWN were excluded.

B. XSleepNet Architecture Modification

We enhanced the state-of-the-art XSleepNet architecture to provide interpretability without losing recognition performances.

XSleepNet [26] is a state-of-the-art sequence-to-sequence DNN architecture primarily designed to analyze sleep stage data. In the sequence-to-sequence paradigm, the network processes sequences of L-sleep epochs (each sleep epoch conists of 30 seconds EEG window both as time-series and time-frequency image), to capture the temporal dependency between different sleep epochs. The network consists in 2 different encoders: an epoch-encoder which process a single epoch and a sequence-encoder which process epoch-encodings of L consecutive sleep epochs into L sequence-encodings. Then each one of the L-sequence encoding its used to predict L different classes which correspond to each epoch.

To enhance the network's interpretability while preserving its overall performance, we introduce a key modification. Specifically, we add a normalization layer immediately before the classification layer. Hence we slightly modify the sequence encoder to output a normalized version of the

[1] https://www.physionet.org/content/sleep-edfx/1.0.0/

sequence encoding. This modification is intended to ensure that the latency space, evaluated at the normalization layer, is suited to compute distances-based losses, such as contrastive learning losses, in order to organize the DNN latent space based on the similarities of the input instances exploited by the network during classification.

C. Loss Function

Our loss function comprises two fundamental components: the contrastive loss (L_{CON}) and the categorical cross-entropy loss (L_{CE}).

$$L = L_{CON} + L_{CE}$$

The incorporation of L_{CE} serves multiple critical purposes within our methodology: first it is introduced to maintain the structural and functional integrity of the neural network, ensuring that the end-user experiences no perceivable difference compared to using the original XSleepNet architecture.

L_{CON} is introduced to impose an organization to the latent space evaluated at the normalization layer. The goal of the contrastive loss is to encourage the network to map similar input instances from the same class close to each other in the latent space while pushing instances from different classes apart. This structured representation facilitate subsequent explanation extraction.

To calculate the contrastive loss, we need to perform (hard) triplet mining, which involves selecting suitable (i.e. from different sleep phases) triplets for loss computation. Given that our batch consists of data with dimensions $batch\_size \times L \times encoding\_length$, performing triplet mining directly on $batch\_size \times L$ data is computationally intensive. To mitigate this computational burden, we randomly sample $batch\_size$ elements from $batch\_size \times L$. This enables us to calculate the contrastive loss on a smaller subset of the data without compromising the effectiveness of the loss term.

By introducing our custom contrastive loss and adopting the proposed efficient triplet mining techniques we aim to create an ordered latent space. This is convenient for the extraction of instance-based explanations while ensuring that the network remains structurally and functionally similar to XSleepNet, allowing the application of explainability algorithms like Integrated Gradients.

D. Instance-Based Explanation Extraction

In this section, we detail our approach to mining the ordered latent space of the DNN.

The purpose of such a mining procedure is to validate the overall network decision-making process, i.e. checking if the network is able to characterize each sleep phase and distinguish between different sleep phases like a physician would do. To this aim, both very-representative and low-representative instances of each sleep phase can be mined and validated via medical domain knowledge.

To identify those instances we analyzed the clusters relative to each sleep phase in the latent space of the DNN. These clusters are the results of the application of our contrastive learning loss. Thus, instances belonging to the same sleep phase should be projected as close to each other, whereas instances belonging to different sleep phases should be projected far apart in the latent space. The centroid of the clusters are considered as very-representative instances of the sleep phase to which the clusters are linked (i.e. most of the element of the cluster belong to that sleep phase) and instances on the boundary of two clusters as low-representative of such a sleep phase.

To mine instances at the center of the clusters of each sleep phase we used the KMedoids algorithm. The value of K has been determined maximising either the silhouette score or the elbow method, allowing us to discover the optimal number of clusters. Each cluster represents a different sleep phase, and the KMedioids [27] method identifies instances at the center of each cluster.

To mine instances at the boundary of the clusters, we used the Support Vector Machine algorithm. Support Vector Machines (SVM) [28] offer an effective approach for identifying instances at the edge of clusters, where the classification model exhibits indecision. SVM is particularly well-suited for this task due to its ability to create decision boundaries and support vectors. SVM effectively identifies support vectors, which are instances close to the decision boundary. The instances at the edge of clusters can be considered as high uncertainty points, as a small variation in their position in the latent space would change their classification outcome.

