Automatic Sleep Staging of EEG Signals: Recent Development, Challenges, and Future Directions

Huy Phan[1,2], **Kaare Mikkelsen**[3]

[1]School of Electronic Engineering & Computer Science, Queen Mary University of London, UK

[2]The Alan Turing Institute, UK

[3]Department of Electrical and Computer Engineering, Aarhus University, Denmark

E-mail: h.phan@qmul.ac.uk, mikkelsen.kaare@ece.au.dk

Abstract.

Modern deep learning holds a great potential to transform clinical practice on human sleep. Teaching a machine to carry out routine tasks would be a tremendous reduction in workload for clinicians. Sleep staging, a fundamental step in sleep practice, is a suitable task for this and will be the focus in this article. Recently, automatic sleep staging systems have been trained to mimic manual scoring, leading to similar performance to human sleep experts, at least on scoring of healthy subjects. Despite tremendous progress, we have not seen automatic sleep scoring adopted widely in clinical environments. This review aims to give a shared view of the authors on the most recent state-of-the-art development in automatic sleep staging, the challenges that still need to be addressed, and the future directions for automatic sleep scoring to achieve clinical value.

1. Introduction

Sleep makes up almost one third of our lives. Good sleep is crucial in maintaining one's mental and physical health [1, 2] while sleep disorders are linked with a host of different ailments [3]. Screening, assessment, and diagnosis for sleep disorders require each 30-second epoch of an overnight polysomnogram (PSG) to be assigned a sleep stage. This procedure is known as sleep staging or scoring. The sequence of sleep stages is critical for measuring parameters of the sleep macrostructure, such as the sleep cycles, the time spend in each stage, sleep latency, wake after sleep onset (WASO), etc. Sleep stages and cycles, which manifest the underlying neuro-physiological processes, are also a rich source for mining diagnostic markers for a wide range of sleep disorders [4–7], from a common one like obstructive sleep apnea (OSA) [8, 9] to a rare one like narcolepsy [6, 7].

Sleep staging is still largely carried out by clinicians in sleep clinics guided by a well established manual, published by the American Academy of Sleep Medicine [10]. This labor-intensive and time-consuming manual scoring is unsuited for handling large-scale data and cannot be scaled to serve the needs of millions suffering from sleep disorders [11–13]. At the same time, there is an increasing need for longitudinal monitoring in home environments. Accurate and cost effective monitoring of sleep not only has great medical value but also allows individuals to self-assess and self-manage their sleep. Thus, it is imperative for sleep staging to be automated. The fact that it follows a predefined set of rules makes sleep staging a perfect task for automation with machine learning. Furthermore, a machine can perform the task thousands of times faster than a human expert, while saving a clinician thousands of hours a year and making sleep assessment and diagnosis more widely available.

Indeed, given the standardization of data through the use of PSG for recordings, a very large volume of methods development has already taken place. Particularly, the existence of increasingly large public data sets online (for example, PhysioNet [14] and the National Sleep Research Resource (NSRR) [15]) has enabled exploitation of deep learning [16, 17] to teach a machine to perform sleep staging using a large amount of training data in the last couple of years. These efforts have led to more advanced and sensible methods [18–24] which have surpassed the agreement level of experts' scoring and achieved a performance acceptable for clinical use. Notwithstanding this tremendous progress, machine sleep scoring still needs to overcome several technical and clinical barriers to be widely adopted and deliver full clinical value. We see great potential for further development. First, from algorithmic perspectives, we consider sleep staging to be an interesting modelling problem where novel methods can be developed to tackle the foreseeable obstacles and pave the way for clinical usage. Second, this is related to the emergence of new recording platforms [25–28] to simplify the sleep setup for home-based monitoring purposes.

In this review article, we begin by discussing the clinical context of automatic sleep scoring, after which we give an overview as well as technical insights of the state-of-the-art methods for automatic scoring of EEG data. Readers should note that a few existing reviews, such as Fiorilli *et al.* [29] and Faust *et al.* 2019 [30], have summed up the topic prior to 2019. A review on broader applications of deep learning on EEG analysis also exists [31]. To avoid re-inventing the wheel, this article focuses on the latest method development in automatic sleep staging. In addition, we limit the scope of this article to fine-grained (*i.e.*, five stages) sleep staging using PSG and modalities directly reading brain activities, such as mobile EEGs, and will not cover research work using other modalities, such as electrocardiogram (ECG)/photoplethysmography (PPG) [32], actigraphy [33], audio [34], video [35], and radar [36, 37]. We then discuss the current challenges, and suggest directions to move forwards. As we shall discuss in the next section, much good work has already been done on this problem. However, as it will become clear in the rest of this review, we believe the field has only solved the first, "entry", problem, and a plethora of new and exciting tasks lie ahead of us.

2. Clinical context

Manual sleep scoring is a somewhat reliable, highly versatile method, which readily yields interpretations and which is standardized across the world. This has made it a good solution, but also a local optimum which is hard to escape. By this we mean that it is not the best possible solution, due to a number of drawbacks:

(i) It is very time consuming (and therefore expensive) to manually score an entire night's recording. Even more so if sleep events are also to be annotated.

(ii) Despite the existence of a sleep scoring standard, there is still variation between individual scorers.

(iii) The sleep scoring manual is based on the polysomnography (PSG) recording setup, which is generally considered to be unwieldy and invasive on sleep. This, combined with the cost of each recording, means that clinicians will usually have to 'make do' with a single (at most two) nights of data, which may not be as representative of the patient's usual nights as one would hope.

It should come as no surprise that the properties of manual sleep scoring have shaped the way sleep monitoring is used - few recordings per subject, qualitative (non-data driven) analysis. This can make it hard, within the clinical reality, to immediately see the benefits of a new method (automatic sleep scoring with other sensor setups). Figuratively speaking, if you have learned to solve all problems using nails, it is hard to see how a screwdriver can compete with your hammer.

Automatic sleep scoring can reduce the costs of existing procedures (PSG recordings either in lab or at home), but also open the door to new ways of using sleep clinically, which today would be infeasible. We can imagine population wide screening for early stages of debilitating diseases (e.g., the REM Sleep Behavior Disorder (RBD) is known to be tightly associated with Parkinson's disease [38]), or routine follow-up procedures quantifying patient sleep after they leave the hospital. These procedures could have very real clinical benefits, but they all require changes to how sleep recordings are used and managed, not to abolish existing procedures, but to supplement them.

An algorithm-first approach to clinical sleep can also solve other problems. First, given the costs of a PSG recording, clinicians may often have to 'make do' with whatever recordings they get, even if the quality is questionable. However, if the standard quantum becomes a week's worth of data, automatic discarding of low quality nights would be trivial. Second, definitions surrounding sleep have been developed and evaluated based on how well they can be used in manual sleep scoring. Computers have far fewer restrictions in this manner, and we can imagine more flexible taxonomies, such as hypnodensity plots [6] or even disease-specific sleep states.

3. The state-of-the-art sleep scoring

Modern deep learning [16, 17] crept into sleep research much slower than other fields such as computer vision, natural language processing, and speech recognition. The use of deep neural networks for automatic sleep staging only started around 5 years ago even though their resurgence has been almost a decade. Nevertheless, in this short period of time, deep neural networks have produced impactful and meaningful results that were never seen with more conventional machine learning methods for a long time.

Transitioning from conventional machine learning, the first attempts to use deep learning for automatic sleep staging mainly employed simple networks in traditional fashion where a short input contexts of one to a few sleep epochs around a target epoch is used to predict the sleep stage of the target epoch. Expectedly, an influx of different variants of typical standalone network architectures, such as DNN [39, 40], CNN [18, 41–51], and RNN [44, 52] (e.g., long short-term memory (LSTM) [53] or gated recurrent unit (GRU) [54]), was seen with limited success. Although these networks are able to learn useful features to represent an input, they are unable to capture long-range dependencies between sleep epochs due to the short input context. The ability of modelling long-range dependency plays an important role in improving sleep staging performance, due to the inherently slow-transition nature of the physiological processes behind sleep stages [55, 56]. In order to compensate for the lack of long-term modelling ability, once such a network has been trained and each epoch is encoded into an epoch-wise feature vector, an additional RNN (e.g. LSTM) [6,18,39,57] was separately trained in a second stage to take into account a long sequence of epoch-wise feature vectors prior to a target epoch to classify it. These hybrid networks with two-stage training initiated by Supratak *et al.* [18] boosted the performance significantly and stood out from the other exiting models at the time.

