physiologia

MDPI

Systematic Review

Advances in Modeling and Interpretability of Deep Neural Sleep Staging: A Systematic Review

Reza Soleimani [1,*], Jeffrey Barahona [1], Yuhan Chen [1], Alper Bozkurt [1], Michael Daniele [1], Vladimir Pozdin [2] and Edgar Lobaton [1]

[1] Department of Electrical and Computer Engineering, North Carolina State University, Engineering Bldg II, 890 Oval Dr, Raleigh, NC 27606, USA; jabaraho@ncsu.edu (J.B.); ychen239@ncsu.edu (Y.C.); aybozkur@ncsu.edu (A.B.); mdaniel6@ncsu.edu (M.D.); ejlobato@ncsu.edu (E.L.)

[2] Department of Electrical and Computer Engineering, Florida International University, 10555 W. Flagler St., Miami, FL 33174, USA; vpozdin@fiu.edu

* Correspondence: rsoleim@ncsu.edu

Abstract: Sleep staging has a very important role in diagnosing patients with sleep disorders. In general, this task is very time-consuming for physicians to perform. Deep learning shows great potential to automate this process and remove physician bias from decision making. In this study, we aim to identify recent trends on performance improvement and the causes for these trends. Recent papers on sleep stage classification and interpretability are investigated to explore different modeling and data manipulation techniques, their efficiency, and recent advances. We identify an improvement in performance up to 12% on standard datasets over the last 5 years. The improvements in performance do not appear to be necessarily correlated to the size of the models, but instead seem to be caused by incorporating new architectural components, such as the use of transformers and contrastive learning.

Keywords: sleep stage classification; EEG; ECG; deep learning; contrastive learning; transfer learning; data augmentation; interpretability

Citation: Soleimani, R.; Barahona, J.; Chen, Y.; Bozkurt, A.; Daniele, M.; Pozdin, V.; Lobaton, E. Advances in Modeling and Interpretability of Deep Neural Sleep Staging: A Systematic Review. *Physiologia* **2024**, *4*, 1–42. https://doi.org/10.3390/physiologia4010001

Academic Editor: Philip J. Atherton

Received: 24 July 2023
Revised: 30 August 2023
Accepted: 11 September 2023
Published: 20 December 2023

1. Introduction

Sleep disorders can impair attention, long-term memory, decision making, and overall negatively affect cognitive performance [1,2]. Diagnosing sleep-related disorders early is therefore important for providing timely treatment, but the task is time-consuming and can be biased by the clinician. Due to the repetitiveness of the task, deep learning techniques can be applied to assist clinicians in sleep scoring and in diagnosing illnesses. Additionally, polysomnographic sleep studies are typically conducted in the clinic using numerous sensors and electrodes, which can be uncomfortable, potentially reducing the quality of sleep data recorded. Besides enabling inference, deep learning techniques may also be applied to reduce the number of sensors needed and to enable the use of wearable sensors for at-home care and preliminary diagnoses.

A polysomnogram (PSG) captures brain waves, oxygen levels, eye and leg movements, heart rate, and breathing throughout the patient's sleep cycle, segmented into 30 s segments referred to as epochs [3]. The American Academy of Sleep Medicine (AASM) and the Rechtschaffen and Kales (RK) manuals are used to define sleep stages. In this paper, we will use the AASM manual definitions. The AASM [4] manual defines five stages of sleep (Wake, N1, N2, N3, REM). Each of these stages has its own characteristics, as shown in Table 1. It can be seen in Table 1 that each stage has its own frequency bands and specific rhythm. N3 occupies the 1–4 Hz band and consists of Delta rhythm. N1 is presented in 4–12 Hz and has Theta and Alpha rhythms. REM is defined by the 4–8 and 16–32 Hz bands. The wake stage is present in 8–12 and 16–32 Hz while N2 is present just in the 12–16 Hz band. The N2 stage typically consists of the K-complex and spindles. The REM stage is typically composed

of saw-tooth waves. Atypical patterns in sleep stages and progression can help diagnose sleep disorders, provide indicators for sleep quality, and provide additional details about the patient's health. Given the large amount of data in a single overnight PSG, clinicians expend significant effort to annotate the data. Furthermore, PSGs are highly variable with multiple sources of aleatoric and epistemic uncertainty, including age, pathology, electrode placement, and annotator bias. This causes even expert human annotators to disagree on epoch labels. With this in mind, machine learning algorithms will never achieve perfect alignment with human annotators because of the level of human disagreement in sleep score labels. At best, we can expect machine learning models to agree with human annotators to the extent that human annotators agree with each other.

Table 1. AASM EEG Frequency Definitions.

Rhythm	Frequency Band (Hz)	Target Stages
Delta	1–4	N3
Theta	4–8	N1, REM
Alpha	8–12	N1, Wake
Sigma	12–16	N2
Beta	16–32	Wake, REM

Automating sleep staging can remove human subjectivity in decision making and reduce the labor considerably by automatically assigning labels based on models trained with multiple annotators. Deep learning models, to a much greater extent than classical models, are capable of capturing information from large datasets. Despite reaching near-human-level performance, deep learning models are opaque, obfuscating the model's decision-making process [5]. Opacity in modeling limits the usefulness of these techniques in the medical community. However, ongoing efforts to make deep learning models more transparent and explainable will enable automatic sleep scoring to be more widely adopted. There have been some efforts in recent years to summarize the different methodologies on sleep studies [5–7]. However, these papers do not detail the modeling efforts (e.g., the reasoning behind the choices of architectures) and do not discuss their nuances and how they affect model performance.

In this review, the main application discussed will be sleep stage scoring, defined as the annotation of sleep signals. We provide an overview of recent research papers between 2018 and 2023 on sleep staging using deep learning. We provide an overview and survey of the sleep classification process and describe in detail the state-of-the-art (SOTA) techniques for sleep staging, including convolutional and recurrent models, transformer-based models, contrastive learning, transfer learning, domain adaptation, and interpretability. We aim to identify recent trends in performance improvement and the causes of these trends. Additionally, we investigate recent models to explore data manipulation techniques, their efficiency, and recent advances.

The rest of the paper is organized as follows. In Section 2, the criteria for the process with which we chose the articles is presented together with a description of the metrics used. In Section 3, we go over sleep-related datasets and some commonly used techniques and algorithms to standardize the signals. In Section 4, different types of modeling schemes are discussed in detail. In Section 5, the explainability for deep neural models is discussed. In Section 6, different models are compared and analyzed. Lastly, we summarize our findings in Section 7.

2. Method

2.1. Design

The literature search for this review was performed according to the PRISMA guidelines [8]. Only papers published between 1 January 2018 and 23 March 2023 were considered. A corpus of machine learning papers were collected from Google Scholar and Web of Science using terms "Deep learning", "Sleep staging", "electroencephalogram (EEG)",

"electrooculogram (EOG)", and "electromyogram (EMG)". The same exact set of terms were used for both search engines. We aimed to exclude papers that focus on hardware design, Internet of Things (IOT), and wearables. We consider only those papers that focus on the modeling and algorithms. To achieve this, we added some exclusion criteria, such as "wearable", "IOT", "artifact", "hardware", and "mobile" to remove the papers that are not relevant to this study. Of these papers, only publications that were cited by multiple researchers (at least 5 citations) or that presented a novel idea in the field were considered for review. The papers that were removed due to low citation did not have novel modeling, while some recent papers with zero citation that presented new techniques were included in the study. These inclusion and exclusion criteria were chosen carefully to make sure all the relevant studies were included. The PRISMA flowchart for article selection is shown in Figure 1. The most commonly used datasets were selected from these papers and are characterized in Section 3.

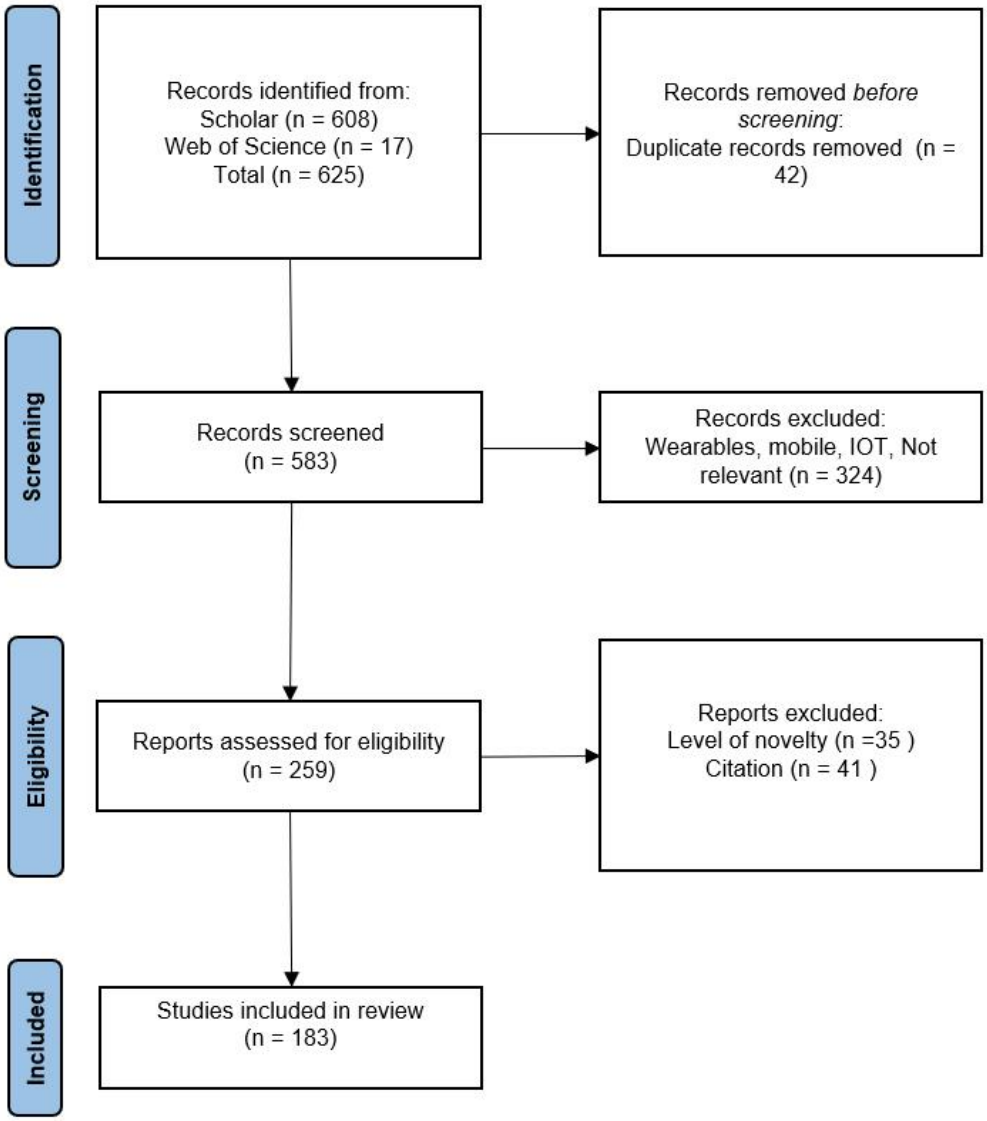


Figure 1. PRISMA flowchart for selecting the papers for deep-learning-based sleep studies.

2.2. Metrics

To better contextualize the performances of these models, here we define the most useful metrics for the task. Cohen's kappa is a measure of reliability between two annotators rating the same thing, correcting for the probability that two raters agree by chance. The metric is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the relative observed agreement among raters, and p_e is the probability of agreement by chance. In other words, p_o is the joint probability of rater agreement, and p_e is the marginal probability of the annotation choices. For reference, to achieve human-level performance, an automated sleep stage scoring algorithm should achieve a Cohen's kappa of 0.76 [9], where 0.76 is the inter-rater agreement between trained scorers using the 2007 AASM scoring rules [10]. However, in the case of unbalanced datasets, Cohen's kappa may be an inadequate measure of classification performance, penalizing judges with different marginal probabilities for annotations [11]. Many papers reviewed here do not provide Cohen's kappa and instead provide an F1 score or an accuracy score.

The F1 score is the harmonic mean of precision and recall, providing a balanced performance measure that takes into account class imbalance. Simple accuracy works well when the classes are balanced in the dataset, but is a less informative metric when used for a dataset with highly imbalanced classes. Sleep datasets are often imbalanced, so papers often pair an accuracy metric with an F1 metric or another balance agnostic metric. Additionally, given the disparity between modalities represented in each of the datasets, comparisons between models will be made based on what dataset they use and the modalities used in the model for the sleep staging task. Furthermore, we will examine the complexity of the models and analyze the complexity of the model versus the performance.

3. Datasets

In this section, the datasets used in the papers reviewed are described. In particular, this review uses the AASM guidelines and makes conversions where necessary. In Table 2, the number of epochs containing each sleep stage as defined by the AASM are presented. As can be seen in the table, the classes are imbalanced, posing additional difficulties for automatic sleep scoring. The effects of this disparity can be seen throughout the presented papers.

Table 2. Percentage of Epochs Assigned to each Target Stage Label in the Different Datasets.

Dataset	Wake%	N1%	N2%	N3%	REM%	Total
Sleep EDF	19.50	6.60	42.07	13.48	18.24	42,308
MASS	9.88	7.88	50.90	13.18	18.15	57,395
SHHS	23.08	4.01	44.20	14.64	15.07	5,421,338
ISRUC (Subgroup 1)	23.62	11.54	32.74	18.69	13.41	90,123
ISRUC (Subgroup 2)	17.18	14.88	33.98	18.80	15.16	14,207
ISRUC (Subgroup 3)	19.53	13.44	30.96	22.79	13.28	8883
UCD	22.60	16.30	33.60	12.80	14.50	20,774

3.1. Data Description

In this section, we briefly introduce popular datasets for sleep staging. Private or niche datasets are not enumerated here.

- SleepEDF (all versions): SleepEDF [12] is a publicly available dataset of physiological signals with corresponding sleep annotations. The initial version of this dataset was small and was originally published in 2002. In 2013, version 1 was published, which was greatly expanded to contain 61 PSGs with accompanying hypnograms. In March 2018, the database was further expanded to version 2 containing 197 PSGs with accompanying hypnograms. Each subject wore a modified Walkman-like cassette-tape recorder to record their normal activities. For each subject, two PSGs of about 20 h each

were recorded during two subsequent day–night periods at the subjects' homes. PSG recordings include EEG from Fpz-Cz and Pz-Oz electrode locations, EOG, and submental chin EMG signals. Corresponding hypnograms (sleep patterns) for these recordings were manually annotated by well-trained technicians according to the Rechtschaffen and Kales manual, and are also available. Annotation consists of sleep stages W, R, 1, 2, 3, 4, M (Movement time), and "not scored". The Sleep EDF dataset has been used in numerous studies to develop and evaluate algorithms for automatic sleep staging recognition problems, as well as to investigate various aspects of sleep and its disorders. Papers using this dataset include [13–22].

- MASS: This is an open-access and collaborative database of laboratory-based PSG recordings [23]. The aim of the dataset is to offer a consistent and readily available data source for evaluating different systems developed to automate sleep analysis. The MASS dataset is unique in its size and scope, with the cohort consisting of 200 complete nights of polysomnograms, recorded from a diverse group of individuals (97 males and 103 females) with an average age of 40.6 years (ranging from 18 to 76 years). The dataset provides a valuable resource for researchers and practitioners working in the field of sleep analysis, with a focus on benchmarking and improving automated sleep analysis systems. This dataset includes polysomnographic recordings of sleep patterns, as well as related data such as participants' medical histories and sleep questionnaires. All recordings feature a sampling frequency of 256 Hz and an EEG montage of 4–20 channels plus standard EOG, EMG, ECG, and respiratory signals. Papers using this dataset include [14,15,24,25].
- SHHS: The Sleep Heart Health Study (SHHS) dataset [26] was collected as part of a study aimed at investigating the relationship between sleep-disordered breathing and cardiovascular disease. This dataset collected data from over 6441 healthy participants without treatment of sleep apnea, with an age over 40 years for the first round (SHHS-1). A second polysomnogram (SHHS-2) was obtained from 3295 of the participants. The dataset includes PSG recordings, as well as data on demographics, medical history, and sleep habits. The PSG recordings were scored by certified technicians using standard criteria for sleep staging, respiratory events, and other sleep-related events. The dataset also includes information on the presence and severity of sleep-disordered breathing, such as apnea and hypopnea. The SHHS dataset has been used in numerous studies investigating various aspects of sleep and sleep-disordered breathing. The dataset is publicly available through the National Sleep Research Resource. Papers using this dataset include [9,18,27–31].
- CAP: The Cyclic Alternating Pattern (CAP) is a periodic EEG activity observed during non-REM (NREM) sleep. The CAP Sleep dataset [32] is comprised of 108 polysomnographic recordings collected at the Sleep Disorders Center of the Ospedale Maggiore in Parma, Italy. These recordings encompass a minimum of three EEG channels, EOG, EMG of the submentalis muscle, bilateral anterior tibial EMG, respiratory signals, and EKG. The CAP Sleep dataset includes 108 polysomnographic recordings from the Sleep Disorders Center of Ospedale Maggiore in Parma, Italy. The data include recordings from 16 healthy subjects and 92 pathological recordings with various sleep disorders, such as Nocturnal Frontal Lobe Epilepsy (NFLE), REM Sleep Behavior Disorder (RBD), Periodic Limb Movement (PLM), insomnia, narcolepsy, Sleep-Disordered Breathing (SDB), and bruxism. The recordings contain data from multiple sources including EEG, EOG, and ECG, along with muscle and respiration signals.
- ISRUC: The ISRUC-Sleep dataset [33] is a comprehensive polysomnographic (PSG) resource aimed at aiding sleep research. This dataset features data from adults, including healthy individuals and those suffering from sleep disorders, with some on sleep medication. The dataset is organized to accommodate different research goals, with information from 100 subjects each with one recording session, eight subjects with two sessions each for tracking changes over time, and 10 healthy subjects in a single session for comparison studies. Each PSG recording contains electrophysiological and

pneumological signals, among other contextual details, all of which have been visually scored by two human experts. Papers using this dataset include [16,24].

- MESA: The Multi-Ethnic Study of Atherosclerosis (MESA) [34,35] is a long-term study sponsored by the National Heart Lung and Blood Institute (NHLBI), focusing on the development and progression of subclinical cardiovascular disease in over 6800 ethnically diverse individuals. Between 2010 and 2012, a subset of participants undertook a Sleep Exam (MESA Sleep), which studied how sleep variations and disorders correlate with subclinical atherosclerosis. The raw sleep data, including polysomnography and actigraphy, are publicly accessible for further research. Papers using this dataset include [27].
- Bonn database: The Bonn-Barcelona micro- and macro- EEG database [36] comprises 960 multichannel EEG signals, each with a duration of 32 s, extracted from long-term EEG data. The selection process did not involve any clinical criteria, such as the presence or absence of epileptiform activity, and all data are de-identified. Papers using this dataset include [37].
- UCD: The St. Vincent's University Hospital/University College Dublin Sleep Apnea Database [38] consists of 25 full overnight polysomnograms, accompanied by a simultaneous three-channel Holter ECG, collected from adult subjects suspected of having sleep-disordered breathing. The subjects, aged between 28 and 68 years with no known cardiac disease or autonomic dysfunction, were randomly selected over six months from patients referred to St Vincent's University Hospital's Sleep Disorders Clinic. Polysomnograms were obtained using the Jaeger–Toennies system, capturing a variety of signals including EEG, EOG, EMG, ECG, and body position, among others. Additionally, three-channel Holter ECGs were recorded using a Reynolds Lifecard CF system. All data are available in EDF format. Papers using this dataset include [39,40].
- MIT-BIH Polysomnographic: The MIT-BIH Polysomnographic Database [41] contains over 80 h of four-, six-, and seven-channel polysomnographic recordings. Each recording includes a beat-by-beat annotated ECG signal, as well as EEG and respiration signals annotated in relation to sleep stages and apnea. These data are collected in Boston's Beth Israel Hospital Sleep Laboratory. Papers using this dataset include [29].
- Apnea-ECG: Apnea-ECG [42] consists of 70 records, divided into a learning set of 35 records, and a test set of 35 records. Each recording includes a continuous digitized ECG signal, a set of apnea annotations and machine-generated QRS annotations. Papers using this dataset include [40].
- DREAMS: The DREAMS Databases [43] consist of recordings annotated in microevents or in sleep stages by several experts. They were acquired in a sleep laboratory of a Belgium hospital using a digital 32-channel polygraph (BrainnetTM System of MEDATEC, Brussels, Belgium). This database is split into 8 databases according to the annotation carried out. Papers using this dataset include [20,21].
- You Snooze You Win: The PhysioNet/Computing in Cardiology Challenge 2018: This dataset was contributed by the Massachusetts General Hospital's (MGH) Computational Clinical Neurophysiology Laboratory (CCNL) and the Clinical Data Animation Laboratory (CDAC) [44]. The dataset consists of 1985 subjects and the sleep stages were annotated by clinical staff based on the American Academy of Sleep Medicine manual with six sleep stages noted in 30 s intervals: wakefulness, stage 1, stage 2, stage 3, rapid eye movement (REM), and undefined. Certified technologists also annotated waveforms for the presence of arousals interrupting sleep, classifying them into various categories such as spontaneous arousals, respiratory effort related arousals (RERA), bruxisms, and others. The physiological signals recorded during the subjects' sleep include EEG, EOG, EMG, ECG, and oxygen saturation (SaO_2). All signals, excluding SaO_2, were sampled at 200 Hz and measured in microvolts. SaO_2 was resampled to 200 Hz and measured as a percentage. Papers using this dataset include [31].