E. Explaining the Instance-Based Explanations

In this section, we detail the methodology for explaining the instance-based explanations identified via the procedure detailed in the last section. We employ the Integrated Gradients (IG) [29] algorithm to understand the importance of individual input features in the decision-making process of a neural network. IG provides a way to measure the influence of each feature on the model's output. IG works by integrating the gradients of the model's output with respect to input features along a path from a baseline to the actual input. The formula for Integrated Gradients is as follows:

$$IG_i(x) = (x_i - x_i') \cdot \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \cdot (x - x'))}{\partial x_i} \, d\alpha$$

Where i is the feature index, x is the actual input instance, x' is the chosen baseline input instance, F is the output of the model and ∇F represents the gradient of the model's output.

In our methodology, we employ two different baseline strategies for central and boundary instances: for central instances, which are representative of specific sleep phases, we employ a standard baseline of 0. This baseline allows us to analyze the absolute influence of each input feature on the classification, providing insights into the essential contributors to a particular sleep phase classification. For instances residing at the boundary between two sleep phases, we use the adjacent instance of the other sleep phase on the boundary of the neighboring sleep phase as the baseline. This choice enables us to understand the specific contributions that characterize the transition from one sleep phase to another.

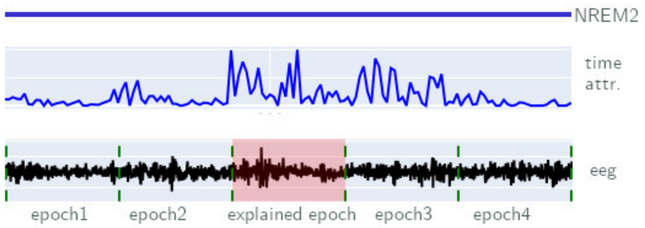

Fig. 2: Temporal attribution maps for an instance based explanation located at the center of the NREM2 sleep phase cluster. The attributions maps have been provided also considering variations in the gradients over the entire sleep sequence. The center epoch is located in sequences with no transition between one sleep phase and another. Hence the center of the cluster of the sleep phase stays in the middle of the sleep phase.

The attributions obtained through Integrated Gradients are then grouped in the time and frequency domains. This aggregation enables us to understand the contributions of specific temporal patterns (time) and frequency components (frequency) that drive the model's classification decisions. Analyzing the contributions in these domains offers valuable insights into the features that influence sleep phase predictions at different scales.

III. Results and Discussion

TABLE I: Performance comparison between our approach and the state of the art architecture XSleepNet

Model	Accuracy	Sensitivity	Specificity
XSleepNet	83.9%	78.6%	95.5%
Our Appr.	81.5%	81.5%	95.4%

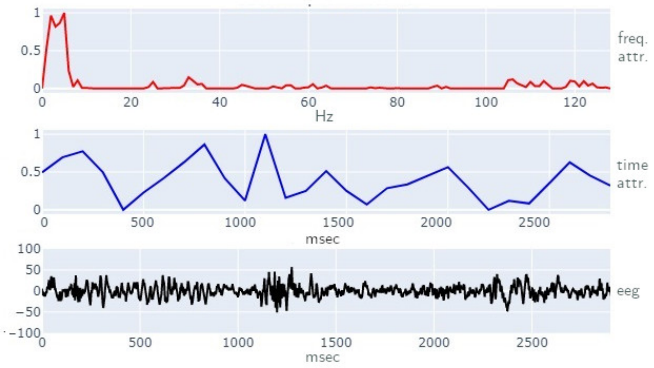

Fig. 1: Attribution maps (time-frequency) for an instance based explanation located at the center of the NREM2 sleep phase cluster. Frequency attributions identifies θ waves $(4-8Hz)$ and time attributions the presence around sample 2400 of a K-Complex.

In this section, we present and discuss the results obtained with our approach. We cover (i) the recognition performance of our custom XSleepNet model, (ii) its ability to organize its latent space, and (iii) the validation of the decision-making process of the DNN by employing the medical domain knowledge. The reported results were obtained using a leave-one-subject-out validation schema.