In fact, the positive effect introduced by long-term modelling in the above-mentioned two-stage training scheme is not a surprise. It resembles how manual scoring is done by sleep experts who normally need to attend to a much larger context around a target epoch in order to determine its label [10]. From modelling perspective, this has commonly been accomplished by Hidden Markov Models (HMM) (see [58] for example) before the evolution of deep learning. However, the early works [6, 18, 39, 57] came with some limitations. First, the independent two-stage training of two subnetworks is sub-optimal since it does not account for the interaction between the epoch-wise feature-learning network and the sequential-modelling counterpart, let alone its inconvenience. Second, even the sequential-modelling network (*i.e.*, the bidirectional RNN) is structured to receive a sequence of epochs as input, it classifies only one target epoch at a time which is usually the last epoch in the input sequence. That is, it is tasked to encode the left-side context of the target epoch in order to make a prediction. Olesen *et al.* in [21] showed that this left-side context often results in lower accuracy than when more balanced one is used.

Nevertheless, these initial results underscore the essence of long-term modelling in

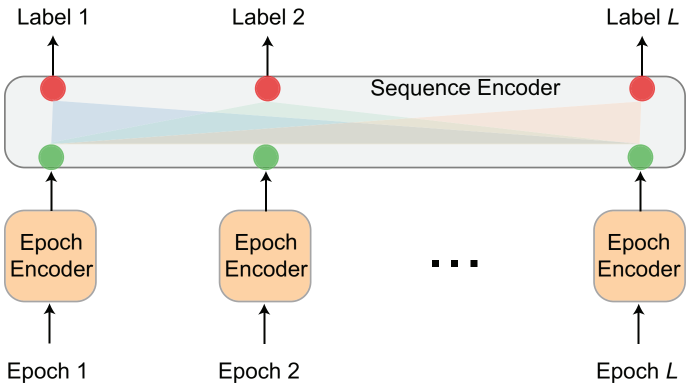

Figure 1: A schematic diagram of sequence-to-sequence sleep staging. Effectively, different epochs in the input sequence is influenced by different contexts, illustrated by the shaded regions in the sequence encoder block, depending on their absolute position in the sequence. The epoch encoder plays the role of the epoch-wise feature extractor that transforms an epoch into a feature vector representation, illustrated by a green circle. The sequence encoder enriches the presentation, illustrated by a red circle, by incorporating interaction of each epoch with other epochs in its context.

automatic sleep staging. Inspired by these results, since late 2018, the community has been witnessing an influx of advanced network architectures with built-in long-term modelling capacity. These networks can be generalized neatly in a common framework, namely *sequence-to-sequence* sleep staging framework [59]. Formally, let us denote an input sequence of L epochs as $(\mathbf{S}_1, \ldots, \mathbf{S}_L)$ where $\mathbf{S}_\ell$ is the ℓ-th epoch, $1 \leq \ell \leq L$. In general, the epochs can be in any form, such as raw signals or time-frequency images and they can be single- or multi-channel. A network adhering to the framework typically consists of two main components: the epoch encoder $\mathcal{F}_E$ and the sequence encoder $\mathcal{F}_S$ as illustrated in Figure 1. The epoch encoder $\mathcal{F}_E : \mathbf{S} \mapsto \mathbf{x}$ acts as an epoch-wise feature extractor which transforms an input epoch $\mathbf{S}$ in the input sequence into a feature vector $\mathbf{x}$ for representation. As a result, the input sequence is transformed into a sequence of feature vectors $(\mathbf{x}_1, \ldots, \mathbf{x}_L)$ (represented by the green circles in the figure). Of note, $\mathcal{F}_S$ can be a hard-coded hand-crafted feature extractor, however, in the deep learning context, it is oftentimes a neural network (e.g., a CNN or an RNN) that learns the feature presentation $\mathbf{x}$ automatically from low-level input signals. In turn, at the sequence level, the sequence encoder $\mathcal{F}_S : (\mathbf{x}_1, \ldots, \mathbf{x}_L) \mapsto (\mathbf{z}_1, \ldots, \mathbf{z}_L)$ transforms the sequence $(\mathbf{x}_1, \ldots, \mathbf{x}_L)$ into another sequence $(\mathbf{z}_1, \ldots, \mathbf{z}_L)$ (represented by the red circles in the figure). In intuition, $\mathbf{z}_\ell$ is a richer representation for the ℓ-th epoch than $\mathbf{x}_\ell$ as it not only encompasses information of the epoch but also encodes its interaction with other epochs in the sequence. More specifically, $\mathbf{z}_\ell$ is derived from $\mathbf{x}_\ell$, taking into account the left context $(\mathbf{x}_1, \ldots, \mathbf{x}_{\ell-1})$ and the right context $(\mathbf{x}_{\ell+1}, \ldots, \mathbf{x}_L)$. Eventually, the vectors $\mathbf{z}_1, \ldots, \mathbf{z}_L$ are used for classification purpose to obtain the sequence of predicted sleep stages, one for each epoch in the input sequence.

The framework features two advantages, making them overcome the limitations of the earlier proposals in [6, 18, 39, 57]. First, both $\mathcal{F}_E$ and $\mathcal{F}_S$ are optimized jointly in an *end-to-end* training fashion, allowing the interaction of the two network components. Second, they are tasked to solve a sequence-to-sequence classification problem, *i.e.*, sequence-to-sequence sleep staging. In other words, a network classifies all the epochs in an input sequence at once rather than only targeting the last epoch. Due to this sequence-to-sequence scheme, different epochs in the input sequence is, in essence, influenced by different contexts depending on their absolute position in the sequence. This is illustrated by the shaded regions in Figure 1. Leveraging this property, sampling and advancing the sequence by one epoch at a time will result in L decisions for a particular epoch. These decisions are associated with diverging contexts; thus, forming an ensemble from them has been shown to lead to performance improvement [19, 23].

In Table 1, we give an overview of automatic sleep staging systems that possess capacity of long-term context modelling. The systems are presented in chronological order. On the one hand, most of the systems expectedly exploited CNN, the cornerstone of deep learning algorithms, for the epoch encoder. The spectrum of the CNN architectures varies from a very basic one [18, 60] to specialized ones, such as ResNet [21], U-Net [61, 62], and U^2-Net [63]. Epoch-wise features can also be learned by capturing sequential information within 30-second signals using RNNs solely (*e.g.*, LSTM [53] and GRU [54]) [19,22,64] or hybrid networks (*e.g.*, CRNN [60,65]). Emerging network architectures like graph convolutional network (GCN) [66] and Transformer [24] have also been shown to be useful for epoch encoding. On the other hand, RNNs have primarily been employed for the sequence encoder due to its well-established capability in sequential modelling. However, inter-epoch sequence modelling can also be accomplished by non-recursive architectures, such as dilated CNN [63], self-attention [67], and Transformer [24]. It should be noted that not all of the networks in the table are strictly sequence-to-sequence (*e.g.*, DeepSleepNet [18], Stephansen *et al.* [6], and GraphSleepNet [66]) or end-to-end (*e.g.*, DeepSleepNet [18], Stephansen *et al.* [6] and Sun *et al.* [57]). However, in principle, they can be framed into the sequence-to-sequence framework and trained end-to-end as what was done with the end-to-end sequence-to-sequence variant of DeepSleepNet in [19].