A summary of selected datasets with their respective modalities is presented in Table A1.

3.2. Data Preprocessing

To use any of these datasets for automatic scoring, disorder diagnoses, and other medical applications, the data must be preprocessed into a form more amenable for use with deep learning. In this section, an overview of commonly used preprocessing techniques are presented.

Many authors [28,45–47] preprocess the data by removing frequency components outside of the accepted ranges of brainwave activity. In EEG, this is usually performed with a Butterworth band-pass filter tuned to 0.4–30 Hz. This removes outside noise from the input signal. Additionally, the data is often normalized using a variety of techniques. In [47], the EEG data is normalized using the 5th and 95th quantiles of each PSG individually. More commonly, the EEG is normalized using standard methods like computing z-scores based on PSG statistics. Statistical and handcrafted features [21] are also commonly extracted from PSG recordings as a pre-processing step.

Some authors [45,48,49] provide semantic structure to EEG data by segmenting the input signals into representative sub-bands containing the alpha, beta, delta, and theta waves. These bands were used as direct inputs to deep learning methods or used to generate handcrafted statistical features for analysis. Methodologies like these inject human insight into the modeling schemes which can improve results by simplifying the task. The raw EEG signals may be transformed using spectrograms [50], empirical mode decomposition, and discrete cosine transforms [51] to add additional structure to the data, allow for transfer learning techniques to be used, or for interpretability reasons.

3.3. Addressing Dataset Deficiencies

In this section, we investigate different techniques that have been used in the literature to overcome label imbalance, poor data quality, and shortage. For classification, these dataset deficiencies can reduce final performance and model reliability. To address deficiencies, techniques for data augmentation have been used, such as techniques for minimizing these issues through data augmentation from statistical methods, generative adversarial methods [52], resampling, etc. Data augmentation of EEG signals should not alter the semantic meaning of the signal, limiting the options available for augmentation.

A popular augmentation method is to add stochasticity to data [47,50,53–59]. Jittering, masking, flipping, scaling, and spectrogram reflection are some augmentation methods that have been used in the literature. In jittering, a random uniform noise is added to the physiological signal. With masking, parts of the signal will be randomly set to a predefined value. In [60], the authors proposed four different augmentation techniques, three of which we will explain since one of the methods (horizontal flipping) has already been described. The first proposed method is *Epoch Overlapping* . In this method, the authors start the current epoch, x_i, by using $2/3$ or $4/5$ of the previous epoch, x_{i-1}, as the starting point to include 30 s worth of recordings, with the label of the current epoch. The second approach is called *Mixup*. In this setup, two randomly selected signals (x_i, x_j) with their corresponding labels (y_i, y_j) are linearly mixed as follows:

$$x_{mix} = \lambda x_j + (1 - \lambda)x_i$$

$$y_{mix} = \lambda y_j + (1 - \lambda)y_i$$

where $\lambda \sim Beta(\alpha, \alpha)$ with $\alpha = 0.25$. The last method is called *Random Cutout*. In this method, the signal will be randomly zeroed out for 6 s in each epoch. This is used to simulate sensor error, such as the times when the leads have been detached from the skin.

While these methods add variety to the dataset, they do not address underlying issues with class imbalance. Common methods for dealing with class imbalances include resampling, bootstrapping [61], and deep learning techniques. In [62], the authors proposed using two methods: (1) generative adversarial network (GAN) and Gaussian white noise (GWN) to increase N1 stage samples; (2) balance the relationship between the trained model and the original imbalanced dataset by having different weights for different classes.

The authors argue that the reason for adding GWN to the signals is twofold. The first reason is that this type of noise is present in EEG acquisition. So adding this noise with different intensities will imitate the true signal. The second reason is that the new data with noise addition can provide new features that help with generalization. Mitigating class imbalance through modeling rather than data augmentation (DA) would be more natural since the imbalance exists in the true distribution of the sleep signals for different classes. Other deep techniques used for balancing class imbalances include modifications to the loss [16,51], which reduce the impact of over-represented classes and increase the impact of underrepresented classes while preserving as much information from the data as possible.

4. Modeling

Deep learning models have revolutionized how time series data are modeled, capturing complex non-linear relationships that are not representable with classical modeling schemes. For sleep studies, models include convolutional neural networks (CNN), recurrent neural networks (RNN), long-short-term memory models (LSTM), graph neural networks (GNNs), and transformers. More recent efforts have made use of contrastive learning to further enforce feature differentiation via instance-level comparisons, allowing models to learn inter- and intra-group descriptors simultaneously. In this section, we explore the types of models used, training schemes, and learning techniques applied.

4.1. Transfer Learning and Domain Adaptation

In recent years, large models have been developed and have shown very promising results in different fields such as computer vision (CV) and natural language processing (NLP). In particular, in NLP, large models such as BERT [63] and GPT [64] have shown significant performance boosts on downstream tasks such as text classification or summarization by having millions, if not billions, of parameters. The base of such popular models is a transformer [65], which revolutionized the overall field of NLP. Since not all researchers have access to large clusters of GPUs to train big models, transfer learning is becoming more and more popular. By having large models as a feature extractor or an embedding space, researchers can perform downstream tasks having to train the model from scratch. This method has been adapted for physiological signals as well. In recent years, researchers have been utilizing deep learning models to perform classification for sleep staging. Most models are trained on a specific dataset and they may not generalize well to other datasets, which shows some of the limitations and challenges that we will explain in subsequent sections. In this section, we review the papers where the authors study transfer learning for sleep studies, and also, we study domain adaptation to investigate the generalizability of the models against different datasets.

The purpose of transfer learning is to leverage models learned in another domain for a different, related domain, providing a stronger initialization for the model in the target domain. Domain, here, can refer to a difference in tasks, datasets, or modalities. This provides numerous advantages than just training on the target domain including faster convergence, greater performance, or a reduction in computational capacity required for training, especially if the target dataset is size-constrained [66]. Furthermore, a source domain for transfer learning need not be directly related to the target domain; models in computer vision and natural language processing have successfully been used to provide strong starting points in seemingly unrelated domains including sleep stage scoring [66,67]. In this section, techniques for and applications of transfer learning and domain adaptation in sleep stage scoring are explored.

Earlier efforts in transfer learning focused on using models trained on large public sleep datasets to improve model performance for small datasets. In [66], the authors trained DeepSleepNet [68] on large public sleep scoring datasets and fine-tuned the model on a small private sleep scoring dataset, reporting significant improvement over training on the small dataset alone. Transfer learning can also be performed in the form of knowledge distillation [69] for model compression or wearable use cases.

Performance gain from using transfer learning can be limited due to loss of local domain information, different data distributions, changes in class information, etc. Even in the simplest cases, such as using different source datasets with the same task for transfer, the differences in source and target datasets can result in limited performance gains. In [70], the transferability of sleep stage scoring datasets with TinySleepNet [15] were assessed using a relative performance improvement ranking metric. Datasets were differentiated by their recording characteristics, the health of the patients represented in the dataset, and the recording environments. They found that, the greater differences between dataset characteristics, the worse the transfer. To reduce the performance loss from domain mismatch, researchers have proposed multiple techniques for adapting source to target domains.

The researchers in [71] used an adversarial learning scheme to reduce loss of domain-specific information from sharing a model between domains and to reduce the classification accuracy degradation. First, they proposed an unshared attention mechanism to maintain domain-specific information for target and source domains. Second, they designed an iterative self-training algorithm to enhance classification accuracy on the target domain by using pseudo labels from the target domain. They also proposed dual distinct classifiers to increase the robustness and quality of the pseudo labels. Other adversarial domain adaptation techniques have been proposed and applied by [60,72–74], dealing with various pain points including cross-subject adaptation and population health differences. A domain mismatch may also result from data-distribution-dependent model parameters like batch normalization. The authors of [75] deal with this by learning the batch normalization characteristics of the target dataset and replacing them in the target model.

By training the model on multiple heterogeneous datasets or on a single large, diverse dataset [76,77], the researchers were able to demonstrate that deep learning models are capable of adapting to the differences in the datasets in the form of differing input shapes, electrode positioning differences, and distribution shifts.

4.2. Convolutional and Recurrent Models

In this section, we explore different types of architectures that involve CNNs, RNNs, and GNNs. We will start by investigating the most influential papers.

In [68], the authors proposed a model named DeepSleepNet with publicly available code. This model consists of two different filter sizes that learn local time-invariant features at different resolutions to have a richer representation space. A bidirectional LSTM is appended to the CNN layer to encode the temporal information and sleep transition in each epoch of data. This model is trained in two steps. In the first step, the CNN layers are trained to extract the representation for the LSTM layer. This helps the model to be better conditioned on the data itself, rather than random initialization for the second step of the training. In the second step, the whole model is fine-tuned by using the pre-trained layer.

In [78], the authors proposed XSleepNet, which is a sequence-to-sequence sleep staging model capable of learning joint representation from multi-modal raw signals and time-frequency images (also known as a spectrogram), as shown in Figure 2. The model blocks are a combination of CNNs, RNNs, and attention mechanisms. In particular, $\mathcal{F}_1$ is a fully-connected CNN which extracts features from raw inputs, and $\mathcal{F}_2$ is an attention-based RNN with learnable filterbank layers to convert spectrograms into high-level features. In the last layer, there are three branches for raw, spectrogram, and a mixture of them to calculate the loss. Authors argue that different views (raw and spectrogram) may generalize or overfit at different rates. For more information, please refer to Appendix A.1.

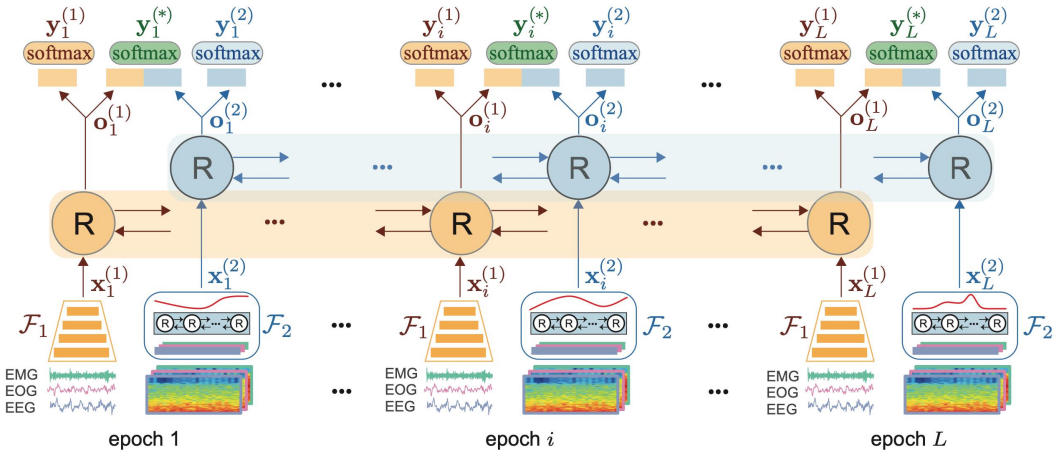


Figure 2. The model architecture for XsleepNet. Image taken from [78].

In [79], the authors proposed a new architecture consisting of a modified ResNet-50 layer and a two-layered bidirectional LSTM named IITNet. In this model, a feature representation is extracted at sub-epoch level (less than 30 s) by using residual neural networks and extracting intra- and inter-epoch temporal context from the time series by using bidirectional RNNs. The latter helps to learn the transition rules between different epochs. If only the target epoch is used as an input to the model, the intra-epoch features will be learned and utilized for decision making. In order to capture intra- and inter-epoch temporal context, the target epoch and its previous neighboring epochs are fed to the model. This architecture mimics real-time sleep scoring where we cannot use any epoch from future epochs.

A mixture of convolution and attention layers were proposed in [80]. The CNN layer extracts local features and the multi-layer attention module is used to learn salient intra- and inter-epoch features, as shown in Figure 3. For the unbalanced dataset, the authors proposed to use a weighted loss function during training to improve model performance on minority classes. In order to learn some short-term features, such as K-complex and spindles, which last around 0.5–1.5 s, each epoch is divided into multiple windows (29 in the paper), as shown in Figure 3a. In Figure 3b, more details on how each window is treated are shown, and Figure 3c shows the attention module.

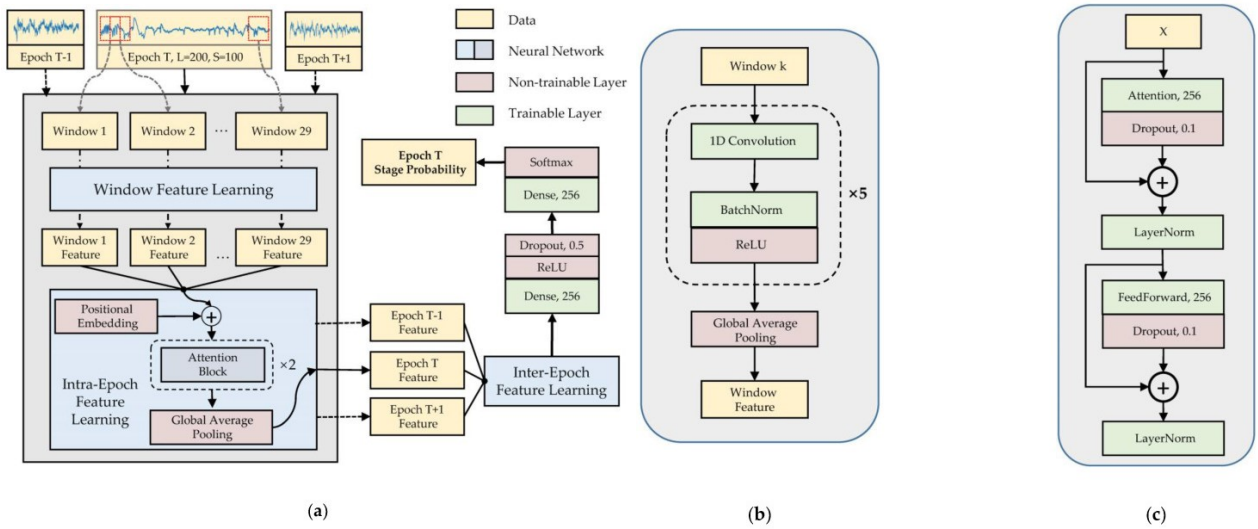


Figure 3. (**a**) Model architecture. (**b**) Feature learning for per window. (**c**) Attention module. Image taken from [80].

The authors of [81] proposed a novel framework based on local pattern transformation (LPT) methods and CNNs for automatic sleep stage scoring. Unlike previous works in other fields, these methods were not employed for manual feature extraction, which requires expert knowledge and the pipeline behind it might bias results. The transformed signals were directly fed into a CNN model (called EpochNet) that can accept multiple successive epochs. The model learns features from multiple input epochs and considers inter-epoch context during classification. The study investigates the role of LPT methods prior to the classification. Four LPT methods are used for both single and multi-channel setup. More details on this model can be found in Appendix A.2.

In [82], deviating from the previous architectures, the authors proposed a model called U-Time. This model is a fully convolutional encoder–decoder network based on U-Net architecture [83], which originally was used for image segmentation. U-Time adopts key concepts from U-Net for 1D time-series segmentation and maps sequential inputs of arbitrary length to sequences of class labels. To achieve this, every individual time-point in the signal is classified and aggregated over fixed intervals to reach the final prediction. This model is easier to train when compared to [68,84], it is more robust due to not using an RNN layer, and it has a smaller number of parameters.

The authors of [85] proposed to use a Pearson-correlation-based graph attention network (PearNet) to address the external relationship of electrodes in different regions of the brain and internal relationships between segments of electrodes within specific brain region problems. Graph nodes are produced by spatial-temporal convolutional layers with SE blocks followed by Pearson correlation block. The graph is then passed on to another module to adaptively learn the node connection for graph structure. The whole structure is shown in Figure 4. The spatial and temporal convolution networks are depicted in Figures 5 and 6, respectively. The VIF (variance inflation loss) determines the degree of multi-collinearity [86].

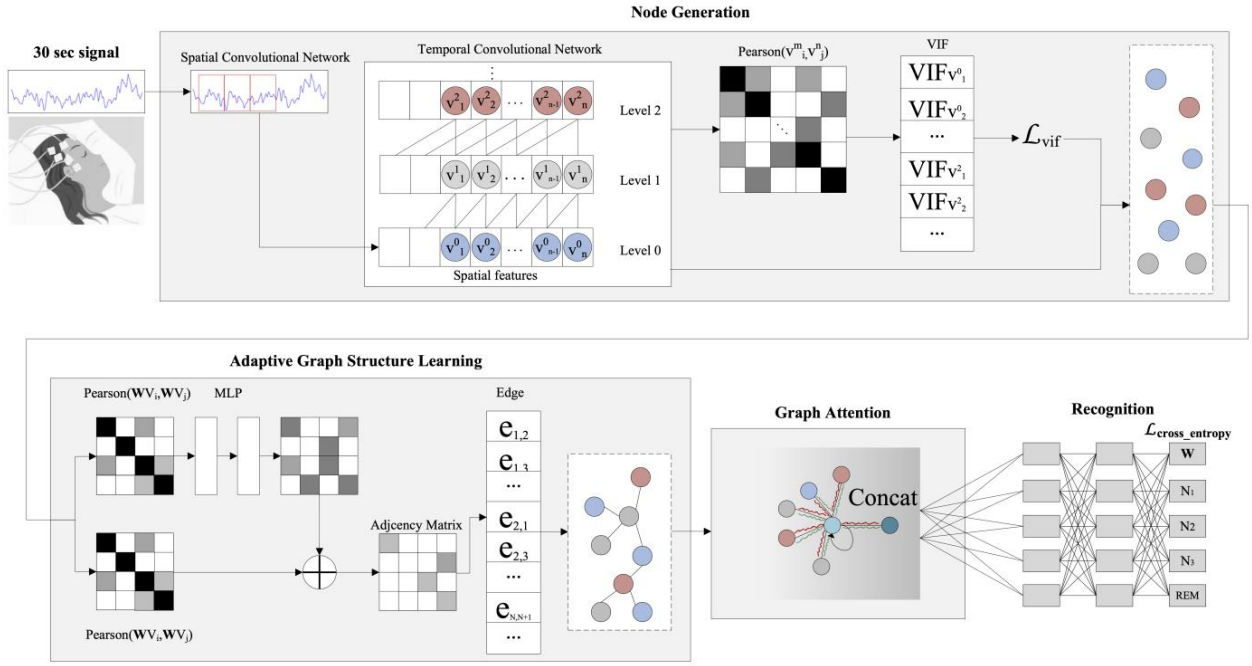


Figure 4. Model architecture for PearNet [85]. Image taken from [85].

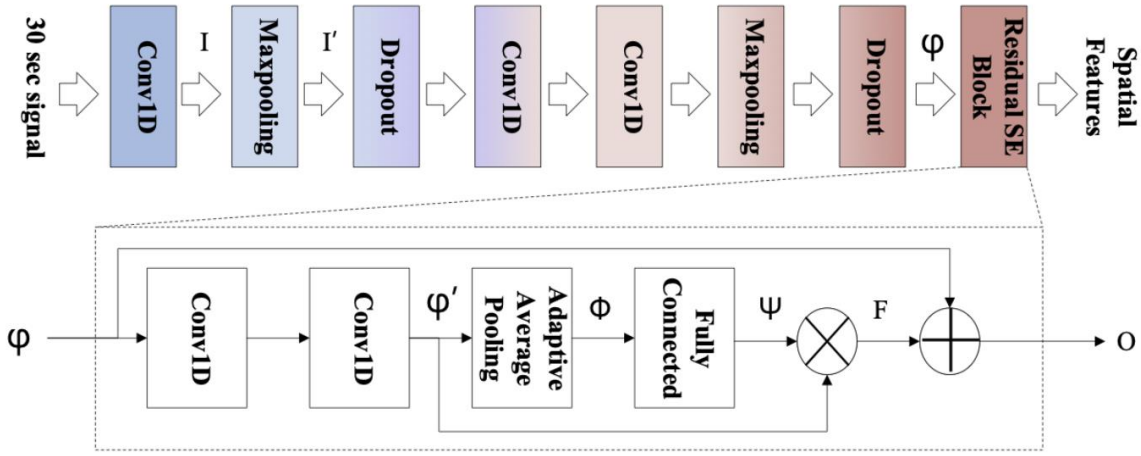


Figure 5. Spatial convolution network. Image taken from [85].

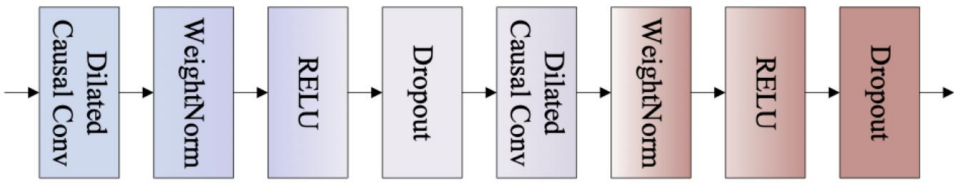


Figure 6. Temporal convolution network. Image taken from [85].