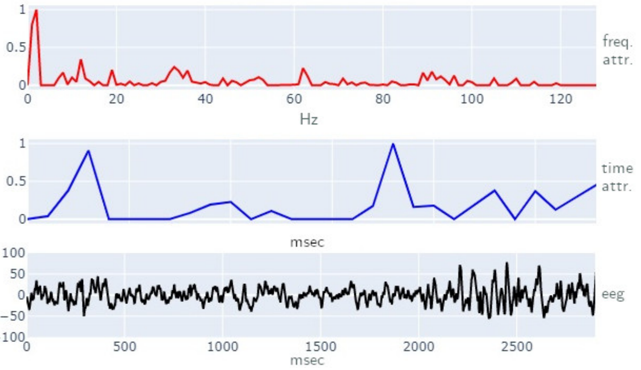

Fig. 3: Attribution maps (time-frequency) for an instance based explanation located at the boundary of the NREM2 sleep phase cluster with the NREM3 sleep phase cluster. The instance has been classified correctly as a NREM2. Frequency attributions identifies δ waves $(0.5-3Hz)$ which are typical of NREM3, hence the high level of uncertainty of the network.

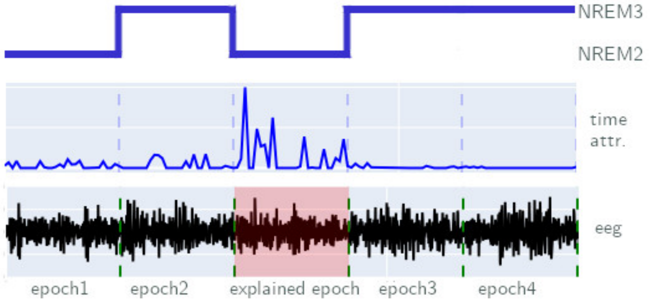

Fig. 4: Temporal attribution maps for an instance based explanation located at the bounday of the NREM2 sleep phase cluster with the NREM3 sleep phase cluster. The attributions maps have been provided also considering variations in the gradients over the entire sleep sequence. The element at the edge of the cluster stay on a temporal transition between sleep phase NREM2 and NREM3. The correct prediction (NREM2) is hence done considering only elements of the actual epoch and not leveraging the sequence information (attributions flatted).

Our primary objective is to introduce interpretability without significantly compromising the performance of the base XSleepNet model. As indicated in Table I, our approach maintains competitive performance across various evaluation metrics. Although there is a minor decrease in accuracy and F1 score, it remains within an comparable range.

The key strength of our approach lies in its ability to organize the latent space of the DNN, to provide some insight into the decision-making process of the network. To this end, we determine the number of centroids individuated by the KMedoids algorithm through the elbow method and silhouette score of clusters generated with different values of *K*. Our results showed that increasing *K* significantly reduced both silhouette and elbow values, suggesting that points are well-distinguished between different sleep phases, forming a single cluster when considering each individual sleep phase, hence we identified $K = 4$ clusters corresponding to each sleep phase (WAKE instances were removed from the analysis to reduce the noise in the latent space).

Subsequently, we characterized the points at the center of each cluster and those at the boundaries between different phases. The central points are highly representative of their respective sleep phases, while the boundary points represent instances where the network exhibits indecision.

We conducted an empirical evaluation [30] using Integrated Gradients (IG) with a focus on two aspects: (i) characterizing the network's decision based on the sequence of L-sleep epochs that the network takes as input and (ii) focusing on the single sleep epoch to which the centroid refers.

Thanks to our approach we were able to understand that the network placed elements in the middle of a sleep phase in the center of the cluster referred to as that sleep phase; and placed sleep epochs near a transition between two different sleep phases on the boundary of the clusters related to those sleep phases, aligning with the expectations of sleep experts.

Moreover while considering the single sleep epochs by themselves, it is possible to understand how elements on the center of the clusters represent "standard" cases of such sleep phase, clearly covering the domain knowledge. On the other hand input elements on the boundary between two clusters (i.e. two different sleep phases) share features of both the sleep phases, and sometimes represent sleep epochs during which a transition from two sleep phases is occurring (e.g, Fig. 3 around 20 sec there is a transition from NREM2 to NREM3).

For instance, Figures 2 and 4 show the difference in the sequence of sleep epochs between an element in the center of the NREM2 cluster, Fig. 2, and an element near to the boundary between NREM2 and NREM3, Fig. 3. As the figures show the element at the center of the cluster is located in the middle of its sleep phase and the element on the boundary is located on a transition between the two sleep phases.