We also collate the performance of the systems on common public sleep databases in Table 1. These results are obtained either from the original works or in other works where the systems are evaluated. We can see a Cohen's kappa of ≥ 0.81, *i.e.*, "almost perfect" agreement level according to the interpretation of Cohen's kappa [68], achieved on databases with a majority of healthy subjects, for example, EDF-20, MASS, SHHS, DOD-H, DOD-O, and CHAT. However, the performance is still substandard on databases associated with sleep pathologies, such as ISRUC and CAP. It is worth stressing that the performance presented in the table should not be used out-of-context to justify a network's efficacy or compare one to another. The rational is that the potential discrepancies in evaluation setup (*e.g.*, data subsets, the number of channels, etc.) and modelling (*e.g.*, scratch training vs. domain adaptation, supervised vs. unsupervised learning, etc.), renders such a comparison meaningless.

Table 1: Automatic sleep staging systems that has capacity of long-term context modelling published since late 2018 and sorted in chronological order. The reported performance are also presented in term of Cohen's kappa [68]. Alternatively, macro F1-score (indicated with the subscript ^{f}) and overall accuracy (indicated with the subscript ^{a}) are presented where Cohen's kappa is not available. Note that not all of networks here are strictly sequence-to-sequence and/or end-to-end.

Network	Year	Input	Epoch Encoder	Sequence Encoder	EDF-20 [69]	EDF-78 [69]	MASS [70]	Physio-2018 [71]	SHHS [72]	DOD-H [64]	DOD-O [64]	ISRUC [73]	CAP [74]	SVUH-UCD [14]	MESA [75]	MrOS [76,77]	CHAT [78,79]	Other
DeepSleepNet [18,19]	2017*	Raw	CNN	RNN	0.760	0.702	0.800	–	–	0.843	0.804	–	–	–	–	–	0.848	–
SeqSleepNet [19]	2018	Time-freq.	RNN	RNN	0.809	0.776	0.815	0.733	0.838	0.804	0.772	–	–	–	–	–	0.854	–
Stephansen *et al.* [6]	2018	Corr. encoding	CNN	RNN	–	–	–	–	–	–	–	–	–	–	–	–	–	0.868^a
Biswal *et al.* [20]	2018	Time-freq.	CNN	RNN	–	–	–	–	–	–	–	–	–	–	–	–	–	0.805
SleepEEGNet [80]	2019	Raw	CNN	RNN	0.790	0.730	–	–	–	–	–	–	–	–	–	–	–	–
SimpleSleepNet [64]	2019	Time-freq.	RNN	RNN	–	–	–	–	–	84.6	82.3	–	–	–	–	–	–	–
Chen *et al.* [81]†	2019	Raw	CNN	RNN, CRF	0.820	–	–	–	–	–	–	–	–	–	–	–	–	0.670
IITNet [60]	2019	Raw	CRNN	RNN	0.780	0.790	–	–	0.810	–	–	–	–	–	–	–	–	–
U-Time [61]/U-Sleep [62]	2019	Raw	U-Net	CNN	0.790^f	0.760^f	0.800^f	0.770^f	0.800^f	0.820^f	0.790^f	0.770^f	0.680^f	0.730^f	0.790^f	0.770^f	0.850^f	0.850^f
TinySleepNet [82]	2020	Raw	CNN	RNN	0.800	0.770	0.782	–	–	–	–	–	–	–	–	–	–	–
GraphSleepNet [66]	2020	Raw	GCN	Attention	–	–	0.834	–	–	–	–	–	–	–	–	–	–	–
Olesen *et al.* [21]	2020	Raw	ResNet	RNN	–	–	–	–	0.871^a	–	–	0.740^a	–	–	–	0.864^a	–	0.864^a
Jaoude *et al.* [83]	2020	Raw	CNN	RNN	–	–	–	–	–	–	–	–	–	–	–	–	–	0.740
Sun *et al.* [57]	2020	Raw, hand-crafted	CNN	RNN	–	–	0.795	–	–	–	–	–	–	–	–	–	–	–
Korkalainen *et al.* [84]	2020	Raw	CNN	RNN	–	0.780	–	–	–	–	–	–	–	–	–	–	–	0.790
Qu *et al.* [85]	2020	Raw	CNN, ResNet	Self-attention	0.780	–	0.800	–	–	–	–	–	–	–	–	–	–	–
HNSleepNet [86]	2020	Raw	CNN	RNN, Attention	0.780	–	0.810	–	–	–	–	–	–	–	–	–	–	–
Li *et al.* [87]	2020	Raw	CNN	RNN, Attention	0.790	–	–	–	–	–	–	–	–	–	–	–	–	–
FCNN+RNN [23]	2020	Raw	Fully CNN	RNN	0.775	0.759	0.806	0.738	80.9	–	–	–	–	–	–	–	0.847	–
XSleepNet [23]	2020	Raw, time-freq.	CNN, RNN	RNN	0.813	0.778	0.823	0.746	0.847	–	–	–	–	–	–	–	0.857	–
RobustSleepNet [22]	2021	Time-freq.	RNN	RNN	0.817^f	0.779^f	0.825^f	–	0.800^f	0.851^f	0.827^f	–	0.738^f	–	0.795^f	0.756^f	–	–
CCRRSleepNet [65]	2021	Raw	CRNN	RNN	0.780	–	–	–	–	–	–	–	–	–	–	–	–	–
Eldele *et al.* [67]	2021	Raw	CNN	Self-attention	0.790	0.740	–	–	0.780	–	–	–	–	–	–	–	–	–
RecSleepNet [88]	2021	Raw	CNN	RNN	0.813^f	0.779^f	–	–	–	–	–	0.779^f	–	0.743^f	–	–	–	–
Coon *et al.* [89]	2021	Raw	CNN	RNN	–	–	–	–	–	–	–	–	–	–	–	–	–	0.769^a
SalientSleepNet [63]	2021	Raw	U^2-Net	Dilated CNN	0.830^f	0.795^f	–	–	–	–	–	–	–	–	–	–	–	–
SleepTransformer [24]	2021	Time-freq.	Transformer	Transformer	–	0.789	–	–	0.828	–	–	–	–	–	–	–	0.842	–

*DeepSleepNet was introduced by Supratak *et al.* [18] in 2017 and the end-to-end version was presented as a baseline in the SeqSleepNet work [19] by Phan *et al.* in 2018.

†Chen *et al.* [81] conducted 4-stage classification rather 5-stage classification in other works in the table.

4. Challenges and future directions

In our view, automatic sleep staging with PSG on healthy people has basically been solved, not only for adults [20, 23] but also for children [90]. The comparative study by Phan *et al.* [90] showed that different network architectures under the same sequence-to-sequence framework result in a similar "almost perfect" consensus level (according to the interpretation of Cohen's kappa [68]) and little discrepancy was seen among their staging outcomes. This suggests that there is probably little room for accuracy improvement within the same sequence-to-sequence framework. Furthermore, the improvement, if any, is not necessarily meaningful.

While automatic sleep scoring of PSG recordings has come a very long way as described above, there are still challenges to be overcome. These, of course, should be viewed as opportunities for innovation. In Figure 2, we give an overview of the challenges around two critical applications, sleep scoring with PSG in clinical spaces and sleep monitoring with wearable EEG in daily living environments, and directions for future works to address the challenges. Before discussing the individual challenges, we feel it is valuable to highlight the overarching contexts of the two applications:

Clinical PSG scoring: It is not sufficient to have high quality PSG scoring of healthy people. In a clinical setting, the tools applied should be equally capable when confronted with non-textbook sleep phenotypes, where the sleep EEG may either be masked by disease-related artifacts, or where the sleep EEG itself may be drastically changed by the patient's condition that a correct sleep scoring either requires specialized routines or may even be impossible. An automatic sleep scoring algorithm must be able to handle this situation transparently and reliably. Thus, to obtain more widespread adoption clinically, automatic sleep scoring should be as robust as manual scoring, and deliver outputs which are easy to fit into clinical workflows. In Figure 2, this relates particularly to 'data heterogeneity', 'model interpretability', 'learning with noisy labels' and 'tailored algorithms'.