In [87], the proposed model (SleepExpertNet) aims to classify sleep stages in a similar way to how the experts evaluate PSG recordings by using the relationship between the EEG signals and the patterns of changes in sleep stages by combining a time series-based multi-head attention mechanism and bidirectional LSTM. To extract the temporal-frequency features from raw EEG signals, a bandpass filter is applied to the single-EEG channel in various frequency bands. Embedded vector generation, including a novel CNN-based architecture, is proposed to consider the neurophysiological characteristics of EEG signals. To solve the class imbalance problem, the cost-sensitive learning of focal loss and a novel weighted random sampling method are applied in the proposed model structure. To improve the classification performance, an ensemble decision algorithm, which can be used in the proposed model without additional training, is developed. The model structure is shown in Figure 7.

In [88], a novel multi-view fusion model named MVF-sleepNet was proposed. This model accepts EEG, ECG, and EOG signals as input. In order to properly encode the relationship between different modalities, the authors proposed constructing two views of time-frequency images (TF images) and graph-learned graphs (GL graphs). In Figure 8, the model framework is shown. This model consists of two streams, one for TF images and one for graph features. For sequentially timed TF images, a combination of VGG-16 and GRU with attention mechanism networks is used for spectral and temporal representation, respectively. Furthermore, for graphs, a combination of Chevyshev graph convolution and temporal-spatial attention mechanism, and temporal convolutional network is employed to learn spatial-temporal representation sequentially timed GL graphs.

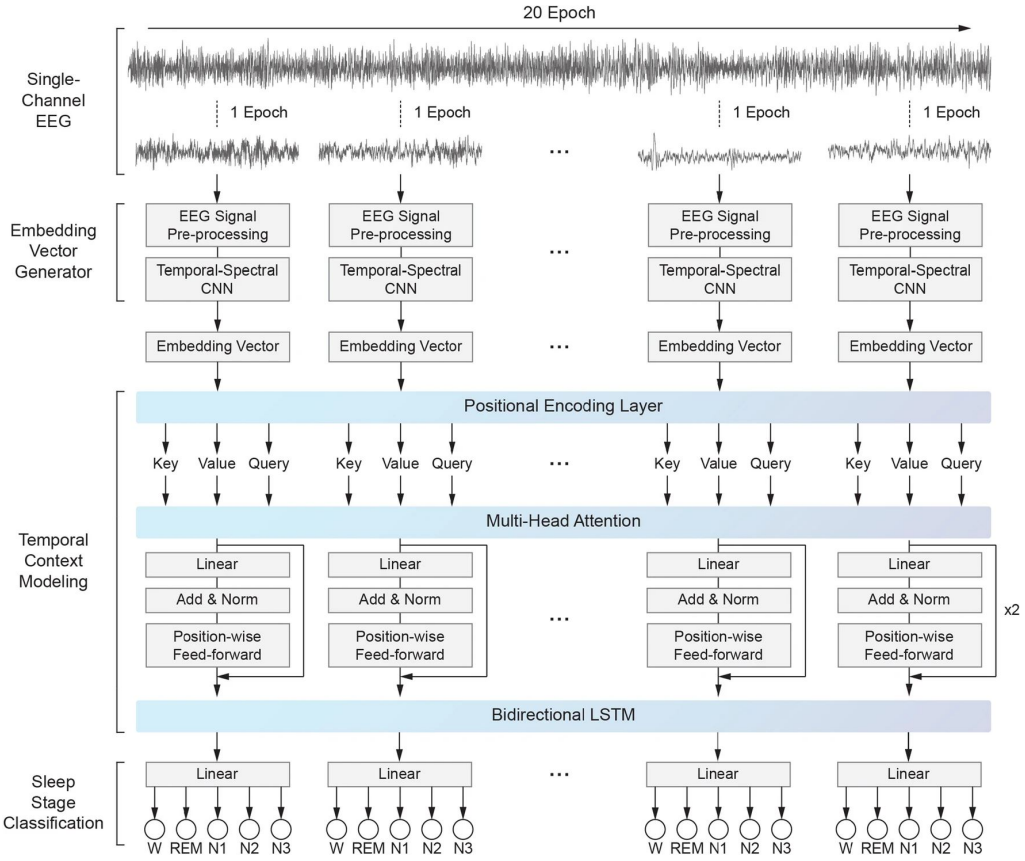


Figure 7. SleepExpertNet architecture. Image taken from [87].

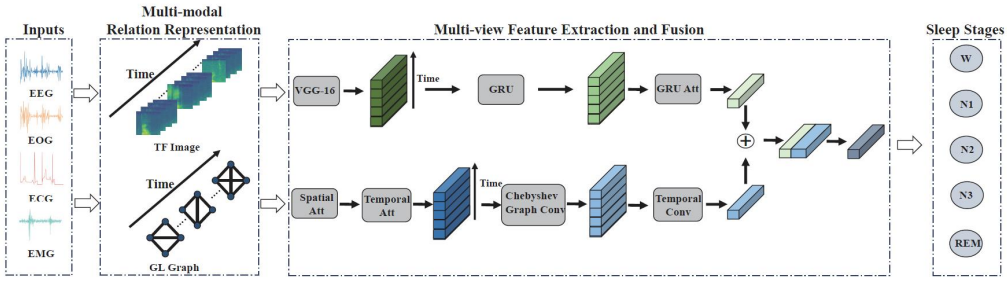


Figure 8. Model architecture for MVF-sleepNet. Image taken from [88].

In [89], the authors proposed a sleep stage classification method using orthogonal convolutional neural networks. They preprocessed the data with a Hilbert Huang transform. Their model consists of three blocks, with each block consisting of a convolutional layer connected to a squeeze excitation network and a rescaling operation. The convolutional network also has a skip connection to the scaling operation. The convolutional layers are initialized with orthogonal weights, and additional regularization is performed to ensure the orthogonality of the weights throughout training.

So far, we have explored models based on CNN and RNN architectures. There are many other papers that, in general, used these types of architecture, with subtle differences in their modeling through loss function and some prepossessing on the data [13,15,22,27,37,46,58,90–123]. In [68], the authors showed the power of multi-scale modeling to capture fast and slow varying features. To improve upon this approach, many papers used different variations of the multi-scale approach to better capture the features for downstream tasks [14,18,124–126,126–135]. Some papers followed the direction in [84] and added attention modules to their models to improve performance [18,30,80,84,126,126,136–148]. Some authors applied post processing algorithms based on the hidden Markov model (HMM), conditional random fields (CRFs), soft-voting, etc. to improve model performance and

correct unreasonable sleep stage transitions [20,133,140,149–153]. Some researchers used time-frequency information to model the data [29,31,40,131,153–158]. The time-frequency information was either solely used as the input to the model or fused with the raw time series in the latent space. In [40,159], the authors used empirical mode decomposition (EMD) [160] to extract intrinsic features from the input signal by outputting different signals for different frequency ranges which can correspond to different sleep stages. In [161], the authors proposed the multi-modal physiological-signals-based Squeeze-and-Excitation Network with Domain Adversarial Learning (SEN-DAL) to capture the features of EEG and EOG for sleep staging. The SEN-DAL consists of a Multi-modal Squeeze-and-Excitation feature fusion module for adaptively utilizing the multi-modal signals and a Domain Adversarial Learning module to extract subject-invariant sleep features.

The current SOTA approaches make extensive use of contrastive learning paradigms, allowing for unsupervised learning mechanism to encode sample-wise and class-wise, by extension, similarity in the underlying feature vector. Recent efforts [17] have also seen success in using contrastive methods as a training step. Advancements in this field also include modifications to the loss function to either enforce certain model behaviors or to address deficiencies in the data, including class imbalance [51,162,163]. Other techniques have also been proposed including manifold-based techniques [164] and incorporating spatio-temporal features.

4.3. Transformer-Based Models

Another class of modeling includes transformers [65] to perform sleep scoring. Because of transformers' tremendous success in NLP, and the fact that texts are considered time series, they attract attention in the physiological set up as well. In the following section, the papers with transformers architecture are investigated.

The authors of [165] proposed a masked-transformer-based model to learn a signal's features in a self-supervised manner. They found that, although their method helps to improve the performance compared to the baseline, this can be due to the effect of the human subjects that they sampled from. An interesting observation from this paper is the effect of masked length on downstream tasks. They reported that a longer mask will result in better performance in the downstream task. They argue that this could be due to the fact that longer masks force the model to learn more details and nuances about the signal.

In [166], the authors proposed a transformer-based modeling, SleepTransformer, for sleep staging. This modeling allows one to add interpretability to the model through the attention layers to have more confidence in decision making and also to be integrated in clinical settings. At epoch level, through attention, the model learns to pick the most relevant parts of the signal. Since this model can be applied at sequence level, it can measure the influence of the neighboring epochs on the target epoch for decision making. We will explain more about the interpretability in Section 5. The model architecture is shown in Figure 9. This model takes time-frequency images as the input, and processes each epoch by a standard transformer block. The output for each epoch is then concatenated and fed into another transformer for fusion. Then, multiple FC layers for sleep staging are utilized.

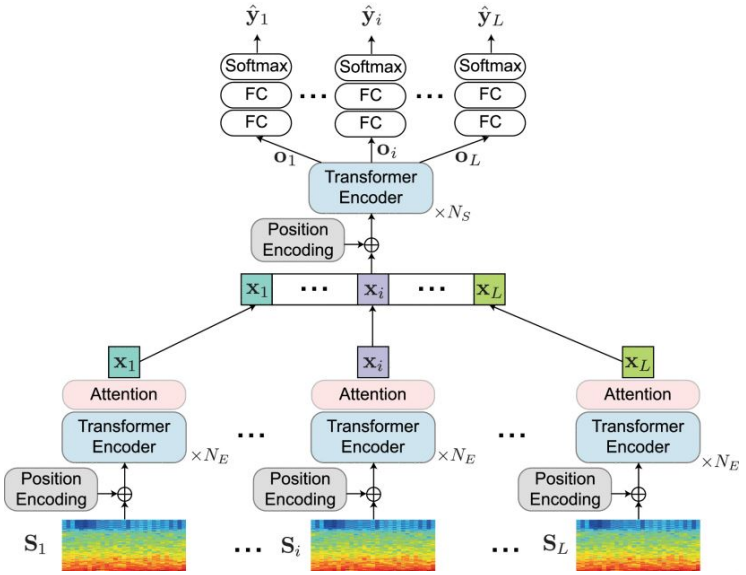


Figure 9. Model architecture for SleepTransformer. Image taken from [166].

Another transformer-based model is composed of a multi-Scale CNN block with intra-modal and cross-modal attention [167]. The main difference (besides the model architecture) between this work and ref. [166] is the cross-modal relationship, which ref. [166] fails to examine. The multi-scale CNN block learns optimal feature representations by considering local and global features. The authors proposed a novel cross-modal transformer model encoder architecture to learn intra-modal temporal attention between time steps within a feature embedding of a modality and cross-modal attention to capture relevant information between each modality. In case of the sequence-based classification, the cross-modal transformer has another block to learn inter-epoch relations. The sequence-based model is depicted Figure 10. More details about this model can be found in Appendix A.3.

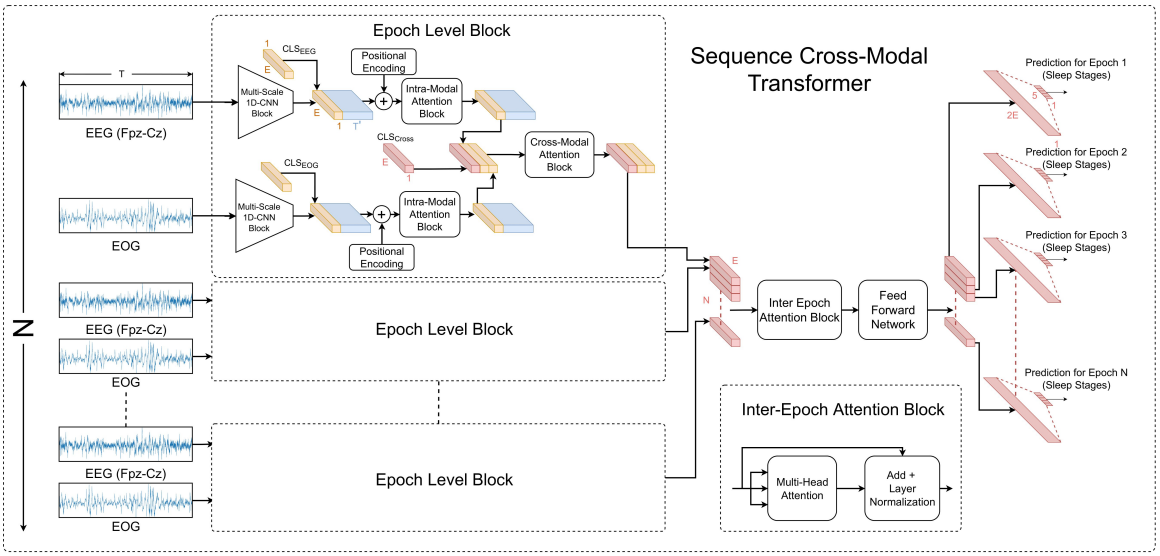


Figure 10. Model architecture for cross-modal transformer. Image taken from [167].

In [168], the authors proposed a multi-modal attention network (MMASleepNet) to perform sleep staging. This model consists of multi-branch feature extraction (MBFE) and an attention-based feature fusion (AFF),and classification layer, as shown in Figure 11. The MBFE takes different modalities and extracts features by simple convolutional layers. The AFF module contains a modal-wise Squeeze-and-Excitation [169] block to adjust the

weights of modalities with more discriminative features and a transformer encoder (TE) to generate attention matrices and extract the inter-dependencies among multi-modal features.

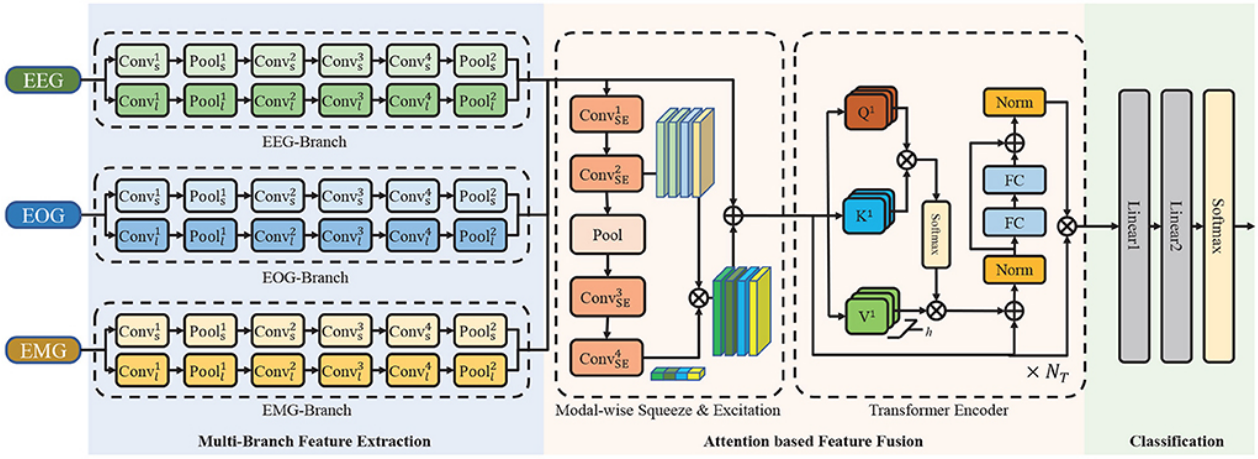


Figure 11. Model architecture for MMASleepNet. Image taken from [168].

In [170], the authors introduced ENGELBERT. This model introduces overlapping attention mechanisms, which helps with efficient matrix formulation, aggregation, and reduces the critical quadratic computational complexity to linear. As a regularization mechanism, the authors applied non-linearity in a dimensionally compressed space for each attention layer. The model architecture is shown in Figure 12. The model has a CNN feature extraction layer that is followed with linear dimensionality expansion to be fed to the attention-based feature fusion layer, which helps increase the representation capacity. One major difference compared to a regular transformer encoder is that the arrangement of the expansion and compression blocks are reorganized, as shown in Figure 13.

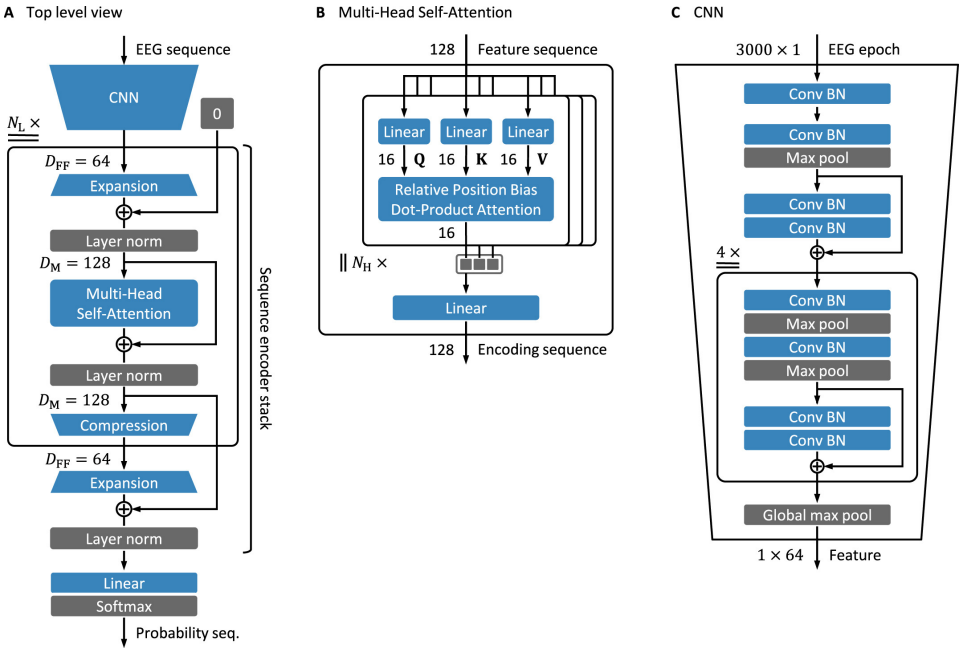


Figure 12. Model architecture for ENGELBERT. Image taken from [170].

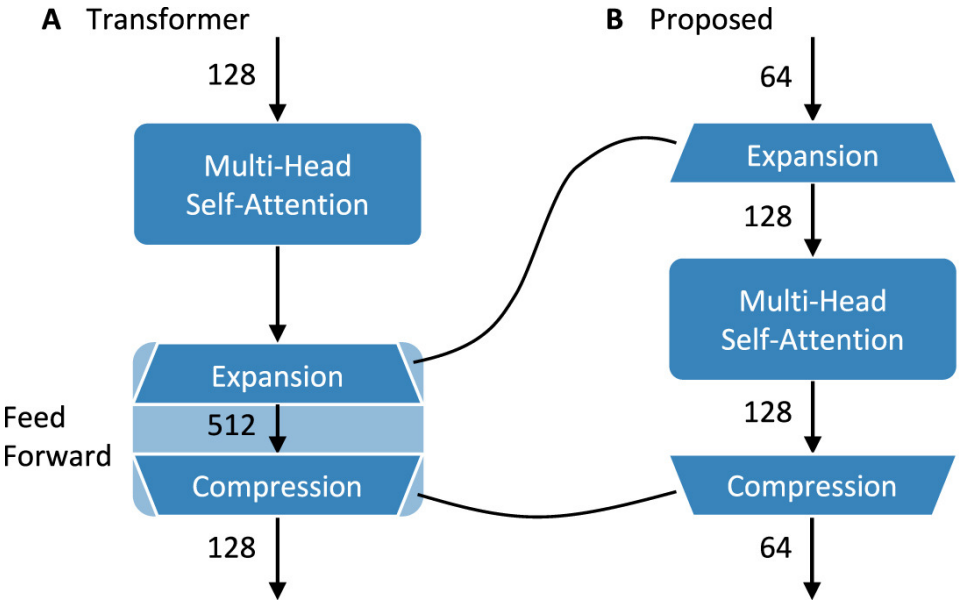


Figure 13. Difference between transformer and ENGELBERT. Image taken from [170].

4.4. Contrastive Learning

Another class of modeling for sleep classification is contrastive learning. In contrastive learning, the goal is to learn a proper representation of the signal by itself (self-supervised) or by having access to partial labels (semi-supervised) to optimally perform downstream tasks such as classification, anomaly detection, etc.

In [171], the authors proposed a novel methodology for fusing multi-channel EEG to provide salient information to a sleep staging model. The logic behind their methodology is that each electrode used for PSG captures a portion of brain activity from different regions of the brain. By structuring the channels as an undirected graph and learning adjacencies, they are able to find strongly related brain activity in terms of both spatial and temporal features. In this paper, the model learns the adjacency matrix, which is utilized to learn graph parameters. More details about the approach are presented in Appendix A.4. The main contributions of this paper are the adaptive adjacency matrix learning and the use of spatial and temporal features for sleep staging.