Figures 1 and 3, provide examples of explanations extracted considering only the actual sleep epoch from the same aforementioned points at the center of NREM2 and at the boundary between NREM2 and NREM3, not the entire sequence. In this case, the network clearly resembles an expert-like reasoning by highlighting the typical patterns of the NREM2 stage, Fig 1. In Fig 3 instead its possible to see how the network correctly determined the start and the end of the sleep phase NREM2 considering the time attributions, but focusing more on θ waves (freq. attributions) which are typical of the NREM3 sleep phase, hence the high level of uncertainty for such sleep epoch.

In summary, our approach balances interpretability with model performance, effectively organizing the latent space and matching the decision-making process of a sleep expert. The results prove the potential of the proposed approach for providing a transparent sleep phase analysis and provide meaningful insights into the network's reasoning.

IV. Conclusion

In this study, we presented an approach that provides interpretability to a cutting edge deep learning architecture for sleep stage classification. Sleep disorders and their diagnosis present critical public health concerns, thus need efficient and accurate sleep stage analysis. Deep learning techniques have demonstrated remarkable success in this field, yet their complex, black-box nature hinders their real-world adoption, particularly in critical applications like medical decision support systems. In this context the explanations of an AI model are not just useful for validating the model but also required by recent regulations on personal data processing (e.g., the General Data Protection Regulation [31]).

With respect to the state-of-the-art AI model for sleep staging, the proposed approach (i) proved to acquire similar or comparable recognition performances; (ii) is able to successfully organize its latent space to allow mining for meaningful instances to explain the model's decision process; and (iii) does align the decision-making process of the AI model to the medical domain knowledge.

Future works will take advantage of the high generalization capability of the proposed approach, which can easily be employed with different and new sleep staging AI architectures. This application will let us gain insights into the actual differences between such highly performative models.

On the other hand, it's worth noticing that for a large variety of medical tasks (e.g., emotion recognition), medical doctors do not perform the labeling directly on EEG instances, and hence extracting human-understandable instances could be troublesome. This could lower the potential applicability of this approach to many medical applications. Hence future works will focus on this limitation.

The ability to understand and validate the reasoning of AI models in sleep analysis not only enhances the trust in these models but also enables their adoption in real-world applications. For this reason, we believe that this approach can have a significant impact on the future of sleep disorder diagnosis, making it more accurate, efficient, and interpretable.

References

[1] V. K. Chattu, M. D. Manzar, S. Kumary, D. Burman, D. W. Spence, and S. R. Pandi-Perumal, "The global problem of insufficient sleep and its serious public health implications," in *Healthcare*, vol. 7, p. 1, MDPI, 2018.

[2] K. B. Mikkelsen, J. K. Ebajemito, M. A. Bonmati-Carrion, N. Santhi, V. L. Revell, G. Atzori, C. Della Monica, S. Debener, D.-J. Dijk, A. Sterr, *et al.*, "Machine-learning-derived sleep–wake staging from around-the-ear electroencephalogram outperforms manual scoring and actigraphy," *Journal of sleep research*, vol. 28, no. 2, p. e12786, 2019.

[3] R. B. Berry, C. E. Gamaldo, S. M. Harding, R. Brooks, R. M. Lloyd, B. V. Vaughn, and C. L. Marcus, "Aasm scoring manual version 2.2 updates: new chapters for scoring infant sleep staging and home sleep apnea testing," 2015.

[4] A. Malhotra, M. Younes, S. T. Kuna, R. Benca, C. A. Kushida, J. Walsh, A. Hanlon, B. Staley, A. I. Pack, and G. W. Pien, "Performance of an automated polysomnography scoring system versus computer-assisted manual scoring," *Sleep*, vol. 36, no. 4, pp. 573–582, 2013.

[5] O. Tsinalis, P. M. Matthews, and Y. Guo, "Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders," *Annals of biomedical engineering*, vol. 44, pp. 1587–1597, 2016.

[6] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 4, pp. 758–769, 2018.

[7] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Automatic sleep stage classification using single-channel eeg: Learning sequential features with attention-based recurrent neural networks," in *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pp. 1452–1455, IEEE, 2018.