Wearable EEG scoring: Medical grade mobile sleep monitoring has great potentials for revolutionizing healthcare both in screening, diagnosing and follow-up. A high quality monitoring platform would allow easy recording of weeks of sleep from each individual, without incurring higher healthcare costs or discomfort the patient. Such data would be much more representative of the patient's actual sleeping patterns, and alleviate the present issues with phenomena which are only periodic, or which may be impacted by patients sleeping in unfamiliar environments. To reap the full benefits from such a monitoring device, we need analysis tools which can be tailored to the individual, which can detect subtle changes in sleep patterns, and which can describe a patient's sleep in different, quantitative terms than the present single-night hypnograms.

The discussion below is structured following the overview outlined in Figure 2.

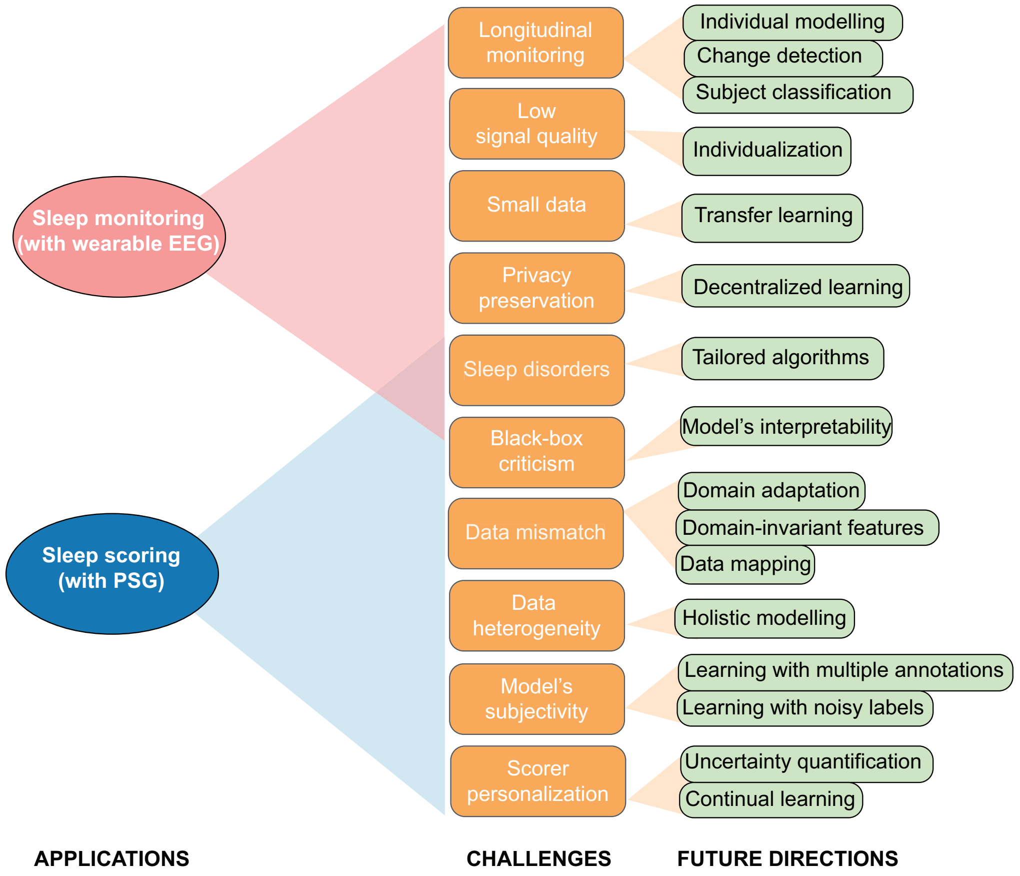

Figure 2: Applications, challenges, and directions in automatic sleep staging.

4.1. Longitudinal monitoring

A particularly interesting development in sleep monitoring, closely related to automatic scoring, is the feasibility of long term monitoring. With conventional PSG setups, scored by hand, it is generally too expensive, not to say inconvenient for the wearer, to perform nightly recordings for weeks on end. However, when scoring is free, and when the hardware can be self-applied [25–27] and is relatively unobtrusive, longitudinal monitoring becomes much more feasible. Interested readers are recommended to refer to [91] for a complete review on devices for wearable sleep staging. These types of devices, and the data sets recorded with them, open new, very interesting avenues of research:

4.1.1. Individual modelling Given a large number of nights from an individual, it seems both possible and advantageous to create personal models using this data. Indeed, we may be inspired by more general approaches for semi-supervised machine learning [92–94]. We could imagine that a successful approach would not only lead to an individual model, but a model which could keep up with the possible "concept drift" that is caused by the long-term changes in a person's sleep patterns. At the same time, such an adaptive approach should be resistant to "catastrophic forgetting" [95], in which the sequential learning of a neural network causes it to forget how to solve previous problems, in this case how to deal with types of sleep or events which only happen very infrequently for the given subject. We suspect that simply requiring a long "memory" is sufficient to solve this problem. Given the performance of models trained as leave-one-subject-out, the size of standard data sets, the fact that inter-subject variation is estimated to be a significant driver of sleep data variation [96–99] and studies evaluating the performance of individualized models [27, 100], we expect that on the order of a 100 nights would be sufficiently long.

4.1.2. Change detection A natural extension to this discussion is detection of sudden (between nights), serious changes to an individual's sleep. Given the large variation in sleep in a given subject, this is not readily feasible based on a few nights. However, based on perhaps weeks of data before and after a significant event (*e.g.*, surgery, change in medication, disease onset, etc.), it seems possible that we could reliably detect that a patient had started to sleep differently. Conceivably, such change detection will be aided by the development of a continuously updating model (since this should entail detecting the need for significant updates).

4.1.3. Subject classification It has been shown that despite the significant night-to-night variation in an individual's sleep, it is possible to define "trait-like characteristics", special to the individual [97–99, 101]. Ideally, this could mean that given a sufficient amount of sleep recordings from an individual, those could be transformed into a reliable biomarker, which could be used to determine not only changes in sleep patterns, but also whether a given patient's sleep was indicative of certain specific diseases, such as RBD.

4.2. Low signal quality - device limitations for mobile sleep monitoring

Mobile sleep monitoring devices have great potential for helping both healthy and sick users get increased knowledge about their own sleep. However, as has been shown in multiple studies [26, 27], even the best studies do not achieve the same inter-scorer reliability measures as PSG-based approaches (state-of-the-art values for Cohen's kappa seem to be about 0.75 for mobile devices [26, 102], while PSG data leads to values above 0.8). A lower signal-to-noise ratio for these types of data (relative to PSG recordings) is likely the main cause for the degraded performance in mobile solutions. Multiple studies [28, 102] have achieved much better performance from the same number of PSG nights as mobile EEG nights (using concurrent recordings). This means that while small data sets can definitely be an issue as discussed in Section 4.3, it seems that making many recordings is likely not sufficient. We can model this phenomenon by imagining the mobile data set as a projection of the high-dimensional PSG data into a lower-dimensional space, with resulting information loss. Given the differences in neuroanatomy, it is reasonable to consider this projection to be subject dependent, and depending on the device design, we could also expect there to be a variation between recording days (because the device may be mounted differently each time).

Several studies have shown that including subject specific information increases sleep scoring performance [27, 100, 103], which is also in line with general EEG studies showing significant differences between individuals [104,105]. Mikkelsen *et al.* [27] found that the differences between algorithms trained on data just from the same subject, and data from both the same subject and many other subjects were minimal. This indicates that the benefit from personalizing sleep scoring algorithms is not that irrelevant data is excluded, but rather that maximally relevant data is included. If this trend were to scale to much larger cohorts, one could imagine that for sufficiently large cohorts, the personalized and general algorithms would perform similarly. However, achieving such a broad training set may be unrealistic in practice.

Individualization of algorithms have been done in multiple ways. Some studies have used random forest ensemble models [27,103,106], which can be trained using very little data. This makes it possible to create full sleep scoring models using only a single or few nights of data from an individual. Other studies, using deep neural networks, have instead resorted to variations of fine tuning population models to individuals. Phan *et al.* [100] explored a technique where the model, during fine tuning, was penalized for making large changes to the output in the source domain, effectively limiting the risk of overfitting to the fine tuning data set.