In [172], a contrastive representation learning model named SleepPriorCL based on prior knowledge is proposed. The goal of this model is to learn a meaningful semantic representation for similar samples in self-supervised manner, without using any labels. It is known that each sleep stage occupies a certain frequency range. The authors leveraged this fact and calculated the energy for each of these frequency packets, and used them as prior knowledge for training. Since similar signals have similar energy, the absolute value of their difference can serve as a similarity measure. The model has to put more weight towards similar samples and less towards dissimilar ones. This is achieved by using the absolute value of the difference between energies of the sample to calibrate the temperature in a contrastive loss function. Moreover, to ensure there are multiple positive (samples) in each mini-batch, an augmentation on each signal is performed in the paper. It should be mentioned that this formulation through prior knowledge helps the model not be biased to positive samples that are just generated through augmentation, and also considers other positive samples in the mini-batch that are not the result of augmentation.

The authors of [173] proposed a new model, called CoSleep, to extract semantic representation for physiological signals through multi-view modeling. The architecture is shown in Figure 14. The CoSleep technique focuses on uncovering connections among instances within the same semantic class, regardless of their temporal proximity. It achieves this by attracting multiple positive instances simultaneously, enhancing representation modeling at the semantic-class level. The approach employs a multi-view co-training mechanism to identify positive instances, drawing from both temporal and spectral perspectives. The

technique extends the Dense Predictive Coding (DPC) framework with a dedicated memory module, which enlarges the pool of negative candidates. This expansion improves DPC training and the extraction of positive instances. More details about model are presented in Appendix A.6.

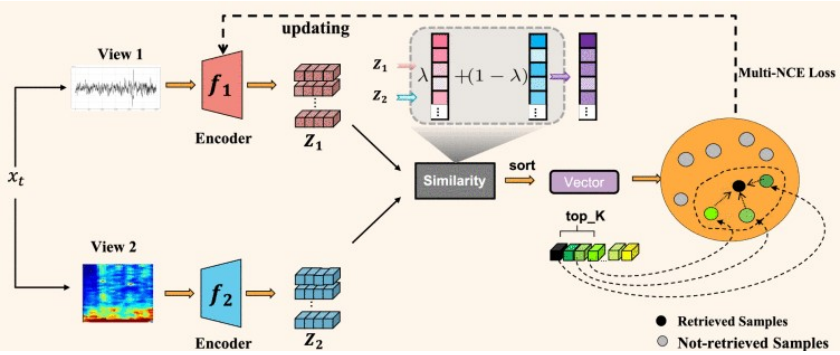


Figure 14. The proposed framework for CoSleep. Image taken from [173].

In [174], the authors proposed SleePyCo for EEG-based sleep stage scoring. The model uses a multi-scale feature pyramid and contrastive learning to learn intra-class and inter-class features, using them to better separate sleep stages in feature space. The feature pyramid consists of CNN layers with varying kernel sizes. These layers are connected laterally to linear projectors to match the dimensions of all of the feature vectors. The resulting feature matrix is then fed to a transformer network to perform the classification. The backbone is pre-trained using a contrastive learning strategy. The second stage is learned through multi-scale temporal context learning, freezing the pyramid network during training to ensure the retention of features learned during the contrastive learning stage. Early stopping was used in both stages to ensure parsimony.

The authors of [175] proposed a multi-task contrastive learning strategy for semi-supervised sleep recognition (MtCLSS). In Figure 15, there are two similar branches, one for transformed signals and the other for the original signals. The model has two architecturally similar branches like a Siamese network. Branch 1 receives the raw EEG signal and Branch 2 receives a transformed EEG signal. Branch 2 predicts sleep stages and transformations. Branch 1 only predicts sleep stages. The features of the two branches are used to perform contrastive learning. The authors argue that the transformation identification helps with exploring the unlabeled data better and also improves the discriminative abilities and robustness of the model.

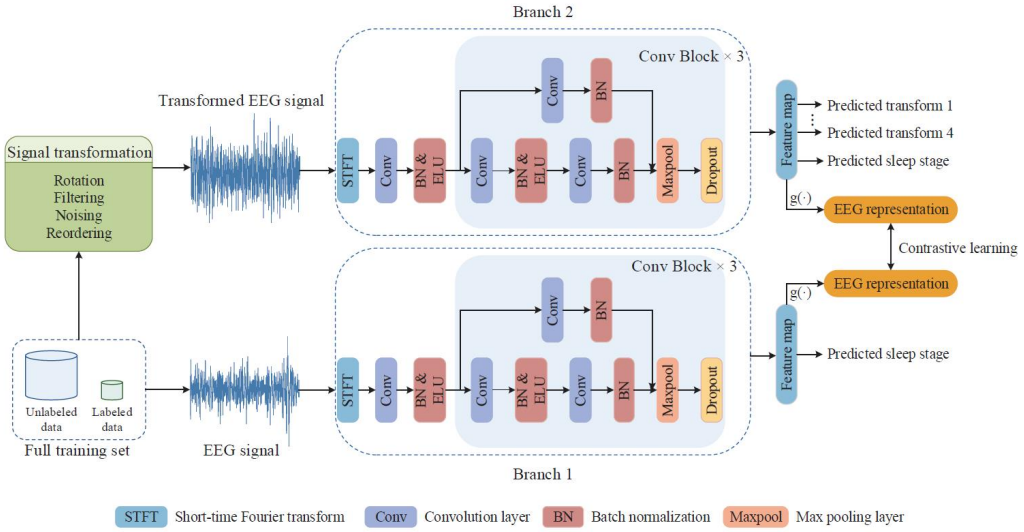


Figure 15. Model architecture for MtCLSS. Image taken from [175].

In [176], the authors proposed a semi-supervised sleep staging approach called co-attention meta sleep staging network (CMS2-Net). This framework aims to mitigate inter-class disparity and intra-class selection in sleep scoring. To resolve the inter-class problem, the authors proposed a co-attention module with second-order statistics to improve latent feature representation. To solve the intra-class problem, a triple-classifier is introduced to resolve the domain shift and re-parameterization.

The authors of [177] proposed to use both expert features and DNN features. They proposed a contrastive-based cross-attention model to predict sleep apnea through ECG signals. The cross-attention module is used to fuse expert knowledge and deep features, which helps with weighting the important features. The authors also performed classification with the contrastive task, which helps with better separation in feature space to enhance the downstream classification task. In order to achieve the goal above, a new hybrid loss has been proposed, which is composed of contrastive loss and cross-entropy. In Figure 16, it can be seen how both raw signal and expert features are fed to their corresponding models to extract features to input to the cross-attention layer to create a fused feature vector. Then, the fused feature vector goes through the projection for contrastive loss and it is passed to the classifiers for apnea detection.

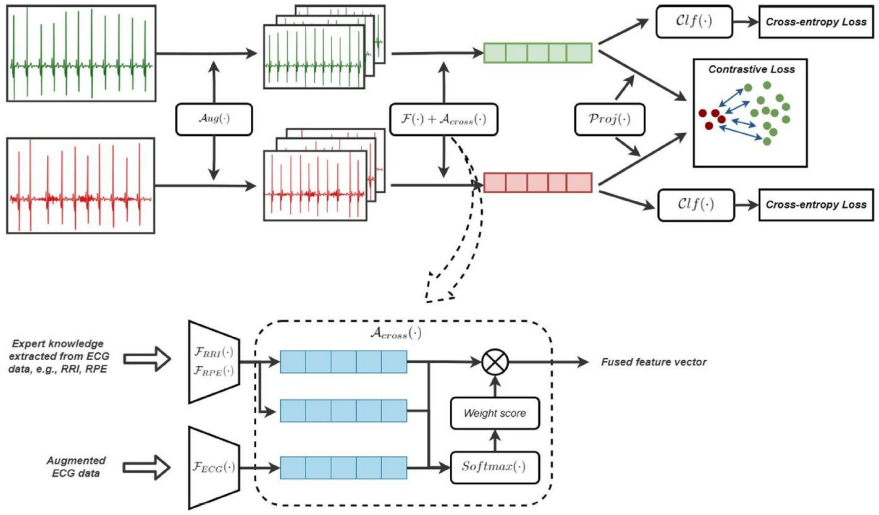


Figure 16. Model architecture for [177]. Image taken from [177].

In [178], the authors proposed the SCL-SSC (Supervised Contrastive Learning for Sleep Stage Classification) model, which combines contrastive learning with classification tasks shown in Figure 17. This task is performed in two steps: (1) feature learning and (2) classification. The feature learner block is trained separately to disentangle positive and negative samples from each other on the representation space, which translate to having more distance between samples from different classes and less distance between samples from the same class in Euclidean space. This is achieved by using the following triplet loss

$$L(R_a, R_p, R_n) = \max(0, \alpha + d(R_a, R_p) - d(R_a, R_n))$$

where R_a, R_p, and R_n are anchor, positive, and negative feature representations, respectively. The function d is the Euclidean distance. The variable α is a bias term to minimize the distance between the same sleep stages and maximize the distance otherwise. It is expected to have $d(R_a, R_p) > d(R_a, R_n)$ for a meaningful learning. The authors used a similar architecture as in DeepSleepNet [68] for the feature block (f_θ). For the classification portion, the authors used an encoder–decoder model with an attention mechanism, shown in Figure 18. This model takes in the features from the pre-trained feature block and feeds it to a sequence of RNNs to build a context vector to be fed to another RNN layer, which then goes through an attention block. This sequential modeling would encapsulate information

from adjacent epochs to improve embedding space and overall model performance. The authors used the loss function as in [84] for classification.

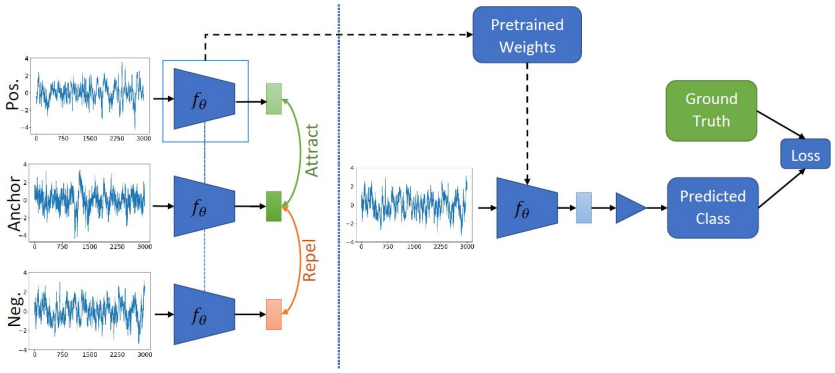


Figure 17. Model framework for SCL-SSC [178]. Image taken from [178].

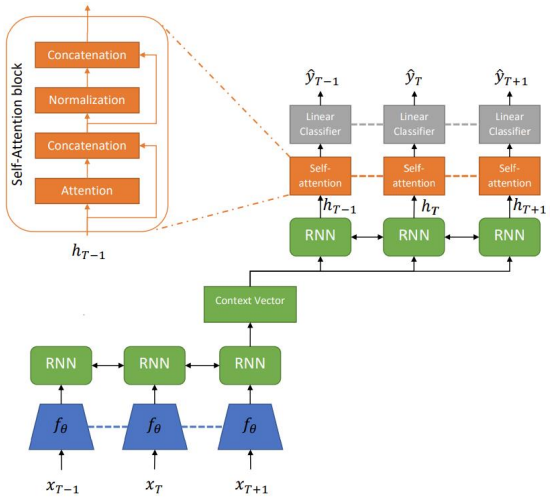
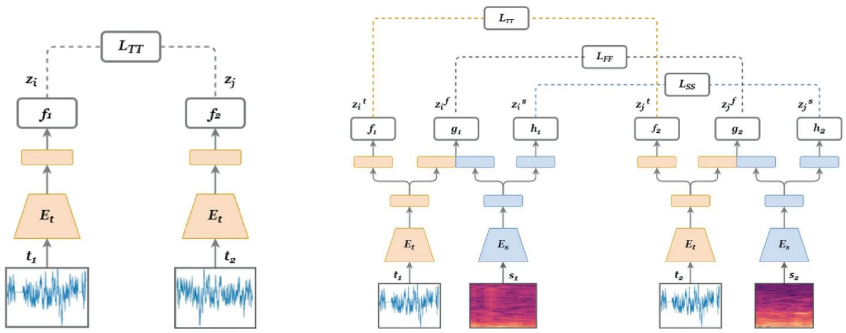


Figure 18. An encoder–decoder-based classifier with self-attention. Image taken from [178].

In [56], the authors proposed a novel multi-view self-supervised method (mulEEG) for unsupervised EEG representation learning. The goal is to design a setup where the model can effectively utilize the complementary information in multiple views to improve the representation space. To achieve this goal, the authors introduced the diverse loss to encourage the model to use complementary information from different views. The model is shown in Figure 19. Apart from some augmentation techniques used in this paper, the main difference compared to CoSleep [173] is how the authors set up their loss function to impose similarity on the positive and negative samples. More details about the model can be found in Appendix A.5.



(a) Single-View (b) mulEEG

Figure 19. Model architecture for mulEEG. Image taken from [56].

5. Interpretability

In Section 4, we mentioned that models normally have thousands, if not millions, of parameters. On top of a big number of parameters, non-linear activation functions add more complexity. Due to these facts, deep learning models are referred to as black-box models because there is no direct way to interpret how exactly information is processed through layers of a network. This is one of the major drawbacks of deep learning models. Another major issue with these models is their vulnerability to adversarial attacks, which can affect the model's output drastically and lead to wrong results. Therefore, it is crucial to study these models thoroughly to reduce the uncertainty in their decision making and also make them more robust towards adversarial attacks. Especially in medical settings where clinicians work with patients, the explainability of the models and how the information goes from the input to the output becomes very important for reliable and trustworthy decision making. Interpretability will help the medical society to accept these models with more trust and certainty. Moreover, explainability can help with diagnosing the portions of the model that contribute the most to wrong decision making, which can later be fixed. We will go through the proposed models that provide some form of interpretability for sleep staging models.

The authors of [179] proposed Sleep staging via Prototypes from Expert Rules (SLEEPER). SLEEPER combines deep learning models with expert-defined rules via a prototype learning framework to generate simple interpretable models such as shallow decision trees and logistic regression models. In particular, SLEEPER utilizes sleep scoring rules and expert defined features to derive prototypes which are embeddings of polysomnogram (PSG) data fragments via convolutional neural networks. The final models are still simple interpretable models like a shallow decision tree or logistic regression defined over those phenotypes. The SLEEPER framework is shown in Figure 20.

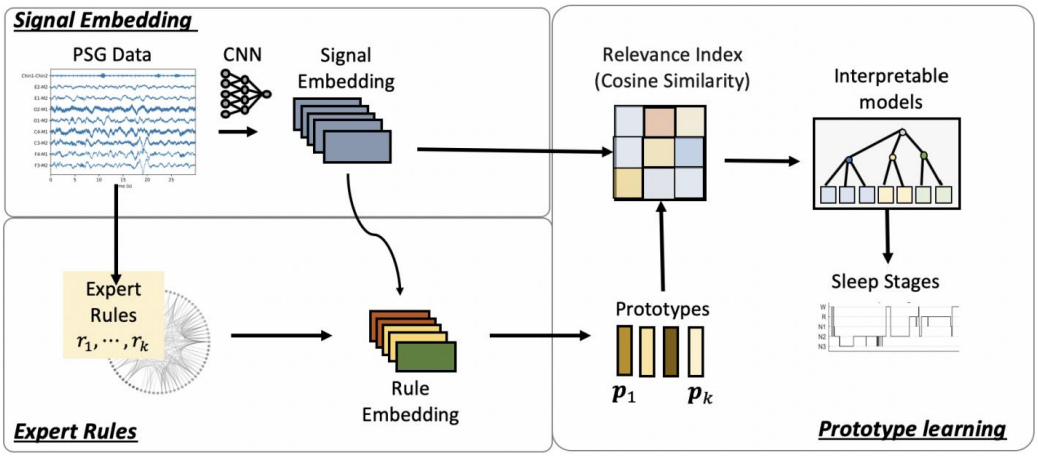


Figure 20. The proposed framework for SLEEPER. Image taken from [179].

In [180], the authors proposed an ablation-based interpretability for a multi-modal CNN-based model. This approach, instead of zeroing out the modalities (masking with zeros), replaces the signals with noise that normally appears in the measurements. They used a 40 Hz sinusoid and a Gaussian noise with mean 0 and standard deviation of 0.1. Before performing these modifications, the F1 score is recorded when all the modalities are intact. After replacing the modalities with the noise mentioned above, the difference in F1 score is reported, as shown in Figure 21. The authors also added a zeroing-out method for comparison.

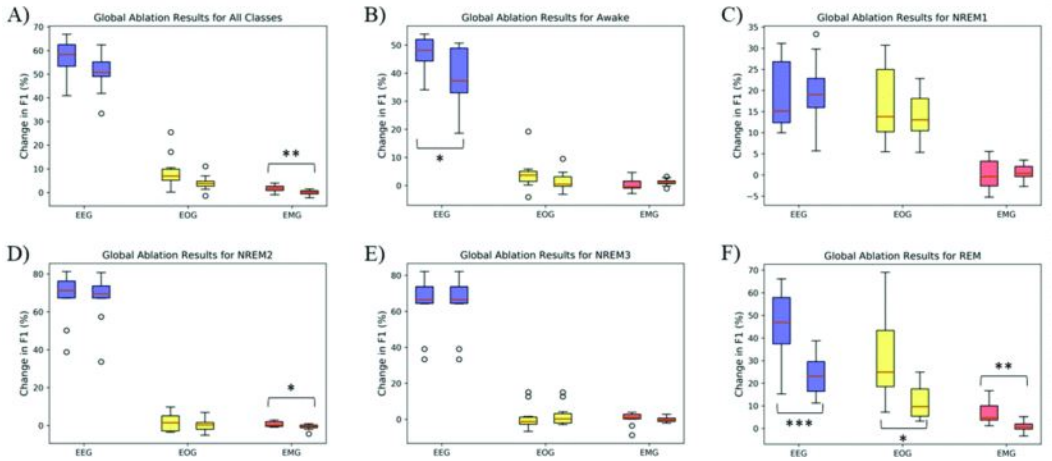


Figure 21. F1 score change for the ablation study. Image taken from [180]. One asterisk (*) indicates p-value smaller than 0.05 ($p < 0.05$). Two asterisks (**) indicate ($p < 0.01$). *** $p < 0.001$.

Figure 21 shows the change in F1 score over all and individual classes. For F1 over all the classes, the EEG shows the most significance. This change can be associated with NREM2 and NREM3, which have the most change. EEG plays an important role for the NREM2, NREM3, and Awake stages. EOG and EMG have more importance for the NREM1 and REM stages. For REM class, both EEG and EOG play an important role in high F1 scores. In general, the EMG does not seem to contribute as much as other modalities. The authors mentioned that the model is originally designed for an EEG signal, which might not be able to properly show the effect of EMG on the classification accuracy.

In [166], the authors proposed SleepTransformer (explained in Section 4.3). They introduced two mechanisms for interpretability: (1) **EEG attention map,** in which the attention score for an epoch is collected and the effect of different regions of an EEG signal can be observed. The goal is to see where the model pays more attention to achieve high accuracy; (2) **Epoch influence bar chart**, in which the model has access to multiple epochs. The influence of each epoch on neighboring epochs can be quantitatively studied. This resembles how clinicians make decisions for sleep scoring. In Figure 22, the heat map for epoch attention score and sequence-level distribution for attention scores is shown. A sequence of 11 epochs is presented for analysis. Figure 22a refers to the epoch influence bar chart, where the influence of other epochs on the target epoch is shown. The arrow at the bottom shows the current epoch. Figure 22b shows the learned EEG features for epoch representation (the time signal is reconstructed by Inverse STFT). In Figure 22c, the epoch-level attention score is depicted on the ground-truth EEG signal, and Figure 22d shows the attention score for the spectrogram.

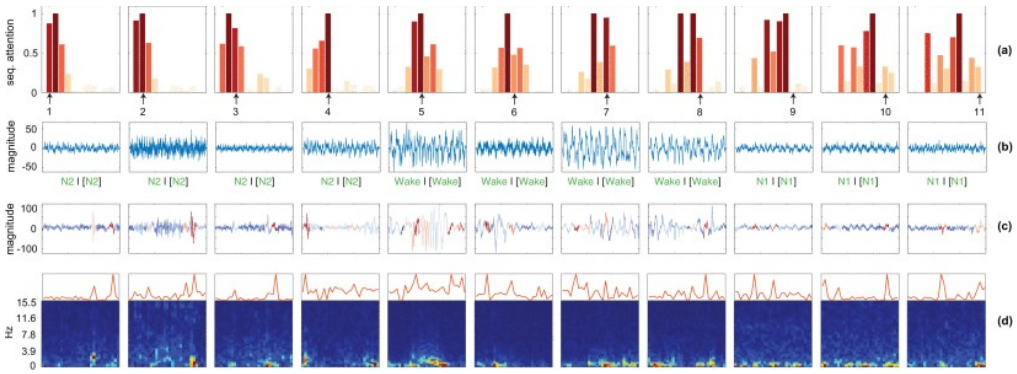


Figure 22. SleepTransformer attention visualization. Image taken from [166].