[8] H. Dong, A. Supratak, W. Pan, C. Wu, P. M. Matthews, and Y. Guo, "Mixed neural network approach for temporal sleep stage classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 2, pp. 324–333, 2017.

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

[10] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. 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*, vol. 27, no. 3, pp. 400–410, 2019.

[11] G. Gagliardi, A. L. Alfeo, M. G. Cimino, G. Valenza, and M. De Vos, "Physioex: a new python library for explainable sleep staging through deep learning," *Physiological Measurement*, vol. 13, no. 2, p. 025006, 2025.

[12] A. L. Alfeo, A. G. Zippo, V. Catrambone, M. G. Cimino, N. Toschi, and G. Valenza, "From local counterfactuals to global feature importance: efficient, robust, and model-agnostic explanations for brain connectivity networks," *Computer Methods and Programs in Biomedicine*, p. 107550, 2023.

[13] A. L. Alfeo, M. G. Cimino, and G. Gagliardi, "Concept-wise granular computing for explainable artificial intelligence," *Granular Computing*, vol. 8, no. 4, pp. 827–838, 2023.

[14] J. M. Schoenborn and K.-D. Althoff, "Recent trends in xai: A broad overview on current approaches, methodologies and interactions.," in *ICCBR Workshops*, pp. 51–60, 2019.

[15] A. L. Alfeo, M. G. Cimino, and G. Gagliardi, "Matching the expert's knowledge via a counterfactual-based feature importance measure," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 71–86, Springer, 2023.

[16] G. Gagliardi, A. L. Alfeo, V. Catrambone, M. G. Cimino, M. De Vos, and G. Valenzal, "Fine-grained emotion recognition using brain-heart interplay measurements and explainable convolutional neural networks," in *2023 11th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 1–6, IEEE, 2023.

[17] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies," *Journal of Biomedical Informatics*, vol. 113, p. 103655, 2021.

[18] E. Delaney, D. Greene, and M. T. Keane, "Instance-based counterfactual explanations for time series classification," in *International Conference on Case-Based Reasoning*, pp. 32–47, Springer, 2021.

[19] R. Guidotti, "Counterfactual explanations and how to find them: literature review and benchmarking," *Data Mining and Knowledge Discovery*, pp. 1–55, 2022.

[20] O. Li, H. Liu, C. Chen, and C. Rudin, "Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[21] R. Crupi, A. Castelnovo, D. Regoli, and B. San Miguel Gonzalez, "Counterfactual explanations as interventions in latent space," *Data Mining and Knowledge Discovery*, pp. 1–37, 2022.

[22] A. L. Alfeo, M. G. Cimino, G. Gagliardi, *et al.*, "Automatic feature extraction for bearings' degradation assessment using minimally preprocessed time series and multi-modal feature learning.," in *IN4PL*, pp. 94–103, 2022.

[23] B. Zhou, D. Bau, A. Oliva, and A. Torralba, "Interpreting deep visual representations via network dissection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 9, pp. 2131–2145, 2018.

[24] I. Stepin, J. M. Alonso, A. Catala, and M. Pereira-Fariña, "A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence," *IEEE Access*, vol. 9, pp. 11974–12001, 2021.

[25] G. Gagliardi, A. L. Alfeo, V. Catrambone, M. G. Cimino, M. De Vos, and G. Valenza, "Using contrastive learning to inject domain-knowledge into neural networks for recognizing emotions," in *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1587–1592, IEEE, 2023.

[26] H. Phan, O. Y. Chén, M. C. Tran, P. Koch, A. Mertins, and M. De Vos, "Xsleepnet: Multi-view sequential model for automatic sleep staging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5903–5915, 2021.

[27] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert systems with applications*, vol. 36, no. 2, pp. 3336–3341, 2009.

[28] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[29] Z. Qi, S. Khorram, and F. Li, "Visualizing deep networks by optimizing with integrated gradients.," in *CVPR Workshops*, vol. 2, pp. 1–4, 2019.

[30] Y. Hailemariam, A. Yazdinejad, R. M. Parizi, G. Srivastava, and A. Dehghantanha, "An empirical evaluation of ai deep explainable tools," in *2020 IEEE Globecom Workshops (GC Wkshps*, pp. 1–6, IEEE, 2020.

[31] M. Foulsham, B. Hitchen, and A. Denley, "Gdpr: how to achieve and maintain compliance," 2019.