Note that while some groups focus on developing the entire device, others have strictly worked on developing sleep scoring algorithms for generic "single channel EEG" data sets, without relating it to a specific device [107, 108]. In this discussion we have lumped the two approaches together. We note also that similar observations, regarding mobile device monitoring, are made in the review by Chriskos *et al.* [109].

4.3. Modelling with a small amount of data

Training a deep neural network generally requires a large amount of data. In fact, deep-learning based sleep staging models only reach expert-level performance when the training cohort is large, *i.e.*, hundreds or thousands of subjects [19, 20, 23, 62]. The networks trained with a small cohort continue exhibiting substandard performance. Unfortunately, in practice, many sleep studies only have access to a small cohort, in the order of a few dozens of subjects, for example when studying a particular sleep disorder [5, 48].

This scenario is particularly common in studies exploring the feasibility of a new monitoring device, for example mobile EEG devices [27,28,110]. While the PSG benefits from being an established standard with an enormous user base, new alternatives, by definition, do not. Add to this that many such devices will undergo multiple generations which may not be compatible. Finally and crucially, new training sets will usually require special recordings of both a device and PSG signals, to obtain the necessary ground truth manual PSG scoring which constitutes the training labels. All together, this means that algorithms for new sleep monitoring devices usually have to be trained with quite small data sets.

The most popular solution to this problem is to use transfer learning, usually in the sense that a neural network is trained on a large sleep data set, often consisting of PSG recordings. This model is then fine-tuned using the new data set [22, 59, 108, 110, 111]. Although model fine tuning often results in better performance than scratch model training, the gains are modest in some cases. The problem is that by fine tuning, we essentially further train a pretrained model with a small amount of data that easily causes overfitting without a proper regularization, especially when the source domain and the target domain are significantly different. This can be remedied by fine tuning just a part of the pretrained model, however, identifying the most relevant layers for fine tuning still remains an unsolved question. EEG data augmentation [112] is another direction to explore.

4.4. Privacy preservation: a note for sleep monitoring

Brain waves are a rich source of information from which deciphering numerous sensitive piece of information has been shown possible, such as identity [104], age [113], gender [114], emotion [115], preference [116], personality [117], etc. This poses a challenge to protect this information from user perspective and to comply with legal restrictions, such as General Data Protection Regulation (GDPR) [118] in European Union and Consumer Privacy Bill of Rights (CPBR) [119] in the US.

Traditionally, when sleep staging models are trained, adapted, and deployed centrally, EEG signals are expected to be sent via some communication means. Data privacy and security concerns should be paid heed in this case, however, it is beyond the scope of this article. From the algorithmic perspective, federated learning [120] emerges as a promising solution to address the fundamental problems of privacy and locality of data in the conventional centralized setting. Instead of bringing data to the (centralized) code, federated learning brings the code to (decentralized) data and exploits the distributed resources to train a model collaboratively. Thus, it gets rid of the need for the data to be transferred to a single server. Several open-source federated learning systems are available, e.g. FATE [121], PaddleFL [122], TensorflowFL [123], and Pysyft [124], that would facilitate future research in this direction. Developing compact deep neural networks [125] for sleep staging also becomes relevant. This particularly fits to longitudinal monitoring as model personalization and development can be done on devices, like wearables, smartphones, or Internet-of-Things (IoT) edge devices and the person's data can stay local. Since these devices operate under run-time energy and memory storage constraints, they can only accommodate compact models due to their reduced energy consumption, memory requirement, and inference latency. First, compact versions of existing state-of-the-art models can be derived via quantization [126, 127] to reduce bit-depth of the weights and pruning [128, 129] to remove redundant weights. Second, a network architecture can be hand-designed to remove redundancy, for example depth-wise convolution [130] and shift-based module [131], and thus improve efficiency. Although this approach often results in good model compactness, there is no guiding principle in hand-engineering a network architecture and most of the existing works are more or less based on trial-and-error. An alternative approach is to automatically search for network architectures, *i.e.*, neural architecture search (NAS) [132, 133], which has seen some success, for example in the image domain [134]

4.5. Disorders affecting sleep structure

If a subject suffers from severe neurological disorders, this can have a drastic impact on their sleep. It may both change the structure, timing and outward characteristics of the individual's sleep [135], and it can also change the physiological features of the various sleep stages [136]. This can eventually lead to different types of issues, which require different types of solutions. First, if the manual scoring of the recordings becomes harder, training and validation of a sleep-staging model becomes equally hard [6, 84]. We are not aware of any methodological studies on how to deal with this issue in sleep scoring, but it is a general issue for most types of medical data. Certainly, some of the methods employed in, for instance, medical imaging analysis [137] could be re-purposed for sequence-to-sequence sleep-staging models. Second, if the underlying issue is that distinct sleep stages are becoming less defined (which could be the case in very advanced brain damage), it is possible that more flexible approaches such as continuous sleep depth estimation [138, 139] are a better fit. Third, even if the above issues do not appear, the changes in transition probabilities or even feature spaces have to be absorbed by a classification algorithm to achieve good sleep-staging performance. Fortunately, some studies have shown that a model's performance can be improved significantly if it is individualized, for instance when the subjects are suffering from epilepsy [106] or RBD [48].

A particular implementation which is interesting in this context is the "ASEEGA" algorithm [140] which extracts frequency-based features and scales them according to the individual night. This results in an algorithm which achieves an average Cohen's kappa of 0.8 for healthy adults and full PSG setups (thus, not quite state-of-the-art), but which, on the other hand, appears to be quite resistant to perturbations caused by sleep disorders [141] (reaching Cohen's kappa values between 0.75 and 0.80, depending on disorders). Another promising approach, similar to the "ASEEGA" algorithm, is to feed an entire night into the algorithm at once. In theory, this could enable the algorithm to detect subject- or disease-specific perturbations, and adjust for them. We have seen the approach applied by Li *et al.* [142] for arousal detection, to great success.

4.6. Black-box criticism and interpretability

4.6.1. Black-box criticism and adoption Trust, a psychological mechanism to deal with uncertainty, is a crucial factor influencing interactions and relationships between human and AI, particularly between clinicians and AI in the healthcare domain. This is the chief mechanism that shapes the use and adoption of AI in healthcare settings where life is involved [143]. With its capabilities, deep learning has demonstrated its benefits to healthcare in many aspects: superior performance, capability of handling large and complex data, data-driven learning ability, etc. Examples are the application of deep learning to image-based diagnosis [144], clinical outcome prediction [145], automatic electrocardiogram (ECG) analysis [146], automatic sleep analysis [6], mental health screening [147], intelligent assistive technologies for dementia care [148], to mention a few. However, the complex nature of these algorithms and their inherent "black-box"-ness have been deterring medical professionals' trust [145, 149]. The way that a deep neural network processes input data through interconnected layers to arrive at its staging decisions poses difficulties in deciphering how it learns to produce the outputs. Expectedly, automatic sleep staging systems are not an exception as black-box skepticism remains one of the main questions around their clinical value and adoption.

4.6.2. Interpretability Interpretability is critical for a trustworthy sleep-staging system due to the fact that sleep stages are often ambiguous and even different human experts tend to disagree to a certain extent [22, 150]. While addressing this interpretability problem is mandatory to unleash the clinical value of deep-learning-based sleep staging algorithms, it requires novel technical approaches to understanding the behaviour of these AI systems (*i.e.*, explainable AI [151])

A few attempts have been made to introduce interpretability to a deep-learning-based automatic sleep staging system. Most of them explain the models using feature visualization methods, such as sensitivity maps [152,153] by Vilamala *et al.* [42], Guided Gradient-weighted Class Activation Maps (Guided Grad-CAM) [154] by Andreotti *et al.* [51], saliency map [155] by Jia *et al.* [63], and self-attention score [156] by Phan *et al.* [24]. In another work, Lee *et al.* [157] proposed to associate a model's learning process with expert-defined EEG patterns. These patterns were used as templates for the first convolutional kernels of a CNN and were located in a test EEG signal via cosine similarity maximization to achieve interpretability. Al-Hussaini *et al.* [158] gained interpretability from the perspective of decision rules by coupling a deep learning model with a regression tree. Prototypes in the high-dimensional embedding space of a CNN were firstly derived and used to generate similarity for each PSG epoch with the expert-defined rules. The similarity scores were then classified by a decision tree. Several resulting splitting rules of the decision tree were found similar to the guidelines for human annotators [10].