It can be seen at the epoch level that the model pays attention to the key parts of the signal. These attention scores can be traced back to each sleep stage frequency range in

which they are present. At sequence level, the attention weights control the contribution of each epoch to perform the classification. In Figure 22b, for stage N2, all the sequences to the left of the target signal contribute the most to the final decision. Interestingly, for the wake stage, the concentration is in the middle. Lastly, for the N1 stage, the attention scores disperse, which can be related to the fact that this signal resembles the N2 and Wake stages. For more examples, please refer to [166].

The authors of [167] proposed a cross-modal transformer, which enables them to use the attention to learn and interpret: (1) intra-modal relationships, (2) cross-modal relationships, and (3) inter-epoch relationships. The intra-modal relations are similar to [166], with a difference, where they introduce the CLS_c token, as discussed in Section 4.3. In a similar procedure, the relationships between modalities are interpreted by the scaled dot-product attention of the representations in the output of the cross-modal attention block corresponding to CLS_c of each modality. Finally, the inter-epoch relationships are calculated by scaled dot-product attention between the CLS_{cross} of each epoch.

In Figure 23, the visualization for interpretability is demonstrated. Five epochs are present in this experiment. Inter-epoch attention plots show the relationships between the five epochs (similar to [166]) and how much impact they have on the current epoch for decision making. The cross-modal attention plot shows how much contribution each modality is providing for decision making. It can be observed in some cases that EEG has more relevance, while in others, EOG dominates the importance for decision making. Intra-modal attention plots are shown as a heat map on the original signals. The darker the shade, the more the model pays attention to that time section. It can be observed that the model highlights the important parts of the signal, which indicates which parts are most relevant to that sleep stage. Please refer to [167] for more examples.

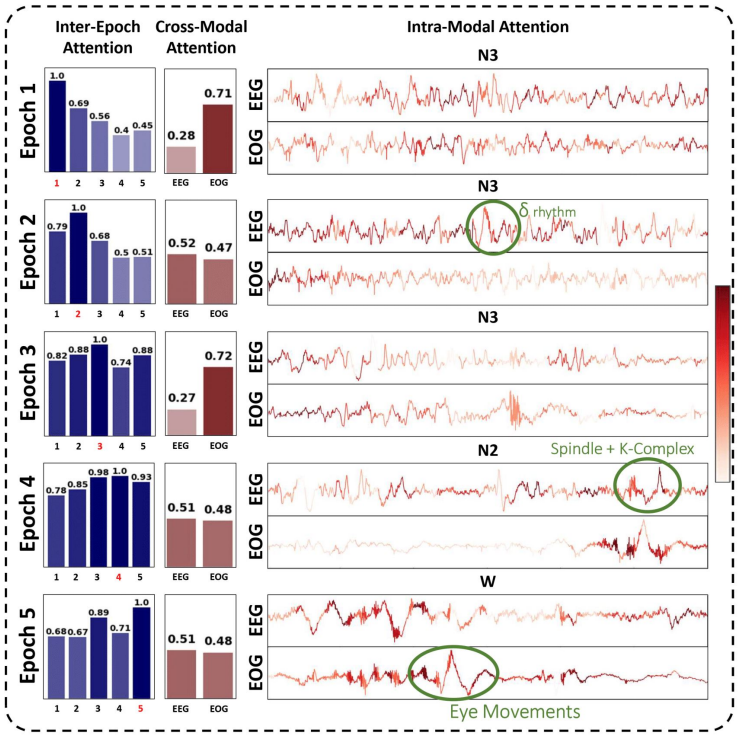


Figure 23. Visualization of inter-epoch and intra- and cross-modal relations. Image taken from [167].

In [159], the authors used empirical mode decomposition (EMD) [160], which is a data-driven algorithm that decomposes the signal into intrinsic mode functions (IMFs). The IMFs are calculated by the noise-assisted bivariate empirical mode decomposition (NA-BEMD) algorithm described in [159]. The EMD is applied to the model input in the first layer, followed by an attention module. The rest of the model is composed of CNN and RNN layers. Each IMF represents a frequency band that is helping with interpretability

of the model by calculating the attention score. Higher attention score to the IMF shows that components of the frequency band of that IMF have more significance towards the sleep staging. The authors showed that IMF2 contains beta and alpha waves, IMF3 alpha and theta waves, IMF4 little alpha, theta, and delta waves, and IMF5 and IMF6 mostly delta waves. In Figure 24, the mean and standard deviation of the attention weights of each IMFs for sleep stages are depicted. IMF2 has the widest spectrum (as expected) and highest mean of attention weight in all stages. IMF2 is the largest for the Wake stage. The attention weights for IMF3 and IMF4 decreased for the Wake stage, which corresponded to low frequency being less present. It can be observed that there are meaningful correlations between the IMFs and the attention weights for each sleep stage.

As shown above, researchers are using different means to be able to interpret the results from the model's output. Next, we will discuss a few other approaches. In [181], the authors used the LIME library [182] on top of their model for interpretability. In [127], the authors proposed something similar to the layout in [68] for model explainability. The model is trained in a semi-supervised sense, which is different from [68], which is completely supervised. In [122], a methodology using linear models was used, achieving similar performance to SOTA deep learning. The techniques used are easily interpreted, yielding greater insight on what factors are most relevant to the analysis. In [183], a decision-tree-based method is proposed for performing sleep stage classification with reasonably accurate results. It is easy to see what factors contribute the most to the classification. This along with [46,122] are examples of verifiable and interpretable techniques. As presented, most of the interpretable approaches such as [24,108], which show significant insight on how the model perceives the data, can help the clinicians trust DNNs to be incorporated in clinics for sleep disorder detection.

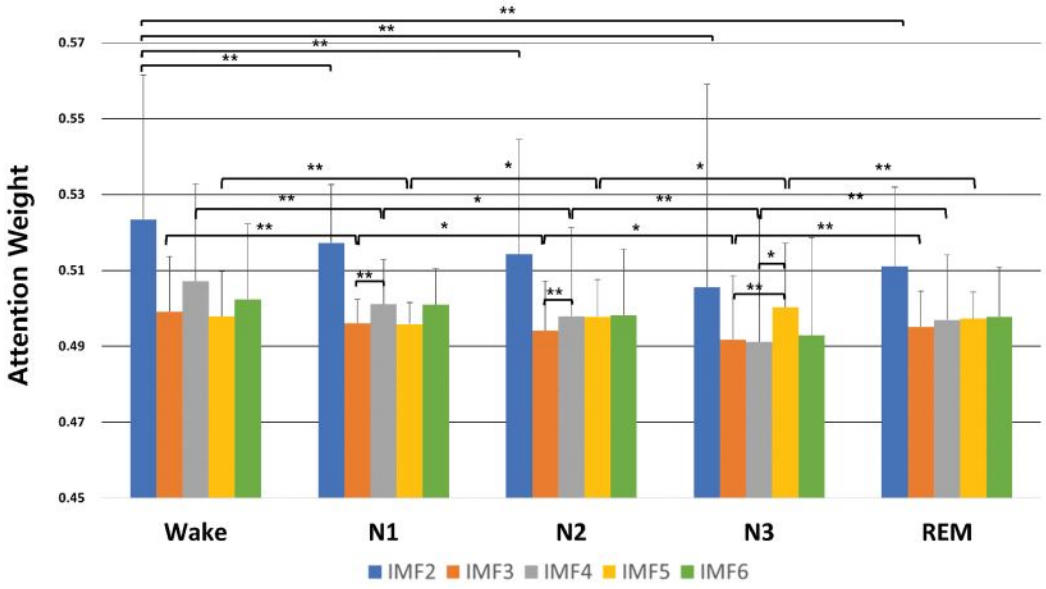


Figure 24. Attention weights of each IMF for sleep stages. One asterisk (*) indicates *p*-value smaller than 0.05 ($p < 0.05$). Two asterisks (**) indicate ($p < 0.01$). Image taken from [159].

6. Discussion

6.1. General Observations

In this section, as shown in Table 3, we summarize the SOTA models from 2018 to 2023 in chronological order with the corresponding reported results for different datasets. In Table 4, we summarize the backbone architectures, number of parameters, and whether interpretability is incorporated in their scheme for all the models presented in Table 3.

Of the methods and techniques described in prior sections, ENGELBERT [51,170], SleepExpertNet [87], MaskSleepNet [163], Xsleepnet [78], and RobustSleepNet [77] achieve highest performance in terms of their respective performance score for most popular

datasets, i.e., EDF-20, EDF-78, MASS, SHHS, and CAP, respectively. Techniques using attention and transformer-style models achieve high performance in this domain, demonstrating the importance of context modeling in classification for this task. The top performing techniques are enumerated in Table 3. Aside from performance, advancements have been made in terms of improving the trustworthiness of these models through interpretability, which is one of the major limiting factors to adoption of these techniques in clinical settings.

From Table 3, we can see that different types of architecture and modeling techniques have been proposed. In early attempts, CNNs and RNNs were utilized for sleep stage classification. In the following years, new modeling and techniques, such as transformers, GNNs, and contrastive learning were explored. Due to the abundance of data, learning how to capture proper features for downstream tasks is critical. The contrastive learning and SSL show strong evidence that these techniques, if applied correctly, can be extremely efficient in feature extraction, as demonstrated in [174,176].

The model performances can vary between different datasets, which can be attributed to data size, data quality, age of the participants, health, and other factors that can affect the data collection. In Table 2, we show the data size for some of the popular datasets. For the SleepEDF dataset, most models have relatively good performance, which can be due to the sufficient size and the quality of the data. The same goes for the MASS and SHHS datasets, which include a large number of high quality recordings. But this is not always the case, as we can see for the CAP dataset. This dataset is very large and it contains many recordings for long hours (please refer to Section 3.1 for more information). Modeling such a dataset is very challenging due to the variability of the data for each participant, which can be directly correlated to each participant's bio-markers that may not be present in other participants, and due to data imbalance. In Section 6.2, we can observe a quantitative comparison between the top performing models for the EDF-20 and EDF-78 datasets.

Another major difference in performance can be attributed to the models' design. For the SleepEDF (EDF-78) datasets, SleepExpertNet [87] has the best performance. As described in Section 4.2, this model is consists of different modules to efficiently extract temporal and sequential information. The first layer is a temporal CNN, which extracts the temporal feature that is fed to a positional encoder. Then, the sequential information is fed to an attention layer to capture the relationship between feature components. The output is processed via LSTM layers for classification. MaskSleepNet [163] has the next best model performance. This model has a multi-scale CNN layer followed by a multi-head attention layer. This approach uses modality masking for training to adapt the model to different scenarios where all modalities, such as EEG, EOG, and EMG, might not be available all at once. This training pushes the model to learn salient information and relationships between different modalities implicitly. Both these models can handle multiple modalities at the same time. The major difference between these two is in their architecture and how they are trained. It should be mentioned that MaskSleepNet [163] has the best performance for the MASS dataset. The second best performing model is BSTT [184]. This is a transformer-based approach that utilizes the Bayesian inference method. This model utilizes two transformer blocks to encode temporal and spatial information based on a raw EEG signal as input. XSleepNet [78] is the best performing model for SHHS dataset. This model, as described in Section 4.2, is a sequential multi-view model that takes in spectrogram and raw EEG signals as the input. The authors proposed to use gradient blending to balance the generalization/overfitting trade-off for classification. The second best model is SeqSleepNet [25], which is a sequential hierarchical RNN model. This model accepts a spectrogram as input to perform classification. These two models have many similarities, but the most prominent difference is that XSleepNet has a completely different training technique.

In Figure 25, for four datasets, the number of parameters (x-axis) and their respective reported metric (y-axis) are plotted. These plots do not contain all the models reported in Tables 3 and 4, because we just know the complexity of some of the models. It should be mentioned that not all the reported metrics in the figure are the same; they are F1-score

and kappa score. We can observe that smaller models can perform on par with bigger models or even better. This shows that better training schemes, as we explained above in the modeling sections, can be very efficient.

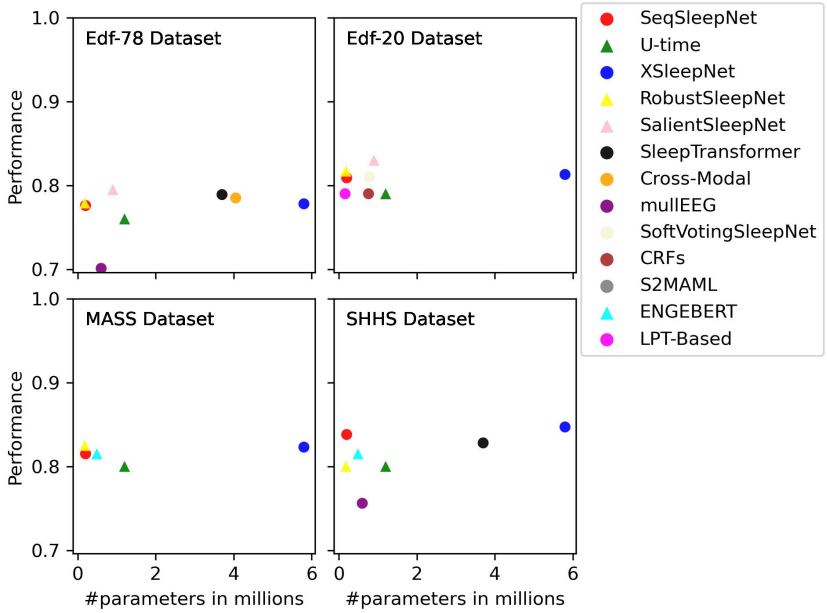


Figure 25. These figures show the complexity vs. accuracy (on 4 datasets) for the models with parameters reported in Table 4. Here, ∘ refers to kappa score and △ refers to F1-score. The models' performances that are not visible are due to the absence of the performance for that particular dataset.

6.2. Quantitative Observations

Since EDF-20 and EDF-78 are the most commonly used datasets for reporting performance, we provide a more quantitative discussion about the models reporting on these datasets. Furthermore, the kappa score was utilized for most models. Hence, we focus on this metric. Given that s_m is the kappa score for model m, s^{top} is the top score for a dataset, and s^{base} is the score for a baseline model, we define the percentage of improvement of the top model over a model m as:

$$p_m^{top} = \frac{(s^{top} - s_m)}{s_m} \times 100\%, \tag{1}$$

and the percentage of improvement of model m over the baseline model as:

$$p_m^{base} = \frac{(s_m - s^{base})}{s^{base}} \times 100\%. \tag{2}$$

We consider the SeqSleepNet [25] model as our baseline since this is the earliest model (published in 2018) that was considered.

For the EDF-20 dataset, we observe that the top ranking model is the ENGELBERT [170]. This model had a $p_m^{base} = 1.7\%$ improvement over the baseline. The second best model was CoSleepNet [51] with a $p_m^{base} = 1.1\%$. The improvement of the top model over CoSleepNet [51] is $p_m^{top} = 0.6\%$. ENGELBERT also reported results for EDF-78 with an improvement of $p_m^{base} = 2.3\%$. Unfortunately, the CoSleepNet [51] model did not report results for the EDF-78 dataset, so it could not be compared against the other models. All other models had either a negligible (around 0.1%) or negative improvement over the baseline.

For the EDF-78 dataset, we observe that the top ranking model is SleepExpertNet [87]. This model had a $p_m^{base} = 12\%$ improvement over the baseline. The second best model was the MaskSleepNet [163] with a $p_m^{base} = 9\%$. The improvement of the top model over MaskSleepNet [163] is about $p_m^{top} = 3\%$. Unfortunately, the MaskSleepNet [163] and

SleepExpertNet [87] models did not provide results for the EDF-20 dataset so they could not be compared to the top ranked models in this dataset. Nevertheless, MaskSleepNet [163] did achieve the top ranking for the MASS dataset. The third highest score model is NAMRTNet [185] with $p_m^{base} = 4\%$ improvement over the baseline. Other models (CrossModel [167], SleepCo [174], SleepTransformer [166], and IITNet [79]) had improvements over the baseline on the EDF-78 dataset between 1% and 2%.

These results (and the ones for EDF-20) indicate that there is not a single model that outperforms the rest in all datasets. However, the models do show an improvement between 1% and 12% over the last five years.

Overall, we can observe that, despite different model architectures and numbers of parameters, there is no guarantee that the models would perform the best over all the datasets. It seems training and data manipulation techniques for models to extract and adapt to the dataset have a bigger role in model efficiency in terms of performance than other factors. This evidence can be seen in Figure 25 and Tables 3 and 4. In the subsequent section, limitations of current approaches are discussed.

Table 3. Comparing SOTA models. The main metric in this table is kappa score. Superscripts *a* and *f* refer to accuracy and macro-F1 score when the kappa score was not reported. The highest scores in terms of macro-F1 and kappa scores are highlighted in bold.

Model	Year	EDF-20	EDF-78	MASS	SHHS	CAP	DRM-SUB	ISRUC	SVUH-UCD	WSC	Other
SeqSleepNet [25]	2018	0.809	0.776	0.815	0.838	–	–	–	–	–	–
U-Time [82]	2019	0.79 ^{f}	0.76 ^{f}	0.80 ^{f}	0.80 ^{f}	0.68 ^{f}	–	0.77 ^{f}	0.73 ^{f}	–	0.85 ^{f}
IITNet [79]	2019	0.78	0.79	–	0.81	–	–	–	–	–	–
GraphSleepNet [171]	2020	–	–	0.834	–	–	–	–	–	–	–
Xsleepnet [78]	2020	0.813	0.778	0.823	**0.847**	–	–	–	–	–	–
RobustSleepNet [77]	2021	0.817 ^{f}	0.779 ^{f}	0.825 ^{f}	0.80 ^{f}	**0.738** ^{f}	–	–	–	–	–
CCRRSleepNet [132]	2021	0.78	–	–	–	–	–	–	–	–	–
SalientSleepNet [147]	2021	0.83 ^{f}	0.795 ^{f}	–	–	–	–	–	–	–	–
SleepTransformer [166]	2021	–	0.789	–	0.828	–	–	–	–	–	–
Cross-Modal [167]	2022	–	0.785	–	–	–	–	–	–	–	–
MtCLSS [175]	2022	0.80	–	–	–	–	–	–	–	–	0.74
MAtt [146]	2022	-	–	–	–	–	–	–	–	–	0.7471 ^{a}
PearNet [85]	2022	0.793 ^{f}	0.753 ^{f}	–	–	–	–	–	–	–	–
CMS2-Net [176]	2022	–	–	–	–	–	–	–	–	–	0.71
SleepContextNet [59]	2022	0.79	0.76	–	0.81	0.71	–	–	–	–	–
mulEEG [56]	2022	–	0.6850	–	0.7366	–	–	–	–	–	–
Liu et al. [18]	2022	0.862 ^{f}	0.852 ^{f}	–	0.835 ^{f}	–	–	–	–	–	–
MVF-SleepNet [88]	2022	–	–	–	–	–	–	0.795	–	–	–
TrustSleepNet [30]	2022	–	–	–	–	–	–	–	–	–	0.82
SoftVotingSleepNet [152]	2022	0.81	–	–	–	–	–	–	–	–	–
Kim et al. [128]	2022	0.838 ^{f}	–	–	–	–	–	–	–	–	–
CRFs [149]	2022	0.79	–	–	–	–	0.76	–	0.66	–	–
IDNN [159]	2022	0.81	0.76	–	–	–	–	–	–	0.75	–
CAttSleepNet [138]	2022	0.78	0.74	–	–	–	–	–	–	–	–
ISENet [140]	2022	–	0.79	–	–	–	–	–	–	–	–
Bi-RNN [102]	2022	0.8404 ^{a}	–	–	–	–	–	–	–	–	–
S2MAML [186]	2022	–	–	–	–	–	**0.821** ^{f}	**0.888** ^{f}	**0.904** ^{f}	**0.863** ^{f}	–
Pei et al. [103]	2022	–	–	–	0.76	–	–	–	0.58	–	–
SleepExpertNet [87]	2022	–	**0.87**	–	–	–	–	–	–	–	–
CDNN [96]	2022	–	–	–	–	–	–	0.854	–	–	0.734
Van Der Donckt et al. [122]	2022	0.812	0.766	0.803	–	–	–	–	–	–	–
SleepyCo [174]	2022	–	0.787	0.811	0.83	–	–	–	–	–	–
ENGELBERT [170]	2022	**0.823**	0.794	0.799	–	–	–	–	–	–	–
CoSleepNet [51]	2023	0.8181	–	–	–	–	0.7693	0.715	–	–	0.674

Table 3. *Cont.*

Model	Year	EDF-20	EDF-78	MASS	SHHS	CAP	DRM-SUB	ISRUC	SVUH-UCD	WSC	Other
LPT-Based [81]	2023	0.79	–	–	–	–	–	–	–	–	–
MaskSleepNet [163]	2023	–	0.847	**0.847**	–	–	–	–	–	–	0.812
SHNN [127]	2023	–	0.7051	–	–	–	0.6955	–	–	–	–
Siamese AE [187]	2023	0.79	–	0.81	–	–	–	–	–	–	–
BSTT [184]	2023	–	–	0.8437	–	–	–	0.7678	–	–	–
NAMRTNet [185]	2023	–	0.808	–	–	–	–	–	–	–	–

Table 4. Building blocks (CNN, RNN, GNN, transformer and contrastive learning) for each of the models in Table 3, number of parameters (No. Par.), and whether interpretabilty (Intr.) is introduced or not. ✓indicates the presence of aforementioned features.