Ultimately, future research towards interpretable automatic sleep staging could benefit from explainable deep learning research in general. On the one hand, new backpropagation-based methods (e.g., layer-wise relevance propagation [159, 160], Deep Learning Important FeaTures (DeepLIFT) [161], and integrated gradients [162]) and perturbation-based methods (e.g., occlusion sensitivity [163], representation erasure [164], meaningful perturbation [165], and prediction difference analysis [166]) can be explored to improve models' explanation via scientific visualization of characteristics of an input that influence the output of a model. Designing intrinsically explainable deep networks, like [24, 63], that can jointly optimize model performance and provide explanations as part of the model output is another potential direction. These intrinsic methods are probably more desirable than the post-hoc methods that seek explanation of models that were never designed to be explainable in the first place. On the other hand, there are model distillation approaches [167,168] in which the knowledge encoded within a deep learning model (*i.e.*, the "black-box" model) is distilled into a "white-box" model which is meant to identify the decision rules influencing the outputs of the deep learning model, as shown in [158]. The distilled models, potentially simple and interpretable, such as decision tree or logistic regression, offer the explanation power while still achieving reasonable performance. In this way, one could alleviate the compromise between interpretability and prediction accuracy to some extent.

However, it remains an open question how the explainability of a model could be objectively quantified, evaluated, and compared. Our opinion is that an explainable AI system for automatic sleep scoring should be inspired by the way a sleep expert performs manual scoring to provide interpretability to (1) whether the features and the rules resulted from an algorithm are clinically relevant to and underpin sleep and (2) how the decision on a target epoch is made under the influence of its neighboring epochs given their strong dependency due to the continuous nature of sleep. However, answering these questions in automatic sleep staging is tricky. First, most of the existing approaches explain the models via the prism of expert-defined rules and features, however, many of these features are not well-defined while the majority of the rules for human annotators are vague [10]. Second, these features and rules cannot be used in many scenarios, for instance, wearable EEG [27,103,169] since the underlying signals are different from scalp EEGs and not readily interpretable for a human scorer. Third, in practice, many features and rules learned by these networks do not conform to the established features and rules. However, these challenges point in the direction of moving beyond interpretability. We envision that efforts should also be spent on understanding the disharmonizing features and rules resulted purely from data. Clinical explanations for them would potentially help us to gain further insights about underlying neurophysiology of human sleep, which, in turn, could be used to update the manual-scoring features and rules [158, 170].

4.7. Data mismatch due to distributional shifts between datasets/cohorts

Sleep data typically come from different sources with a wide range of institutions, demographics, diseases, modalities, devices, and acquisition conditions. As a result, these mismatches violate the data assumption of being independent and identically distributed (i.i.d.) required for a machine learning system. Even when a deep-learning sleep staging model is trained on large amounts of data, resulting in powerful hierarchical representations, the discrepancies are still computationally significant, degrading the accuracy of sleep staging models on unseen data with a shift (*i.e.*, mismatch) in their distribution [59]. A naive solution for this problem is to form training data from as many conditions as possible [21], ideally from all types of conditions that will be foreseeably encountered in the deployment phase. However, this is expensive, time-consuming, and infeasible. In addition, novel setups will likely emerge in the study of particular sleep disorders [5, 6] or when exploring the feasibility of new monitoring devices [26, 28, 103, 110, 171].

As the data mismatches cannot be simply reversed by signal preprocessing, approaches to migrate a pretrained model to a target cohort with an unseen condition (*e.g.*, via *transductive* transfer learning or domain adaptation [59, 110, 111, 172–174]) have been adopted. While most existing works on this direction utilized a large labelled database for model pretraining in a supervised fashion, semi-supervised [175] and unsupervised (i.e. self-supervised) [176–180] training regimes would further allow leveraging unprecedentedly large amounts of unlabelled data for this purpose. However, before migrating a pretrained model to a target domain, distributional shifts need to be detected and quantified to indicate the variation of the model and whether the migration is necessary or not. Entropy of the probability outputs could potentially serve this purpose [181]. Then, methods for model migration from a source domain to a target domain can be categorized depending on the availability of labelled data in the target domain. In the best case when all data is labelled, supervised domain adaptation methods [182], such as [59, 110, 111], appear to be most sensible. In these methods, a pretrained model needs to undergo a fine tuning process, *i.e.*, the model is further trained in a supervised fashion using the target domain's labelled data. When only a part of the target domain data are labelled, supervised domain adaptation is, in essence, still feasible if the amount of labelled data is sufficient. Otherwise, fine tuning using a small amount of data will be exposed to a great risk of overfitting. In any case, a better solution is to exploit semi-supervised domain adaptation methods, that incorporate semi-supervised learning (SSL) [175] and domain adaptation [182], to leverage both labelled and unlabelled data at the same time. For example, a pretrained model can be fine tuned to simultaneously minimize the sum of supervised classification and unsupervised reconstruction cost functions [183]. In the worst case when all the data are unlabelled, unsupervised domain adaptation will be a natural choice [184]. For example, encouraging results have been reported using adversarial domain adaption methods to match the feature distributions of the source and target domains via gradient reversal from a domain classifier that is tasked to discriminate between the two domains [173, 174, 185]. A pretrained network can be also be adapted to a target domain by modulating the domain-specific statistics of deep features stored in the network's normalization layers like batch normalization [186]. While the reliance on target data labels are costly as human scoring is required, in general, the performance gains are proportional to the amount of labelled data. With the same target-domain data, the gains from supervised transfer learning methods are expected to be higher than semi-supervised ones which are in turn higher than unsupervised ones. However, success of model migration is also subject to training strategies, network-architecture choices and datasets [59].

An alternative approach to the domain adaptation is data mapping. This has remained mostly unexplored for sleep data. The difference is that domain adaptation requires modification of the model parameters while data mapping aims to modify the data to map them from one domain to another. To this end, a mapping function can be learned to map from a certain target domain to the source domain. Evaluating a sleep staging model on a target domain, the test data is firstly fed to the domain mapping function to make them look like the source data before sleep staging takes place. Inspired by the success of generative adversarial networks (GAN) [187] in image-to-image translation, subject-to-subject or sequence-to-sequence PSG mapping could potentially be done similarly with these GAN variants. However, the challenge here is that we may wish to achieve the mapping by modifying the traits of the mismatched factors in data while preserving the sequential nature of the sleep data.

Another approach to align source and target domains is to force a sleep staging model to learn domain-invariant feature representations [110, 188]. In other words, the features learned by the model follow the same distribution no matter whether the input are from the source or target domain, and so representing the underlying sleep stages while being agnostic to other factors. As a consequence, the model trained on the source domain can generalize well to the target domain without the necessity of modification of the model parameters or data mapping. One way to achieve this is to minimize the distances (*e.g.*, Wasserstein distance [189, 190]) between the distributions during training in addition to the classification task. An alternative to distance minimization is to incorporate reconstruction losses [191, 192] to encourage the learned features to reconstruct well either the target domain data or both the source and target domain data. Another possibility is to rely on an adversarial domain classifier which is tasked to discriminate the source and target domain. In light of adversarial training as in a GAN [187], the idea is then to train the sleep staging model to learn the features such that the domain classifier is unable to distinguish from which domain the features originated [188]. However, this approach requires data from both domains to be available at the training phase.

4.8. Heterogeneity: a challenge beyond data mismatch

Apart from the data-mismatch challenge discussed in Section 4.7, heterogeneity is another challenge emerging from the data originated from different sources, or even from different subjects of the same cohort. Typically the number of channels and modalities in PSG recordings varies significantly due to the differences in channel layout and recording setup. This is not a major challenge in many existing network architectures relying on a single channel of one or a few modalities (*i.e.*, EEG, EOG, and EMG) [18, 19, 23, 62]. However, it limits the applicability of those utilizing a larger number of channels [47, 66] when one or more employed channels are missing in test data. In practice, different channels and modalities manifest different perspectives of the underlying neurophysiological processes in human sleep. For example, Alpha rhythm appears most clearly in occipital lobe, sawtooth waves characterizing REM are best captured in central lobe and K-complex events, the hallmark of N2, are best observed in central lobe [10]. As a result, consolidating information from all available channels of PSG data in a holistic view would potentially improve sleep-staging performance. This will also facilitate model building from an unprecedented amount of data gathered from different sources [6, 22, 62].

Fortunately, tackling this challenge probably does not need to design entirely new modelling paradigms but could build upon the current ones. One could devise an intermediate layer that interfaces the heterogeneous raw data and a network. This layer aims to amalgamate all available channels and modalities to form an input with a fixed number of channels which are ready to be fed to any existing network architecture. As a result, existing state-of-the-art models would be invigorated owing to the more informative input. Guillot and Thorey [22] proposed such a layer using a multi-head attention layer with N heads to map an input with varying number of channels to N channels. Potentially, this could also be done using an across-channel pooling operator, *e.g.*, average pooling or max pooling. Channel mapping, for example via a convolutional 1×1 kernel [193], could also map the varying inputs into any fixed number of channels and further enrich the resulting channels via non-linear activations. Ideally, such an interface layer should be integrated to an existing sleep staging model and trained jointly on the classification task in an end-to-end fashion. However, automatic channel quality control should be put in place to exclude bad-quality channels with, for example, highly noisy [6] and excessive data missing [102].

4.9. Subjectivity in model building

It is well-known that manual scoring of PSG is highly subjective. Many previous studies consistently reported the consensus among human scorers around a Cohen's kappa of 0.76 [150, 194]. The consensus is particularly poor on epochs manifesting characteristics of more than one sleep stages. Examples are N1 stage, epochs close to the boundary of two stages, and data from patients with fragmented sleep. So far, it has been a common practice to use the subjective and noisy labels annotated by a single scorer for model training as if they are perfect. Thus, the scorer's subjectivity is unavoidably transferred into the trained model. Reducing such subjectivity is necessary but remains largely uncharted.

The above-mentioned subjectivity can be alleviated by scoring at a fine-grained temporal resolution (*i.e.*, smaller epochs [6, 21] or even samples [61, 62]). However, the fine-grained supervision signals in these works were still derived from 30-second epoch annotation, and therefore, the subjectivity continued to exist. A more proper treatment would be to consider the labels as noisy by nature (in practice, one-hot ground-truth labels are unknown or even not in existence) and robust training methods should be devised to manage the noisy label [195, 196]. It also remains an open question how to train a model with multiple supervision signals (*i.e.*, annotations of two or more human scorers) at the same time rather than a single supervision signal as usual. Doing this will encourage the model to adapt to the scoring style of a cohort of scorers. Note that, this is different from averaging labels of multiple scorers [64] which eventually results in a single supervision signal.

4.10. Scorer personalization

Since it is very likely that a model trained with the annotation from a single scorer as described in Section 4.9 will mimic (*i.e.*, overfit) the scorer's style, the resulting "subjective" model poses another challenge from the end-user perspective that is worth being discussed separately here. Imagine a clinical scenario when a clinician is an end-user of the scoring system. The trained model can be reasonably viewed as a digital twin of the original scorer who labelled the data; hence, it will face disagreement on staging decisions with a new human scorer (*i.e.*, an end-user in this case) as similar as disagreement between two human scorers. For maximum adoption by the clinician, this raises the need for the model to gradually adjust to the scoring style of the new scorer. Readers should note that this scorer personalization problem is orthogonal to the data personalization discussed previously in Section 4.1.1.

Tackling this challenge requires a closed-loop interaction between the model and the end-user [197]. On the one hand, the disagreed staging decisions need to be first identified. This could be done via uncertainty quantification, for example using entropy-based metrics [24, 181] or an ensemble with Monte Carlo dropout [198], as the model's decisions with low confidence are more likely to be disagreed ones. Although this approach can isolate a large portion of wrong decisions, a pitfall here is that an algorithm may output contentious decisions with a very high confidence. This anomalous behavior has been studied in many prior works [199]. While understanding this behavior in the context of human sleep is an interesting subject on its own, methods will also need to be developed to identify wrong decisions associated with high confidence. Model's interpretability (see Section 4.6.2) with the aid of convenient user-interface and user-experience design will be equally important to allow the end-user to scrutinize the potentially wrong decisions and make necessary corrections in an interactive manner. On the other hand, given the end-user's feedback, learning methods will be needed to incrementally adapt the model using the newly labelled data in an open-ended fashion. Approaches for continual learning [200], such as the meta-learning method used in [201], stands out as promising candidates. While these methods are required to learn from sequential, and potentially small, data, they also need to overcome catastrophic forgetting [202], the central issue in this learning setting.

5. Conclusions

Benefits of automatic tools for sleep scoring have driven research in automatic sleep staging for many decades. Promising results recently achieved by deep learning based methods have given incentives for a large number of research studies, now resulting in solutions that have comparable performance to sleep experts on the sleep scoring task, at least on healthy individuals. This methodological development has benefited immensely from the initiatives for free access to large collections of de-identified sleep data and open-source tools and techniques of many experiments available for researchers around the world. We perceive this achievement as a small yet important milestone in the use of AI in sleep research. However, many challenges still have to be overcome for these AI tools to prove their clinical usefulness. Next to the challenge of sleep disorders, issues related to data heterogeneity, model's explainability and subjectivity will require more attention. In order to bring sleep monitoring outside sleep labs to daily living environments, prospective studies also have to be conducted to improve robustness to low signal quality of mobile EEG devices and limited amounts of training data and to address issues related to privacy and longitudinal monitoring. Once these challenges have been overcome, we expect AI-based sleep scoring tools will play an important role in day-to-day sleep practice, complementing the increasing need for trained sleep experts and benefiting millions in need of accurate sleep assessment, monitoring, and treatment options for many sleep disorders.

Author contributions

H. Phan and K. Mikkelsen contributed equally.

Acknowledgement

H. Phan is supported by a Turing Fellowship under the EPSRC grant EP/N510129/1.

References

[1] J. M. Siegel, "Clues to the functions of mammalian sleep," *Nature*, vol. 437, no. 27, pp. 1264–1271, 2005.

[2] P. Maquet, "The role of sleep in learning and memory," *Science*, vol. 294, no. 5544, pp. 1048–1052, 2001.

[3] Ignacio Perez-Pozuelo, Bing Zhai, Joao Palotti, Raghvendra Mall, Michaël Aupetit, Juan M. Garcia-Gomez, Shahrad Taheri, Yu Guan, and Luis Fernandez-Luque, "The future of sleep health: a data-driven revolution in sleep science and medicine," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–15, Mar. 2020, Number: 1 Publisher: Nature Publishing Group.

[4] R. G. Norman, M. A Scott, I. Ayappa, J. A. Walsleben, and D. M. Rapoport, "Sleep continuity measured by survival curve analysis," *Sleep*, vol. 29, no. 12, pp. 1625–1631, 2006.

[5] N. Cooray, F. Andreotti, C. Lo, M. Symmonds, M. T. M. Hu, and M. De Vos, "Detection of rem sleep behaviour disorder by automated polysomnography analysis," *Clinical Neurophysiology*, vol. 130, no. 4, pp. 505–514, 2019.

[6] J. B. Stephansen *et al.*, "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy," *Nature Communications*, vol. 9, no. 1, pp. 5229, 2018.

[7] J. A. E. Christensen, O. Carrillo, E. B. Leary, P. E. Peppard, T. Young, H. B. D. Sorensen, P. Jennum, and E. Mignot, "Sleep-stage transitions during polysomnographic recordings as diagnostic features of type 1 narcolepsy," *Sleep Medicine*, vol. 16, pp. 1558–1566, 2015.