Model	Year	CNN	RNN	GNN	Transf.	Contr. Lear.	Interp.	#No. Par.
SeqSleepNet [25]	2018	–	✓	–	–	–	–	∼0.2 M
U-Time [82]	2019	✓	–	–	–	–	–	∼1.2 M
IITNet [79]	2019	✓	✓	–	–	–	–	–
GraphSleepNet [171]	2020	–	–	✓	–	–	–	–
Xsleepnet [78]	2020	✓	✓	–	–	–	–	∼5.8 M
RobustSleepNet[77]	2021	–	✓	–	–	–	–	∼0.18 M
CCRRSleepNet [132]	2021	✓	✓	–	–	–	–	–
SalientSleepNet [147]	2021	✓	–	–	–	–	–	∼0.9 M
SleepTransformer [166]	2021	–	–	–	✓	–	✓	∼3.7 M
Cross-Modal [167]	2022	✓	–	–	✓	–	✓	∼4.05 M
MtCLSS [175]	2022	✓	–	–	–	✓	–	–
MAtt [146]	2022	✓	–	–	–	–	–	–
PearNet [85]	2022	✓	–	✓	–	–	–	–
CMS2-Net [176]	2022	✓	–	–	–	✓	–	–
SleepContextNet [59]	2022	✓	✓	–	–	–	–	–
mulEEG [56]	2022	✓	–	–	–	✓	–	∼0.6 M
Liu et al. [18]	2022	✓	–		–	–	–	–
MVF-SleepNet [88]	2022	–	–	–	–	–	–	–
TrustSleepNet [30]	2022	–	–	–	–	–	–	–
SoftVotingSleepNet [152]	2022	✓	–	–	–	–	–	∼0.79 M
Kim et al. [128]	2022	✓	–	–	–	–	–	–
CRFs [149]	2022	✓	✓	–	–	–	–	∼0.76 M
IDNN [159]	2022	✓	✓	–	–	–	✓	–
CAttSleepNet [138]	2022	✓	✓	–	–	–	–	–
ISENet [140]	2022	–	–	–	–	–	–	–
Bi-RNN [102]	2022	✓	✓	–	–	–	–	–
S2MAML [186]	2022	✓	–	–	–	–	–	∼0.6 M
Pei et al. [103]	2022	✓	✓	–	–	–	–	–
SleepExpertNet [87]	2022	✓	✓	–	–	–	–	–
CDNN [96]	2022	✓	–	–	–	–	–	–
Van Der Donckt et al. [122]	2022	–	–	–	–	–	✓	–
SleepyCo [174]	2022	✓	–	–	–	✓	–	–
ENGELBERT [170]	2022	✓	–	–	✓	–	–	∼0.49 M
CoSleepNet [51]	2023	✓	✓	–	–	–	–	–
LPT-Based [81]	2023	✓	–	–	–	–	–	∼0.16 M
MaskSleepNet [163]	2023	✓	–	–	–	–	–	–
SHNN [127]	2023	✓	✓	–	–	–	–	–
Siamese AE [187]	2023	✓	✓	–	–	–	–	–
BSTT [184]	2023	–	–	✓	✓	–	–	–
NAMRTNet [185]	2023	✓	–	–	–	–	–	–

7. Conclusions

In this paper, we trace the evolution from early attempts using CNNs and RNNs to recent exploration of methods like transformers and contrastive learning. Proper feature extraction is crucial, with self-supervised learning and contrastive techniques proving efficient. Model performance variations across datasets are attributed to factors such as data size, quality, and participant characteristics. Notably, SleepEDF, MASS, and SHHS datasets yield better results, while the challenging CAP dataset exhibits high participant variability and data imbalance. Model designs greatly affect performance differences, for example, SleepExpertNet [87], while XSleepNet [78] balances generalization using multi-modal input. SleepTransformer [166] offers interpretability and is strong for SHHS dataset, while MaskSleepNet [163] adapts through modality masking for MASS dataset. A parameter–performance plot, Figure 25, demonstrates recent models achieving higher or closer performance to bigger models with fewer parameters. We highlighted evolving

techniques, dataset influence, model design impact, and the trend of smaller models yielding improved results.

We investigated sleep studies from various aspects, such as data manipulation, modeling, transfer learning, and interpretability. Not all of these aspects have been thoroughly investigated. It can be seen in Section 3 that different methods have been utilized for data preprocessing and augmentation. Although some of these techniques, such as adding noise, flipping, stretching, etc., show some promise, to the best of the authors' knowledge, there are no studies to properly study these mechanisms and the changes they cause to sleep profiles in the sense of time and frequency features. In [52], some effort to evaluate the DA methods has been performed, but it needs more proper investigation of feature preservation in both time and frequency for reliable data generation. It should be mentioned that some attempts have been made to properly learn time and frequency features to properly perform data manipulation [50,52,57,58]. The current SOTA models use many parameters (trainable parameters). Since these models are supposed to be used in medical settings, some guarantees will help physicians to be inclined to use the models. As discussed in Section 5, some efforts have been conducted to add explainability to the models to describe their behavior. These efforts are normally restricted to visualizing attention layers, using masking techniques, or feature extraction to be able to understand how the information flows from input to output. Despite the methods mentioned in Section 5, more investigation can be performed to quantitatively measure the interpretability of a model and explore new techniques that are more generalizable and do not rely on specific model architectures.

In summary, we observe that there has been a trend of improvement on model performance (e.g., we observed a 1.7% and 12% improvement on the EDF-20 and EDF-78 datasets over the last five years, respectively). However, these improvements are not necessarily due to models with a higher number of parameters (since we did not see a correlation between number of parameters and performance), but instead due to the inclusion of more sophisticated architectures and data augmentation techniques (e.g., using transformers and contrastive learning).

Author Contributions: Conceptualization, R.S. and E.L.; methodology, R.S.; validation, R.S., J.B., and Y.C.; formal analysis, R.S., J.B., Y.C., and E.L.; investigation, R.S., J.B., Y.C., and E.L.; resources, R.S., J.B., Y.C., and E.L.; data curation, R.S., J.B., and Y.C.; writing—original draft preparation, R.S., J.B., Y.C., E.L., A.B., and M.D.; writing—review and editing, R.S., J.B., Y.C., E.L., A.B., V.P. and M.D.; visualization, R.S., J.B., and Y.C.; supervision, R.S., J.B., Y.C., E.L., A.B., and M.D; project administration, R.S. and E.L.; funding acquisition, E.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Science Foundation (NSF) under award IIS-2037328, and EEC-1160483 (ERC for ASSIST).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Selected Models with Details

In this appendix, we investigate some of the models in more detail.

Appendix A.1. XsleepNet [78]

As discussed in Section 4.2, XsleepNet [78] combines two views that contribute differently to the model performance with different rate of convergence. This can result in overfitting in one view and underfitting in another. To account for this issue, the authors proposed to use adaptive gradient blending [188] to calculate generalization/overfitting

ratio to properly adjust the rates of the different views in the loss function. The calculations for cross-entropy (CE) are

$$\mathcal{L}^{(k)} = -\frac{1}{L}\sum_{l=1}^{L} \mathbf{y}_l \log(\hat{\mathbf{y}}_l^{(k)}),$$

where $k \in \{1, 2, *\}$, which refers to different views and their combination. The loss weights are calculated as

$$w^{(k)} = \frac{1}{Z}\frac{G_k}{O_k^2},$$

where Z is a normalization factor, G_k is the generalization factor that is defined as the gained information about the target distribution in training, and O_k is the overfitting measure that is defined as the gap between the gain on the training set and the target distribution. The weighted loss is defined as

$$\mathcal{L}(n) = \sum_{k\in\{1,2,*\}} w^{(k)}(n)\mathcal{L}^{(k)}(n),$$

where n is the training step.

Appendix A.2. EpochNet [81]

In Section 4.2, the LPT method was introduced. The authors evaluated four LPT methods: one-dimensional local binary pattern (1D-LBP), local neighbor descriptive pattern (LNDP), local gradient pattern (LGP), and local neighbor gradient pattern (LNGP). Both single-channel and multi-channel were extensively used in the analysis. The authors considered single- and multi-channel EEG, single-channel EOG, and combinations of both. Before classification, the model extracts epoch features from many subsequent epochs in parallel and blends them to construct inter-epoch links among extracted features. Since LPT methods transform a given signal into a certain range, the proposed approach does not require normalization or standardization of the signals. Furthermore, the signals are not filtered before the transformation, and the approach relies on the discriminative power of discovered local patterns. The model is shown in Figure A1.

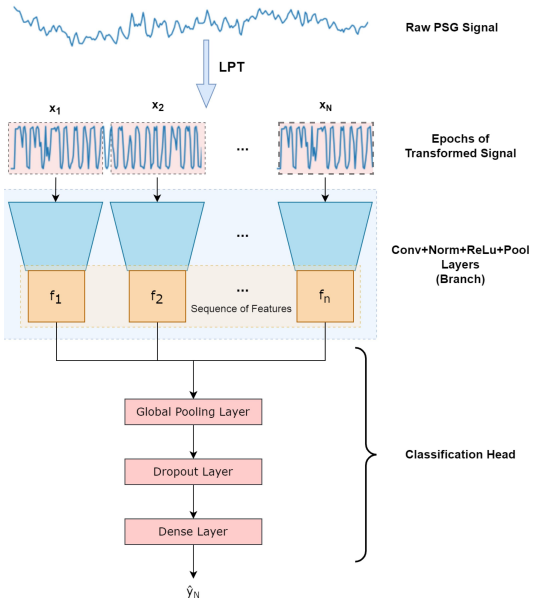


Figure A1. The proposed framework in [81]. Image taken from [81].

Appendix A.3. CrossModal Transformer [167]

As it can be seen for each of the steps in Figure 10, there is a special learnable vector, CLS, that is concatenated to each block output. The CLS vector helps to reduce model

size, to add modalities, and to incorporate interpretability. In the first block, CLS_{EEG} and CLS_{EOG} are vectors for each distinct modalities. CLS_{Cross} vector is added to output of both modalities to account for the cross-modal attention layer and also optimally combine both modalities for decision making. Because of the model structure, the authors proposed an attention-based interpretability for intra-modal, cross-modal, and inter-epoch relationships. We go over this aspect of the model in Section 5.

Appendix A.4. GraphSleepNet [171]

The authors defined a non-negative function $A_{mn} = g(x_m, x_n)$, $n, m \in \{1, 2, 3, \ldots, N\}$, where x_j is a specific input channel from the EEG. This function represents the connection and is implemented by a single-layer neural network with weights vector w. Formally,

$$g(x_m, x_n) = \frac{\exp(\text{ReLU}(w^T|x_m - x_n|))}{\sum_{n=1}^{N} \exp(\text{ReLU}(w^T|x_m - x_n|))}.$$

The weights vector is updated by the following loss function

$$\mathcal{L}_{graph\_learning} = \sum_{m,n=1}^{N} \|x_m - x_n\|_2^2 A_{mn} + \lambda \|A\|_F^2$$

where λ is a hyperparameter controlling the sparsity of the adjacency matrix. The adjacency matrix is fed through a custom spatial-temporal attention layer and a series of graph convolutional layers before being used for classification.

Appendix A.5. mulEEG [56]

The authors used a family of two augmentations, T_1 and T_2, to provide two samples for the same signal, as shown in Figure 19b. The spectrogram for each of these samples is calculated and passed through their respective networks (E_t for time signal, and E_s for spectrograms). For E_t, ResNet-50 [189] with 1D-convolutions is used, and for E_s, the architecture from [190] is used. After the encoders, there are projection layers f, h, and g to map time signal, spectrogram, and fusion to z^t, z^s, and z^f, respectively. The z^t, z^s, and z^f are used to calculate the diverse loss as follows

$$\ell_d(z_k, a, b) = -\log \frac{\exp(\cos(z_k[a], z_k[b])/\tau_d)}{\sum_{i=1}^{4} \mathbf{1}_{[i \neq a]} \exp(\cos(z_k[a], z_k[i])/\tau_d))},$$

$$L_D = \frac{1}{4N} \sum_{k=1}^{N} \ell_d(z_k, 1, 2) + \ell_d(z_k, 2, 1) + \ell_d(z_k, 3, 4) + \ell_d(z_k, 4, 3),$$

$$L_{tot} = \lambda_1 (L_{TT} + L_{FF} + L_{SS}) + \lambda_2 L_D$$

where λ_1 and λ_2 are hyper-parameters to be chosen, a and b are indexes for the elements in z_k, and $z_k = [z_i^t, z_j^t, z_i^s, z_j^s]$ $\forall i = j$. It uses three contrastive losses: L_{TT} accounts for time-series feature space, L_{SS} for spectrogram feature space, and L_{FF} for concatenated features space. L_D further encourages the complementary information to be exploited between time-series and spectrogram views.

Appendix A.6. CoSleep [173]

This model takes time-series and spectrogram as input and passes them through f_1 and f_2 encoders, as shown in Figure 14. Both encoders have residual convolutional networks from [191]. The output of encoders goes through a similarity module to build a linear combination of the similarities of the two views as follows

$$\omega_{sim} = \lambda \cdot \omega_1 + (1 - \lambda) \cdot \omega_2$$

where $\omega_1 \in \mathbb{R}^n$ is a similarity vector between z_t and any sample from the positive set V_1, and similarly, $\omega_2 \in \mathbb{R}^{n*1}$ corresponds to the similarity of the sample to a spectrogram V_2. After this step, the loss, which is called multi-instance InfoNCE (MI-NCE), is calculated as follows

$$\mathcal{L}_{\text{MI-NCE}} = -\sum \log \left[\frac{\sum_{i \in \mathcal{P}_t} \exp(\hat{z}_t \cdot z_i)}{\sum_{i \in \mathcal{P}_t} \exp(\hat{z}_t \cdot z_i) + \sum_{k=1} \exp(\hat{z}_t \cdot z_k)} \right],$$

where $\mathcal{P}_t$ represents the positive set to x_t. $\mathcal{P}_t$ is defined as follows

$$\mathcal{P}_t = \{\psi(x_t), x_k | k \in \text{topK}\{\omega_{sim}\}\},$$

where $\psi(\cdot) \in \{f_1, f_2\}$ represents the up-to-date encoder for each view, and topK finds K nearest neighbors present over all samples in the batch, respective to ω_{sim}. The MI-NCE is inspired by the InfoNCE loss [192], which is defined as follows,

$$\mathcal{L}_{\text{InfoNCE}} = -\sum \log \left[\frac{\exp(\hat{z}_t \cdot z_t)}{\exp(\hat{z}_t \cdot z_t) + \sum_{k=1} \exp(\hat{z}_t \cdot z_k)} \right],$$

where $z_t \cdot \hat{z}_t$ is the cosine similarity between prediction and ground-truth representation. InfoNCE attracts positive samples and repels negative instances. It is worth mentioning that MI-NCE incorporates multiple positives, which help the model to achieve semantic-class-level representation. In the training step, the encoders are trained with InfoNCE first to be conditioned on a good initialization rather than random initialization. After the pre-training, the model is trained based on MI-NCE loss, but to do so, in each iteration, one of the networks is frozen while the other one is being updated.

Appendix B. Data Modalities

Table A1. Different modalities present in selected datasets used in Table 3.

Dataset	EEG	EOG	EMG
SleepEDF	✓	✓	✓
MASS	✓	✓	✓
SHHS	✓	×	×
CAP	✓	✓	✓
ISRUC	✓	✓	✓
DRM-SUB	✓	✓	✓
SVUH-UCD	✓	✓	✓

References

1. Alhola, P.; Polo-Kantola, P. Sleep deprivation: Impact on cognitive performance. *Neuropsychiatr. Dis. Treat.* **2007**, *3*, 553–567.
2. Tramonti Fantozzi, M.P.; Banfi, T.; Di Galante, M.; Ciuti, G.; Faraguna, U. Sleep deprivation-induced changes in baseline brain activity and vigilant attention performance. *Brain Sci.* **2022**, *12*, 1690. [CrossRef] [PubMed]
3. Stowe, R.C.; Afolabi-Brown, O. Pediatric polysomnography—A review of indications, technical aspects, and interpretation. *Paediatr. Respir. Rev.* **2020**, *34*, 9–17. [CrossRef] [PubMed]
4. Berry, R.B.; Brooks, R.; Gamaldo, C.E.; Harding, S.M.; Marcus, C.; Vaughn, B.V. The AASM manual for the scoring of sleep and associated events. In *Rules, Terminology and Technical Specifications*; American Academy of Sleep Medicine: Darien, IL, USA , 2012; Volume 176, p. 2012.
5. Phan, H.; Mikkelsen, K. Automatic sleep staging of EEG signals: Recent development, challenges, and future directions. *Physiol. Meas.* **2022**, *43*, 04TR01. [CrossRef]
6. Craik, A.; He, Y.; Contreras-Vidal, J.L. Deep learning for electroencephalogram (EEG) classification tasks: A review. *J. Neural Eng.* **2019**, *16*, 031001. [CrossRef] [PubMed]
7. Sri, T.R.; Madala, J.; Duddukuru, S.L.; Reddipalli, R.; Polasi, P.K. A Systematic Review on Deep Learning Models for Sleep Stage Classification. In Proceedings of the 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 28–30 April 2022; pp. 1505–1511.
8. Moher, D.; Shamseer, L.; Clarke, M.; Ghersi, D.; Liberati, A.; Petticrew, M.; Shekelle, P.; Stewart, L.A. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst. Rev.* **2015**, *4*, 1–9. [CrossRef] [PubMed]

9. Zhang, L.; Fabbri, D.; Upender, R.; Kent, D. Automated sleep stage scoring of the Sleep Heart Health Study using deep neural networks. *Sleep* **2019**, *42*, zsz159. [CrossRef] [PubMed]
10. Danker-hopfe, H.; Anderer, P.; Zeitlhofer, J.; Boeck, M.; Dorn, H.; Gruber, G.; Heller, E.; Loretz, E.; Moser, D.; Parapatics, S.; et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *J. Sleep Res.* **2009**, *18*, 74–84. [PubMed]
11. Delgado, R.; Tibau, X.A. Why Cohen's Kappa should be avoided as performance measure in classification. *PloS ONE* **2019**, *14*, e0222916. [CrossRef] [PubMed]
12. Kemp, B.; Zwinderman, A.H.; Tuk, B.; Kamphuisen, H.A.; Oberye, J.J. Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG. *IEEE Trans. Biomed. Eng.* **2000**, *47*, 1185–1194. [CrossRef]
13. Yildirim, O.; Baloglu, U.B.; Acharya, U.R. A deep learning model for automated sleep stages classification using PSG signals. *Int. J. Environ. Res. Public Health* **2019**, *16*, 599. [CrossRef] [PubMed]
14. Qu, W.; Wang, Z.; Hong, H.; Chi, Z.; Feng, D.D.; Grunstein, R.; Gordon, C. A residual based attention model for eeg based sleep staging. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 2833–2843. [CrossRef] [PubMed]
15. Supratak, A.; Guo, Y. TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), virtual, 20–24 July 2020; pp. 641–644.
16. Lee, H.; Seong, E.; Chae, D.K. Self-supervised learning with attention-based latent signal augmentation for sleep staging with limited labeled data. In Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI-22), Vienna, Austria, 23–29 July 2022; pp. 3868–3876.
17. Mai, X.; Yu, T. BootstrapNet: An Contrastive Learning Model for Sleep Stage Scoring based on Raw Single-Channel Electroencephalogram. In Proceedings of the 2021 2nd International Conference on Artificial Intelligence and Computer Engineering (ICAICE), Hangzhou, China, 5–7 November 2021; pp. 303–308.
18. Liu, Z.; Luo, S.; Lu, Y.; Zhang, Y.; Jiang, L.; Xiao, H. Extracting multi-scale and salient features by MSE based U-structure and CBAM for sleep staging. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2022**, *31*, 31–38. [CrossRef] [PubMed]
19. Tao, Y.; Yang, Y.; Yang, P.; Nan, F.; Zhang, Y.; Rao, Y.; Du, F. A novel feature relearning method for automatic sleep staging based on single-channel EEG. *Complex Intell. Syst.* **2022**, *9*, 41–50. [CrossRef]
20. Yang, B.; Zhu, X.; Liu, Y.; Liu, H. A single-channel EEG based automatic sleep stage classification method leveraging deep one-dimensional convolutional neural network and hidden Markov model. *Biomed. Signal Process. Control* **2021**, *68*, 102581. [CrossRef]
21. Jain, R.; Ganesan, R.A. Reliable sleep staging of unseen subjects with fusion of multiple EEG features and RUSBoost. *Biomed. Signal Process. Control* **2021**, *70*, 103061. [CrossRef]
22. Zhao, L.; Liu, G.; Tang, X.; Bai, Y.; Li, Y.; Wang, X. Improved Model Accuracy Based on a Simple Frame of Temporal-Correlation Representation Method in Sleep Staging. In Proceedings of the 2022 3rd International Conference on Pattern Recognition and Machine Learning (PRML), Chengdu, China, 22–24 July 2022; pp. 216–222.
23. O'reilly, C.; Gosselin, N.; Carrier, J.; Nielsen, T. Montreal Archive of Sleep Studies: An open-access resource for instrument benchmarking and exploratory research. *J. Sleep Res.* **2014**, *23*, 628–635. [CrossRef]
24. Jia, Z.; Lin, Y.; Wang, J.; Ning, X.; He, Y.; Zhou, R.; Zhou, Y.; Li-wei, H.L. Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2021**, *29*, 1977–1986. [CrossRef]
25. Phan, H.; Andreotti, F.; Cooray, N.; Chén, O.Y.; De Vos, M. SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2019**, *27*, 400–410. [CrossRef]
26. Sleep Heart Health Study (SHHS). Available online: https://biolincc.nhlbi.nih.gov/studies/shhs/ (accessed on 3 January 2018).
27. Sridhar, N.; Shoeb, A.; Stephens, P.; Kharbouch, A.; Shimol, D.B.; Burkart, J.; Ghoreyshi, A.; Myers, L. Deep learning for automated sleep staging using instantaneous heart rate. *NPJ Digit. Med.* **2020**, *3*, 106. [CrossRef]
28. Van Steenkiste, T.; Groenendaal, W.; Deschrijver, D.; Dhaene, T. Automated sleep apnea detection in raw respiratory signals using long short-term memory neural networks. *IEEE J. Biomed. Health Inform.* **2018**, *23*, 2354–2364. [CrossRef] [PubMed]
29. Li, Q.; Li, Q.; Liu, C.; Shashikumar, S.P.; Nemati, S.; Clifford, G.D. Deep learning in the cross-time frequency domain for sleep staging from a single-lead electrocardiogram. *Physiol. Meas.* **2018**, *39*, 124005. [CrossRef] [PubMed]
30. Huang, G.; Ma, F. TrustSleepNet: A Trustable Deep Multimodal Network for Sleep Stage Classification. In Proceedings of the 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Ioannina, Greece, 27–30 September 2022; pp. 1–4.
31. Biswal, S.; Sun, H.; Goparaju, B.; Westover, M.B.; Sun, J.; Bianchi, M.T. Expert-level sleep scoring with deep neural networks. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 1643–1650. [CrossRef]
32. Terzano, M.G.; Parrino, L.; Sherieri, A.; Chervin, R.; Chokroverty, S.; Guilleminault, C.; Hirshkowitz, M.; Mahowald, M.; Moldofsky, H.; Rosa, A.; et al. Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep. *Sleep Med.* **2001**, *2*, 537–554. [CrossRef] [PubMed]
33. Khalighi, S.; Sousa, T.; Santos, J.M.; Nunes, U. ISRUC-Sleep: A comprehensive public dataset for sleep researchers. *Comput. Methods Programs Biomed.* **2016**, *124*, 180–192. [CrossRef] [PubMed]

34. Zhang, G.Q.; Cui, L.; Mueller, R.; Tao, S.; Kim, M.; Rueschman, M.; Mariani, S.; Mobley, D.; Redline, S. The National Sleep Research Resource: Towards a sleep data commons. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 1351–1358. [CrossRef]
35. Chen, X.; Wang, R.; Zee, P.; Lutsey, P.L.; Javaheri, S.; Alcántara, C.; Jackson, C.L.; Williams, M.A.; Redline, S. Racial/ethnic differences in sleep disturbances: The Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep* **2015**, *38*, 877–888. [CrossRef]
36. Martínez, C.G.; Niediek, J.; Mormann, F.; Andrzejak, R.G. Seizure onset zone lateralization using a non-linear analysis of micro vs. macro electroencephalographic recordings during seizure-free stages of the sleep-wake cycle from epilepsy patients. *Front. Neurol.* **2020**, *11*, 553885. [CrossRef]
37. Nagabushanam, P.; Thomas George, S.; Radha, S. EEG signal classification using LSTM and improved neural network algorithms. *Soft Comput.* **2020**, *24*, 9981–10003. [CrossRef]
38. Goldberger, A.L.; Amaral, L.A.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **2000**, *101*, e215–e220. [CrossRef]
39. Loh, H.W.; Ooi, C.P.; Vicnesh, J.; Oh, S.L.; Faust, O.; Gertych, A.; Acharya, U.R. Automated detection of sleep stages using deep learning techniques: A systematic review of the last decade (2010–2020). *Appl. Sci.* **2020**, *10*, 8963. [CrossRef]
40. Mashrur, F.R.; Islam, M.S.; Saha, D.K.; Islam, S.R.; Moni, M.A. SCNN: Scalogram-based convolutional neural network to detect obstructive sleep apnea using single-lead electrocardiogram signals. *Comput. Biol. Med.* **2021**, *134*, 104532. [CrossRef] [PubMed]
41. Ichimaru, Y.; Moody, G. Development of the polysomnographic database on CD-ROM. *Psychiatry Clin. Neurosci.* **1999**, *53*, 175–177. [CrossRef] [PubMed]
42. Penzel, T.; Moody, G.B.; Mark, R.G.; Goldberger, A.L.; Peter, J.H. The apnea-ECG database. In Proceedings of the Computers in Cardiology (CinC), Cambridge, MA, USA, 24–27 September 2000; pp. 255–258.
43. Devuyst, S.; Dutoit, T.; Kerkhofs, M. *The DREAMS Databases and Assessment Algorithm*; Zenodo: Geneva, Switzerland, 2005.
44. Ghassemi, M.M.; Moody, B.E.; Lehman, L.W.H.; Song, C.; Li, Q.; Sun, H.; Mark, R.G.; Westover, M.B.; Clifford, G.D. You snooze, you win: The physionet/computing in cardiology challenge 2018. In Proceedings of the Computers in Cardiology (CinC), Maastricht, The Netherlands, 23–26 September 2018; Volume 45; pp. 1–4.
45. Li, Y.; Xu, Z.; Zhang, Y.; Cao, Z.; Chen, H. Automatic sleep stage classification based on a two-channel electrooculogram and one-channel electromyogram. *Physiol. Meas.* **2022**, *43*, 07NT02. [CrossRef] [PubMed]
46. Chen, Z.; Yang, Z.; Zhu, L.; Chen, W.; Tamura, T.; Ono, N.; Altaf-Ul-Amin, M.; Kanaya, S.; Huang, M. Automated sleep staging via parallel frequency-cut attention. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2023**, *31*, 1974–1985. [CrossRef] [PubMed]
47. Zhong, Q.; Lei, H.; Chen, Q.; Zhou, G. A Multi-scale Residual Convolutional Neural Network for Sleep Staging Based on Single Channel Electroencephalography Signal. 2021, *preprint (Version 1)*. [CrossRef]
48. An, P.; Yuan, Z.; Zhao, J.; Jiang, X.; Wang, Z.; Du, B. Multi-subband and Multi-subepoch Time Series Feature Learning for EEG-based Sleep Stage Classification. In Proceedings of the 2021 IEEE International Joint Conference on Biometrics (IJCB), Shenzhen, China (virtual), 4–7 August 2021; pp. 1–8.
49. Huang, Z.; Ling, B.W.K. Sleeping stage classification based on joint quaternion valued singular spectrum analysis and ensemble empirical mode decomposition. *Biomed. Signal Process. Control* **2022**, *71*, 103086. [CrossRef]
50. Kuo, C.E.; Chen, G.T.; Liao, P.Y. An EEG spectrogram-based automatic sleep stage scoring method via data augmentation, ensemble convolution neural network, and expert knowledge. *Biomed. Signal Process. Control* **2021**, *70*, 102981. [CrossRef]
51. Efe, E.; Ozsen, S. CoSleepNet: Automated sleep staging using a hybrid CNN-LSTM network on imbalanced EEG-EOG datasets. *Biomed. Signal Process. Control* **2023**, *80*, 104299. [CrossRef]
52. Fan, J.; Sun, C.; Chen, C.; Jiang, X.; Liu, X.; Zhao, X.; Meng, L.; Dai, C.; Chen, W. EEG data augmentation: Towards class imbalance problem in sleep staging tasks. *J. Neural Eng.* **2020**, *17*, 056017. [CrossRef]
53. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. Bootstrap your own latent—A new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, virtual, 6–12 December 2020; pp. 21271–21284.
54. Eldele, E.; Ragab, M.; Chen, Z.; Wu, M.; Kwoh, C.K.; Li, X.; Guan, C. Time-series representation learning via temporal and contextual contrasting. *arXiv* **2021**, arXiv:2106.14112.
55. Mohsenvand, M.N.; Izadi, M.R.; Maes, P. Contrastive representation learning for electroencephalogram classification. In Proceedings of the Machine Learning for Health (PMLR), virtual, 11 December 2020; pp. 238–253.
56. Kumar, V.; Reddy, L.; Kumar Sharma, S.; Dadi, K.; Yarra, C.; Bapi, R.S.; Rajendran, S. mulEEG: A multi-view representation learning on EEG signals. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Singapore, 18–22 September 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 398–407.
57. Xu, Q.; Zhou, D.; Wang, J.; Shen, J.; Kettunen, L.; Cong, F. Convolutional Neural Network Based Sleep Stage Classification with Class Imbalance. In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 18–23 July 2022; pp. 1–6.
58. Efe, E.; Özsen, S. A New Approach for Automatic Sleep Staging: Siamese Neural Networks. *Trait. Signal* **2021** . [CrossRef]
59. Zhao, C.; Li, J.; Guo, Y. SleepContextNet: A temporal context network for automatic sleep staging based single-channel EEG. *Comput. Methods Programs Biomed.* **2022**, *220*, 106806. [CrossRef]
60. He, Z.; Du, L.; Wang, P.; Xia, P.; Liu, Z.; Song, Y.; Chen, X.; Fang, Z. Single-channel EEG sleep staging based on data augmentation and cross-subject discrepancy alleviation. *Comput. Biol. Med.* **2022**, *149*, 106044. [CrossRef]

61. Wallace, B.C.; Small, K.; Brodley, C.E.; Trikalinos, T.A. Class imbalance, redux. In Proceedings of the 2011 IEEE 11th International Conference on Data Mining (KDD) Workshops, Vancouver, BC, Canada, 11 December 2011; pp. 754–763.
62. Zhou, D.; Xu, Q.; Wang, J.; Xu, H.; Kettunen, L.; Chang, Z.; Cong, F. Alleviating class imbalance problem in automatic sleep stage classification. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–12. [CrossRef]
63. Kenton, J.D.M.W.C.; Toutanova, L.K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
64. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process* **2020**, *33*, virtual, 6–12 December 2020; pp. 1877–1901.
65. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30. https://dl.acm.org/doi/10.5555/3295222.3295349. [CrossRef]
66. Andreotti, F.; Phan, H.; Cooray, N.; Lo, C.; Hu, M.T.; De Vos, M. Multichannel sleep stage classification and transfer learning using convolutional neural networks. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 17–21 July 2018; pp. 171–174.
67. ElMoaqet, H.; Eid, M.; Ryalat, M.; Penzel, T. A deep transfer learning framework for sleep stage classification with single-channel EEG signals. *Sensors* **2022**, *22*, 8826. [CrossRef]
68. Supratak, A.; Dong, H.; Wu, C.; Guo, Y. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2017**, *25*, 1998–2008. [CrossRef]
69. Zhang, C.; Liao, Y.; Han, S.; Zhang, M.; Wang, Z.; Xie, X. Multichannel Multidomain-Based Knowledge Distillation Algorithm for Sleep Staging With Single-Channel EEG. *IEEE Trans. Circuits Syst. II Express Br.* **2022**, *69*, 4608–4612. [CrossRef]
70. Supratak, A.; Haddawy, P. Quantifying the impact of data characteristics on the transferability of sleep stage scoring models. *Artif. Intell. Med.* **2023**, *139*, 102540. [CrossRef]
71. Eldele, E.; Ragab, M.; Chen, Z.; Wu, M.; Kwoh, C.K.; Li, X.; Guan, C. ADAST: Attentive cross-domain EEG-based sleep staging framework with iterative self-training. *IEEE Trans. Emerg.* **2022**, *7*, 210–221. [CrossRef]
72. Nasiri, S.; Clifford, G.D. Attentive adversarial network for large-scale sleep staging. In Proceedings of the Machine Learning for Healthcare (MLHC) (PMLR), Durham, NC, USA, 7–8 August 2020; pp. 457–478.
73. Zhao, R.; Xia, Y.; Zhang, Y. Unsupervised sleep staging system based on domain adaptation. *Biomed. Signal Process. Control* **2021**, *69*, 102937. [CrossRef]
74. Heremans, E.R.; Phan, H.; Borzée, P.; Buyse, B.; Testelmans, D.; De Vos, M. From unsupervised to semi-supervised adversarial domain adaptation in electroencephalography-based sleep staging. *J. Neural Eng.* **2022**, *19*, 036044. [CrossRef]
75. Fan, J.; Zhu, H.; Jiang, X.; Meng, L.; Chen, C.; Fu, C.; Yu, H.; Dai, C.; Chen, W. Unsupervised domain adaptation by statistics alignment for deep sleep staging networks. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2022**, *30*, 205–216. [CrossRef]
76. Phan, H.; Chén, O.Y.; Koch, P.; Mertins, A.; De Vos, M. Deep transfer learning for single-channel automatic sleep staging with channel mismatch. In Proceedings of the 2019 27th European Signal Processing Conference (EUSIPCO), A CorUña, Spain, 2–6 September 2019; pp. 1–5.
77. Guillot, A.; Thorey, V. RobustSleepNet: Transfer learning for automated sleep staging at scale. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2021**, *29*, 1441–1451. [CrossRef]
78. Phan, H.; Chén, O.Y.; Tran, M.C.; Koch, P.; Mertins, A.; De Vos, M. XSleepNet: Multi-view sequential model for automatic sleep staging. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 5903–5915. [CrossRef] [PubMed]
79. Seo, H.; Back, S.; Lee, S.; Park, D.; Kim, T.; Lee, K. Intra-and inter-epoch temporal context network (IITNet) using sub-epoch features for automatic sleep scoring on raw single-channel EEG. *Biomed. Signal Process. Control* **2020**, *61*, 102037. [CrossRef]
80. Zhu, T.; Luo, W.; Yu, F. Convolution-and attention-based neural network for automated sleep stage classification. *Int. J. Environ. Res. Public Health* **2020**, *17*, 4152. [CrossRef]
81. Zan, H.; Yildiz, A. Local Pattern Transformation-Based convolutional neural network for sleep stage scoring. *Biomed. Signal Process. Control* **2023**, *80*, 104275. [CrossRef]
82. Perslev, M.; Jensen, M.H.; Darkner, S.; Jennum, P.J.; Igel, C. U-Time: A fully convolutional network for time series segmentation applied to sleep staging. In Proceedings of the NeurIPS, Vancouver, BC, Canada, 8–14 December 2019; pp. 4415–4426.
83. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015, Proceedings, Part III 18*; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
84. Mousavi, S.; Afghah, F.; Acharya, U.R. SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach. *PLoS ONE* **2019**, *14*, e0216456. [CrossRef] [PubMed]
85. Lu, J.; Tian, Y.; Wang, S.; Sheng, M.; Zheng, X. PearNet: A Pearson Correlation-based Graph Attention Network for Sleep Stage Recognition. In Proceedings of the 9th of the IEEE International Conference on Data Science and Advanced Analytics (DSAA), virtual, 13–16 October 2022; pp. 1–8.
86. Stine, R.A. Graphical interpretation of variance inflation factors. *Am. Stat.* **1995**, *49*, 53–56.

87. Lee, C.H.; Kim, H.J.; Kim, Y.T.; Kim, H.; Kim, J.B.; Kim, D.J. SleepExpertNet: High-performance and class-balanced deep learning approach inspired from the expert neurologists for sleep stage classification. *Ambient. Intell. Humaniz. Comput.* **2023**, *14*, 8067–8083. [CrossRef]
88. Li, Y.; Chen, J.; Ma, W.; Zhao, G.; Fan, X. MVF-sleepnet: Multi-view fusion network for sleep stage classification. *IEEE J. Biomed. Health Inform.* **2022**, *ahead of print*. [CrossRef]
89. Zhang, J.; Yao, R.; Ge, W.; Gao, J. Orthogonal convolutional neural networks for automatic sleep stage classification based on single-channel EEG. *Comput. Methods Programs Biomed.* **2020**, *183*, 105089. [CrossRef] [PubMed]
90. Nie, H.; Tu, S.; Xu, L. Recsleepnet: An automatic sleep staging model based on feature reconstruction. In Proceedings of the 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), virtual, 9–12 December 2021; pp. 1458–1461.
91. Jia, Z.; Cai, X.; Zheng, G.; Wang, J.; Lin, Y. SleepPrintNet: A multivariate multimodal neural network based on physiological time-series for automatic sleep staging. *IEEE Trans. Artif. Intell.* **2020**, *1*, 248–257. [CrossRef]
92. Jadhav, P.; Rajguru, G.; Datta, D.; Mukhopadhyay, S. Automatic sleep stage classification using time–frequency images of CWT and transfer learning using convolution neural network. *Biocybern. Biomed. Eng.* **2020**, *40*, 494–504. [CrossRef]
93. Zhu, F.; Liang, Q. OCRNN: An orthogonal constrained recurrent neural network for sleep analysis based on EEG data. *Ad. Hoc. Netw.* **2020**, *104*, 102178. [CrossRef]
94. Li, C.; Qi, Y.; Ding, X.; Zhao, J.; Sang, T.; Lee, M. A deep learning method approach for sleep stage classification with eeg spectrogram. *Int. J. Environ. Res. Pub. Health* **2022**, *19*, 6322. [CrossRef]
95. Urtnasan, E.; Park, J.U.; Joo, E.Y.; Lee, K.J. Deep convolutional recurrent model for automatic scoring sleep stages based on single-lead ECG signal. *Diagnostics* **2022**, *12*, 1235. [CrossRef]
96. Kwon, K.; Kwon, S.; Yeo, W.H. Automatic and accurate sleep stage classification via a convolutional deep neural network and nanomembrane electrodes. *Biosensors* **2022**, *12*, 155. [CrossRef] [PubMed]
97. Li, H.; Guan, Y. DeepSleep convolutional neural network allows accurate and fast detection of sleep arousal. *Commun. Biol.* **2021**, *4*, 18. [CrossRef] [PubMed]
98. Olesen, A.N.; Jørgen Jennum, P.; Mignot, E.; Sorensen, H.B.D. Automatic sleep stage classification with deep residual networks in a mixed-cohort setting. *Sleep* **2021**, *44*, zsaa161. [CrossRef]
99. Yan, R.; Li, F.; Zhou, D.; Ristaniemi, T.; Cong, F. A deep learning model for automatic sleep scoring using multimodality time series. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 23–27 August 2021; pp. 1090–1094.
100. Van Der Donckt, J.; Van Der Donckt, J.; Deprost, E.; Vandenbussche, N.; Rademaker, M.; Vandewiele, G.; Van Hoecke, S. Do not sleep on traditional machine learning: Simple and interpretable techniques are competitive to deep learning for sleep scoring. *Biomed. Signal Process. Control* **2023**, *81*, 104429. [CrossRef]
101. Yan, R.; Li, F.; Zhou, D.D.; Ristaniemi, T.; Cong, F. Automatic sleep scoring: A deep learning architecture for multi-modality time series. *J. Neurosci. Methods* **2021**, *348*, 108971. [CrossRef]
102. Fu, Z.; Huang, C.; Zhang, L.; Wang, S.; Zhang, Y. Deep Learning Model of Sleep EEG Signal by Using Bidirectional Recurrent Neural Network Encoding and Decoding. *Electronics* **2022**, *11*, 2644. [CrossRef]
103. Pei, W.; Li, Y.; Siuly, S.; Wen, P. A hybrid deep learning scheme for multi-channel sleep stage classification. *Comput. Mater. Contin.* **2022**, *71*, 889–905.
104. Michielli, N.; Acharya, U.R.; Molinari, F. Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals. *Comput. Biol. Med.* **2019**, *106*, 71–81. [CrossRef]
105. Fiorillo, L.; Favaro, P.; Faraci, F.D. Deepsleepnet-lite: A simplified automatic sleep stage scoring model with uncertainty estimates. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2021**, *29*, 2076–2085. [CrossRef]
106. Zhang, X.; Xu, M.; Li, Y.; Su, M.; Xu, Z.; Wang, C.; Kang, D.; Li, H.; Mu, X.; Ding, X.; et al. Automated multi-model deep neural network for sleep stage scoring with unfiltered clinical data. *Sleep Breath.* **2020**, *24*, 581–590. [CrossRef]
107. Fernandez-Blanco, E.; Rivero, D.; Pazos, A. Convolutional neural networks for sleep stage scoring on a two-channel EEG signal. *Soft Comput.* **2020**, *24*, 4067–4079. [CrossRef]
108. Lee, T.; Hwang, J.; Lee, H. Trier: Template-guided neural networks for robust and interpretable sleep stage identification from eeg recordings. *arXiv* **2020**, arXiv:2009.05407.
109. Barnes, L.D.; Lee, K.; Kempa-Liehr, A.W.; Hallum, L.E. Detection of sleep apnea from single-channel electroencephalogram (EEG) using an explainable convolutional neural network (CNN). *PLoS ONE* **2022**, *17*, e0272167. [CrossRef] [PubMed]
110. Wang, I.N.; Lee, C.H.; Kim, H.J.; Kim, H.; Kim, D.J. An ensemble deep learning approach for sleep stage classification via single-channel EEG and EOG. In Proceedings of the 2020 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Republic of Korea, 21–23 October 2020; pp. 394–398.
111. Li, Y.; Gu, Z.; Lin, Z.; Yu, Z.; Li, Y. An automatic sleep staging model combining feature learning and sequence learning. In Proceedings of the 12th International Conference on Advanced Computational Intelligence (ICACI), Dali, China, 14–16 August 2020; pp. 419–425.
112. Zhang, H.; Wang, X.; Li, H.; Mehendale, S.; Guan, Y. Auto-annotating sleep stages based on polysomnographic data. *Patterns* **2022**, *3*, 100371. [CrossRef]
113. Sors, A.; Bonnet, S.; Mirek, S.; Vercueil, L.; Payen, J.F. A convolutional neural network for sleep stage scoring from raw single-channel EEG. *Biomed. Signal Process. Control* **2018**, *42*, 107–114. [CrossRef]

114. Abou Jaoude, M.; Sun, H.; Pellerin, K.R.; Pavlova, M.; Sarkis, R.A.; Cash, S.S.; Westover, M.B.; Lam, A.D. Expert-level automated sleep staging of long-term scalp electroencephalography recordings using deep learning. *Sleep* **2020**, *43*, zsaa112. [CrossRef]
115. Fan, J.; Sun, C.; Long, M.; Chen, C.; Chen, W. Eognet: A novel deep learning model for sleep stage classification based on single-channel eog signal. *Front. Neurosci.* **2021**, *15*, 573194. [CrossRef]
116. Anandakumar, M.; Pradeepkumar, J.; Kappel, S.L.; Edussooriya, C.U.; De Silva, A.C. A Knowledge Distillation Framework for Enhancing Ear-EEG Based Sleep Staging with Scalp-EEG Data. *arXiv* **2022**, arXiv:2211.02638.
117. Mikkelsen, K.; De Vos, M. Personalizing deep learning models for automatic sleep staging. *arXiv* **2018**, arXiv:1801.02645.
118. Brandmayr, G.; Hartmann, M.; F̀urbass, F.; Dorffner, G. Self-attention long-term dependency modelling in electroencephalography sleep stage prediction. In Proceedings of the International Conference on Neural Information Processing, Sanur, Bali, Indonesia, 8–12 December 2021; Springer: Berlin/Heidelberg, Germany, pp. 379–390.
119. Satapathy, S.K.; Loganathan, D. Automated classification of multi-class sleep stages classification using polysomnography signals: A nine-layer 1D-convolution neural network approach. *Multimed. Tools Appl.* **2023**, *82*, 8049–8091. [CrossRef]
120. Phan, H.; Andreotti, F.; Cooray, N.; Chén, O.Y.; De Vos, M. Joint classification and prediction CNN framework for automatic sleep stage classification. *IEEE. Trans. Biomed. Eng.* **2018**, *66*, 1285–1296. [CrossRef] [PubMed]
121. Li, F.; Yan, R.; Mahini, R.; Wei, L.; Wang, Z.; Mathiak, K.; Liu, R.; Cong, F. End-to-end sleep staging using convolutional neural network in raw single-channel EEG. *Biomed. Signal Process. Control* **2021**, *63*, 102203. [CrossRef]
122. Van Der Donckt, J.; Van Der Donckt, J.; Deprost, E.; Rademaker, M.; Vandewiele, G.; Van Hoecke, S. Do not sleep on linear models: Simple and interpretable techniques outperform deep learning for sleep scoring. *arXiv* **2022**, arXiv:2207.07753.
123. Parekh, N.; Dave, B.; Shah, R.; Srivastava, K. Automatic sleep stage scoring on raw single-channel EEG: A comparative analysis of CNN architectures. In Proceedings of the 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT), Erode, India, 15–17 September 2021; pp. 1–8.
124. Zhao, R.; Xia, Y.; Wang, Q. Dual-modal and multi-scale deep neural networks for sleep staging using EEG and ECG signals. *Biomed. Signal Process. Control* **2021**, *66*, 102455. [CrossRef]
125. Khalili, E.; Asl, B.M. Automatic sleep stage classification using temporal convolutional neural network and new data augmentation technique from raw single-channel EEG. *Comput. Methods Programs Biomed.* **2021**, *204*, 106063. [CrossRef] [PubMed]
126. Wang, H.; Lu, C.; Zhang, Q.; Hu, Z.; Yuan, X.; Zhang, P.; Liu, W. A novel sleep staging network based on multi-scale dual attention. *Biomed. Signal Process. Control* **2022**, *74*, 103486. [CrossRef]
127. Zhang, Y.; Cao, W.; Feng, L.; Wang, M.; Geng, T.; Zhou, J.; Gao, D. Shnn: A single-channel eeg sleep staging model based on semi-supervised learning. *Expert Syst. Appl.* **2023**, *213*, 119288. [CrossRef]
128. Kim, H.; Lee, S.M.; Choi, S. Automatic sleep stages classification using multi-level fusion. *Biomed. Eng. Lett.* **2022**, *12*, 413–420. [CrossRef]
129. Jadhav, P.; Mukhopadhyay, S. Automated sleep stage scoring using time-frequency spectra convolution neural network. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–9. [CrossRef]
130. Zhou, D.; Wang, J.; Hu, G.; Zhang, J.; Li, F.; Yan, R.; Kettunen, L.; Chang, Z.; Xu, Q.; Cong, F. SingleChannelNet: A model for automatic sleep stage classification with raw single-channel EEG. *Biomed. Signal Process. Control* **2022**, *75*, 103592. [CrossRef]
131. Fang, Y.; Xia, Y.; Chen, P.; Zhang, J.; Zhang, Y. A dual-stream deep neural network integrated with adaptive boosting for sleep staging. *Biomed. Signal Process. Control* **2023**, *79*, 104150. [CrossRef]
132. Neng, W.; Lu, J.; Xu, L. CCRRSleepNet: A hybrid relational inductive biases network for automatic sleep stage classification on raw single-channel eeg. *Brain Sci.* **2021**, *11*, 456. [CrossRef]
133. Jiang, X.; Zhao, J.; Bo, D.; Panfeng, A.; Guo, H.; Yuan, Z. MRNet: A Multi-scale Residual Network for EEG-based Sleep Staging. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18-22 July 2021; pp. 1–8.
134. Goshtasbi, N.; Boostani, R.; Sanei, S. SleepFCN: A fully convolutional deep learning framework for sleep stage classification using single-channel electroencephalograms. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2022**, *30*, 2088–2096. [CrossRef]
135. Sun, C.; Chen, C.; Li, W.; Fan, J.; Chen, W. A hierarchical neural network for sleep stage classification based on comprehensive feature learning and multi-flow sequence learning. *IEEE J. Biomed. Health Inform.* **2019**, *24*, 1351–1366. [CrossRef]
136. Phyo, J.; Ko, W.; Jeon, E.; Suk, H.I. TransSleep: Transitioning-Aware attention-based deep neural network for sleep staging. *IEEE Trans. Cybern.* **2022**, *53*, 4500–4510. [CrossRef] [PubMed]
137. Zhang, C.; Liu, S.; Han, F.; Nie, Z.; Lo, B.; Zhang, Y. Hybrid manifold-deep convolutional neural network for sleep staging. *Methods* **2022**, *202*, 164–172. [CrossRef] [PubMed]
138. Li, T.; Zhang, B.; Lv, H.; Hu, S.; Xu, Z.; Tuergong, Y. CAttSleepNet: Automatic end-to-end sleep staging using attention-based deep neural networks on single-channel EEG. *Int. J. Environ. Res. Public Health* **2022**, *19*, 5199. [CrossRef]
139. Banville, H.; Wood, S.U.; Aimone, C.; Engemann, D.A.; Gramfort, A. Robust learning from corrupted EEG with dynamic spatial filtering. *NeuroImage* **2022**, *251*, 118994. [CrossRef] [PubMed]
140. Huang, J.; Ren, L.; Zhou, X.; Yan, K. An improved neural network based on SENet for sleep stage classification. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 4948–4956. [CrossRef]
141. Phan, H.; Andreotti, F.; Cooray, N.; Chén, O.Y.; De Vos, M. Automatic sleep stage classification using single-channel eeg: Learning sequential features with attention-based recurrent neural networks. In Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 17–21 July 2018; pp. 1452–1455.

142. Kuo, C.E.; Liao, P.Y.; Lin, Y.S. A self-attention-based ensemble convolution neural network approach for sleep stage classification with merged spectrogram. In Proceedings of the 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, 14–17 December 2021; pp. 1262–1268.
143. Yuan, Y.; Jia, K.; Ma, F.; Xun, G.; Wang, Y.; Su, L.; Zhang, A. A hybrid self-attention deep learning framework for multivariate sleep stage classification. *BMC Bioinform.* **2019**, *20*, 1–10. [CrossRef]
144. Yuan, Y.; Xun, G.; Ma, F.; Suo, Q.; Xue, H.; Jia, K.; Zhang, A. A novel channel-aware attention framework for multi-channel eeg seizure detection via multi-view deep learning. In Proceedings of the Biomedical and Health Informatics (BHI), Las Vegas, NV, USA, 4–7 March 2018; pp. 206–209.
145. Ma, F.; Chitta, R.; Zhou, J.; You, Q.; Sun, T.; Gao, J. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In Proceedings of the Knowledge Discovery and Data Mining (KDD), Halifax, NS, Canada, 13–17 2017; pp. 1903–1911.
146. Pan, Y.T.; Chou, J.L.; Wei, C.S. MAtt: A Manifold Attention Network for EEG Decoding. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 31116–31129.
147. Jia, Z.; Lin, Y.; Wang, J.; Wang, X.; Xie, P.; Zhang, Y. SalientSleepNet: Multimodal salient wave detection network for sleep staging. *arXiv* **2021**, arXiv:2105.13864. .
148. Eldele, E.; Chen, Z.; Liu, C.; Wu, M.; Kwoh, C.K.; Li, X.; Guan, C. An attention-based deep learning approach for sleep stage classification with single-channel EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2021**, *29*, 809–818. [CrossRef] [PubMed]
149. Yang, B.; Wu, W.; Liu, Y.; Liu, H. A novel sleep stage contextual refinement algorithm leveraging conditional random fields. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–13. [CrossRef]
150. Wu, Y.; Fang, X.; Li, J.; Zhang, L.; Chen, Z.; Wang, Y. A deep learning approach with conditional random field for automatic sleep stage scoring. In Proceedings of the Conference on Electronic Information Technology and Computer Engineering (EITCE), Xiamen, China, 22–24 October 2021; pp. 901–906.
151. Hong, J.K.; Lee, T.; Delos Reyes, R.D.; Hong, J.; Tran, H.H.; Lee, D.; Jung, J.; Yoon, I.Y. Confidence-Based Framework Using Deep Learning for Automated Sleep Stage Scoring. *Nat. Sci. Sleep* **2021**, *13*, 2239–2250. [CrossRef]
152. Neshov, N.; Tonchev, K.; Velchev, Y.; Manolova, A.; Poulkov, V. SoftVotingSleepNet: Majority Vote of Deep Learning Models for Sleep Stage Classification from Raw Single EEG Channel. In Proceedings of the International Black Sea Conference on Communications and Networking (BlackSeaCom), Sofia, Bulgaria, 6–9 June 2022; pp. 298–302.
153. Zhang, L.; Chen, D.; Chen, P.; Li, W.; Li, X. Dual-CNN based multi-modal sleep scoring with temporal correlation driven fine-tuning. *Neurocomputing* **2021**, *420*, 317–328. [CrossRef]
154. Khare, S.K.; Bajaj, V.; Taran, S.; Sinha, G. Multiclass sleep stage classification using artificial intelligence based time-frequency distribution and CNN. In *Artificial Intelligence-Based Brain-Computer Interface*; Elsevier: Amsterdam, The Netherlands, 2022; pp. 1–21.
155. Erdenebayar, U.; Kim, Y.J.; Park, J.U.; Joo, E.Y.; Lee, K.J. Deep learning approaches for automatic detection of sleep apnea events from an electrocardiogram. *Comput. Methods Programs Biomed.* **2019**, *180*, 105001. [CrossRef]
156. Zhou, D.; Xu, Q.; Wang, J.; Zhang, J.; Hu, G.; Kettunen, L.; Chang, Z.; Cong, F. LightSleepNet: A lightweight deep model for rapid sleep stage classification with spectrograms. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Virtual, 1–5 November 2021; pp. 43–46.
157. Sanghavi, S.; Vaid, P.; Rathod, P.; Srivastava, K. SpectroTemporalNet: Automated Sleep Stage Scoring with Stacked Generalization. In Proceedings of the 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 4–6 August 2021; pp. 1256–1263.
158. Begawan, I.A.; Djamal, E.C.; Djajasasmita, D.; Kasyidi, F.; Nugraha, F. Sleep Stage Identification Based on EEG Signals Using Parallel Convolutional Neural Network and Recurrent Neural Network. In Proceedings of the 2022 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia, 1–3 October 2022; pp. 39–44.
159. Baek, J.; Lee, C.; Yu, H.; Baek, S.; Lee, S.; Lee, S.; Park, C. Automatic Sleep Scoring Using Intrinsic Mode Based on Interpretable Deep Neural Networks. *IEEE Access* **2022**, *10*, 36895–36906. [CrossRef]
160. Kim, Y.; Ryu, J.; Kim, K.K.; Took, C.C.; Mandic, D.P.; Park, C. Motor imagery classification using mu and beta rhythms of EEG with strong uncorrelating transform based complex common spatial patterns. *Comput. Intell. Neurosci.* **2016**, *2016*, 1489692. [CrossRef]
161. Jia, Z.; Cai, X.; Jiao, Z. Multi-modal physiological signals based squeeze-and-excitation network with domain adversarial learning for sleep staging. *IEEE Sens. J.* **2022**, *22*, 3464–3471. [CrossRef]
162. Wang, S.; Liu, W.; Wu, J.; Cao, L.; Meng, Q.; Kennedy, P.J. Training deep neural networks on imbalanced data sets. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 4368–4374.
163. Zhu, H.; Zhou, W.; Fu, C.; Wu, Y.; Shen, N.; Shu, F.; Yu, H.; Chen, C.; Chen, W. MaskSleepNet: A Cross-modality Adaptation Neural Network for Heterogeneous Signals Processing in Sleep Staging. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 2353–2364. [CrossRef] [PubMed]
164. Jeong, S.; Ko, W.; Mulyadi, A.W.; Suk, H.I. Efficient continuous manifold learning for time series modeling. *arXiv* **2021**, arXiv:2112.03379.
165. Chien, H.Y.S.; Goh, H.; Sandino, C.M.; Cheng, J.Y. MAEEG: Masked Auto-encoder for EEG Representation Learning. *arXiv* **2022**, arXiv:2211.02625.
166. Phan, H.; Mikkelsen, K.; Chén, O.Y.; Koch, P.; Mertins, A.; De Vos, M. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE. Trans. Biomed. Eng.* **2022**, *69*, 2456–2467. [CrossRef] [PubMed]

167. Pradeepkumar, J.; Anandakumar, M.; Kugathasan, V.; Suntharalingham, D.; Kappel, S.L.; De Silva, A.C.; Edussooriya, C.U. Towards interpretable sleep stage classification using cross-modal transformers. *arXiv* **2022**, arXiv:2208.06991.
168. Yubo, Z.; Yingying, L.; Bing, Z.; Lin, Z.; Lei, L. MMASleepNet: A multimodal attention network based on electrophysiological signals for automatic sleep staging. *Front. Neurosci.* **2022**, *16*, 973761. [CrossRef]
169. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
170. Brandmayr, G.; Hartmann, M.; F̈urbass, F.; Matz, G.; Samwald, M.; Kluge, T.; Dorffner, G. Relational local electroencephalography representations for sleep scoring. *Neural Netw.* **2022**, *154*, 310–322. [CrossRef]
171. Jia, Z.; Lin, Y.; Wang, J.; Zhou, R.; Ning, X.; He, Y.; Zhao, Y. GraphSleepNet: Adaptive Spatial-Temporal Graph Convolutional Networks for Sleep Stage Classification. In Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI), Yokohama, Japan, 7–15 January 2021; Volume 2021; pp. 1324–1330.
172. Zhang, H.; Wang, J.; Xiao, Q.; Deng, J.; Lin, Y. Sleeppriorcl: Contrastive representation learning with prior knowledge-based positive mining and adaptive temperature for sleep staging. *arXiv* **2021**, arXiv:2110.09966.
173. Ye, J.; Xiao, Q.; Wang, J.; Zhang, H.; Deng, J.; Lin, Y. Cosleep: A multi-view representation learning framework for self-supervised learning of sleep stage classification. *IEEE Signal Process. Lett.* **2021**, *29*, 189–193. [CrossRef]
174. Lee, S.; Yu, Y.; Back, S.; Seo, H.; Lee, K. SleePyCo: Automatic Sleep Scoring with Feature Pyramid and Contrastive Learning. *arXiv* **2022**, arXiv:2209.09452.
175. Li, Y.; Luo, S.; Zhang, H.; Zhang, Y.; Zhang, Y.; Lo, B. MtCLSS: Multi-Task Contrastive Learning for Semi-Supervised Pediatric Sleep Staging. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 2647–2655. [CrossRef] [PubMed]
176. Zhang, C.; Yu, W.; Li, Y.; Sun, H.; Zhang, Y.; De Vos, M. CMS2-net: Semi-supervised sleep staging for diverse obstructive sleep apnea severity. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 3447–3457. [CrossRef]
177. Huang, G.; Ma, F. Concad: Contrastive learning-based cross attention for sleep apnea detection. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2021, Bilbao, Spain, 13–17 September 2021, Proceedings, Part V 21*; Springer International Publishing: Cham, Switzerland, 2021; pp. 68–84.
178. Kumar, C.B.; Mondal, A.K.; Bhatia, M.; Panigrahi, B.K.; Gandhi, T.K. SCL-SSC: Supervised Contrastive Learning for Sleep Stage Classification. *arXiv* **2022**, arXiv:2109.07839
179. Al-Hussaini, I.; Xiao, C.; Westover, M.B.; Sun, J. SLEEPER: Interpretable Sleep staging via Prototypes from Expert Rules. In Proceedings of the 4th Machine Learning for Healthcare Conference (PMLR), Ann Arbor, MI, USA, 8–10 August 2019; pp. 721–739.
180. Ellis, C.A.; Zhang, R.; Carbajal, D.A.; Miller, R.L.; Calhoun, V.D.; Wang, M.D. Explainable Sleep Stage Classification with Multimodal Electrophysiology Time-series. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Virtual, 1–5 November 2021; pp. 2363–2366.
181. Troncoso-García, A.; Martínez-Ballesteros, M.; Martínez-Álvarez, F.; Troncoso, A. Explainable machine learning for sleep apnea prediction. *Procedia Comput. Sci.* **2022**, *207*, 2930–2939. [CrossRef]
182. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
183. Al-Hussaini, I.; Mitchell, C.S. Serf: Interpretable sleep staging using embeddings, rules, and features. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM), Atlanta, GA, USA, 17–21 October 2022; pp. 3791–3795.
184. Liu, Y.; Jia, Z. Bstt: A bayesian spatial-temporal transformer for sleep staging. In Proceedings of the 11st International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.
185. Xu, X.; Chen, C.; Meng, K.; Lu, L.; Cheng, X.; Fan, H. NAMRTNet: Automatic Classification of Sleep Stages Based on Improved ResNet-TCN Network and Attention Mechanism. *Appl. Sci.* **2023**, *13*, 6788. [CrossRef]
186. Lemkhenter, A.; Favaro, P. Towards sleep scoring generalization through self-supervised meta-learning. In Proceedings of the 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, UK, 11–15 July 2022; pp. 2961–2966.
187. You, Y.; Guo, X.; Yang, Z.; Shan, W. A Siamese Network-Based Method for Improving the Performance of Sleep Staging with Single-Channel EEG. *Biomedicines* **2023**, *11*, 327. [CrossRef]
188. Wang, W.; Tran, D.; Feiszli, M. What makes training multi-modal classification networks hard? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), virtual, 14–19 June 2020; pp. 12695–12705.
189. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
190. Yang, C.; Xiao, D.; Westover, M.B.; Sun, J. Self-supervised eeg representation learning for automatic sleep staging. *arXiv* **2021**, arXiv:2110.15278.

191. Huang, N.E.; Shen, Z.; Long, S.R.; Wu, M.C.; Shih, H.H.; Zheng, Q.; Yen, N.C.; Tung, C.C.; Liu, H.H. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. Math. Phys. Eng. Sci.* **1998**, *454*, 903–995. [CrossRef]
192. Sarkar, P.; Etemad, A. Self-supervised learning for ecg-based emotion recognition. In Proceedings of the International Conference on Acoustics, Speech, & Signal Processing (ICASSP), virtual, 4–8 May 2020; pp. 3217–3221.