[8] C. V. Senaratna, Jennifer L Perret, Caroline J Lodge, Adrian J Lowe, Brittany E Campbell, Melanie C Matheson, Garun S Hamilton, and Shyamali C Dharmage, "Prevalence of obstructive sleep apnea in the general population: A systematic review," *Sleep Med Rev*, vol. 34, pp. 70–81, 2017.

[9] S. Redline, P. V. Tishler, M. Schluchter, J. Aylor, K. Clark, and G. Graham, "Risk factors for sleep-disordered breathing in children. associations with obesity, race, and respiratory problems," *American Journal of Respiratory and Critical Care Medicine*, vol. 159, pp. 1527–1532, 1999.

[10] Richard B. Berry, Rita Brooks, Charlene Gamaldo, Susan M. Harding, Robin M. Lloyd, C. L. Marcus, and Bradley V. Vaughn, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*, American Academy of Sleep Medicine, 2.3 edition, 2016.

[11] A. C. Krieger, Ed., *Social and Economic Dimensions of Sleep Disorders, An Issue of Sleep Medicine Clinics*, Elsevier, 2017.

[12] V. K. Chattu *et al.*, "The global problem of insufficient sleep and its serious public health implications," *Healthcare (Basel)*, vol. 7, no. 1, 2019.

[13] Institute of Medicine, *Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem*, Washington DC: The National Academies Press, 2006.

[14] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, pp. e215–e220, 2000.

[15] G. Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, and S. Redline, "The national sleep research resource: towards a sleep data commons," *J Am Med Inform Assoc.*, vol. 25, no. 10, pp. 1351–1358, 2018.

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[17] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, MIT Press, 2016.

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

[19] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging," *IEEE Trans. on Neural Systems and Rehabilitation Engineering (TNSRE)*, vol. 27, no. 3, pp. 400–410, 2019.

[20] S. Biswal *et al.*, "Expert-level sleep scoring with deep neural networks," *J Am Med Inform Assoc.*, vol. 25, no. 12, pp. 1643–1650, 2018.

[21] A. N. Olesen *et al.*, "Automatic sleep stage classification with deep residual networks in a mixed-cohort setting," *Sleep*, vol. 44, no. 1, pp. zsaa161, 2021.

[22] Antoine Guillot and Valentin Thorey, "RobustSleepNet: Transfer learning for automated sleep staging at scale," *arXiv:2101.02452 [cs, eess, stat]*, Jan. 2021, arXiv: 2101.02452.

[23] H. Phan *et al.*, "XSleepNet: Multi-view sequential model for automatic sleep staging," *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[24] H. Phan *et al.*, "SleepTransformer: Automatic sleep staging with interpretability and uncertainty quantification," *IEEE Trans. on Biomedical Engineering (TBME)*, 2022.

[25] Tomi Miettinen, Katja Myllymaa, Susanna Westeren-Punnonen, Jari Ahlberg, Taina Hukkanen, Juha Toyras, Reijo Lappalainen, Esa Mervaala, Kirsi Sipila, Sami Myllymaa, Tomi Miettinen, Katja Myllymaa, Susanna Westeren-Punnonen, Jari Ahlberg, Taina Hukkanen, Juha Toyras, Reijo Lappalainen, Esa Mervaala, Kirsi Sipila, and Sami Myllymaa, "Success Rate and Technical Quality of Home Polysomnography With Self-Applicable Electrode Set in Subjects With Possible Sleep Bruxism," *IEEE journal of biomedical and health informatics*, vol. 22, no. 4, pp. 1124–1132, July 2018.

[26] Pierrick J. Arnal, Valentin Thorey, Michael E. Ballard, Albert Bou Hernandez, Antoine Guillot, Hugo Jourde, Mason Harris, Mathias Guillard, Pascal Van Beers, Mounir Chennaoui, and Fabien Sauvet, "The Dreem Headband as an Alternative to Polysomnography for EEG Signal Acquisition and Sleep Staging," *bioRxiv*, p. 662734, June 2019.

[27] Kaare B. Mikkelsen, Yousef R. Tabar, Simon L. Kappel, Christian B. Christensen, Hans O. Toft, Martin C. Hemmsen, Mike L. Rank, Marit Otto, and Preben Kidmose, "Accurate whole-night sleep monitoring with dry-contact ear-EEG," *Scientific Reports*, vol. 9, no. 1, pp. 16824, Nov. 2019, Number: 1 Publisher: Nature Publishing Group.

[28] Kaare B. Mikkelsen, James K. Ebajemito, Maria A. Bonmati-Carrion, Nayantara Santhi, Victoria L. Revell, Giuseppe Atzori, Ciro della Monica, Stefan Debener, Derk-Jan Dijk, Annette Sterr, and Maarten de Vos, "Machine-learning-derived sleep–wake staging from around-the-ear electroencephalogram outperforms manual scoring and actigraphy," *Journal of Sleep Research*, vol. 0, no. 0, pp. e12786, Nov. 2018.

[29] L. Fiorillo *et al.*, "Automated sleep scoring: A review of the latest approaches," *Sleep Medicine Reviews*, vol. 48, pp. 101204, 2019.

[30] O. Faust *et al.*, "A review of automated sleep stage scoring based on physiological signals for the new millennia," *Comput Methods Programs Biomed*, vol. 176, pp. 81–91, 2019.

[31] Y. Roy *et al.*, "Deep learning-based electroencephalography analysis: a systematic review," *J Neural Eng*, vol. 16, no. 5, pp. 051001, 2019.

[32] M. Radha, P. Fonseca, A. Moreau, M. Ross, A. Cerny, P. Anderer, X. Long, and R. M. Aarts, "A deep transfer learning approach for wearable sleep stage classification with photoplethysmography," *npj Digital Medicine*, vol. 4, no. 135, 2021.

[33] B. Zhai, I. Perez-Pozuelo, E. A. D. Clifton, J. Palotti, and Y. Guan, "Making sense of sleep: Multimodal sleep stage classification in a large, diverse population using movement and cardiac sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 2, pp. 1–33, 2020.

[34] E. Dafna, A. Tarasiuk, and Y. Zigel, "Sleep staging using nocturnal sound analysis," *Scientific Reports*, vol. 8, no. 13474, 2018.

[35] X. Long, R. Otte, E. van der Sanden, J. Werth, and T. Tan, "Video-based actigraphy for monitoring wake and sleep in healthy infants: A laboratory study," *Sensors*, vol. 19, no. 5, pp. 1075, 2019.

[36] S. Toften, S. Pallesen, M. Hrozanovad, F. Moen, and J. Gronli, "Validation of sleep stage classification using non-contact radar technology and machine learning (somnofy)," *Sleep Medicine*, pp. 54–61, 2020.

[37] M. Piriyajitakonkij, P. Warin, P. Lakhan, P. Leelaarporn, N. Kumchaiseemak, S. Suwajanakorn, T. Pianpanit, N. Niparnan, S. C. Mukhopadhyay, and T. Wilaiprasitporn, "SleepPoseNet: Multi-view learning for sleep postural transition recognition using uwb," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 4, pp. 1305–1314, 2021.

[38] Yi-Qi Lin and Sheng-Di Chen, "RBD: A red flag for cognitive impairment in Parkinson's disease?," *Sleep Medicine*, vol. 44, pp. 38–44, Apr. 2018.

[39] H. Dong *et al.*, "Mixed neural network approach for temporal sleep stage classification," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 2, pp. 324–333, 2018.

[40] R. Wei, X. Zhang, J. Wang, and X. Dang, "The research of sleep staging based on single-lead electrocardiogram and deep neural network," *Biomed Eng Lett.*, vol. 8, no. 1, pp. 87–93, 2018.

[41] O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou, "Automatic sleep stage scoring with single-channel EEG using convolutional neural networks," *arXiv:1610.01683*, 2016.

[42] Albert Vilamala, Kristoffer H. Madsen, and Lars K. Hansen, "Deep convolutional neural networks for interpretable analysis of EEG sleep stage scoring," in *Proc. MLSP*, 2017.

[43] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "DNN filter bank improves 1-max pooling CNN for single-channel EEG automatic sleep stage classification," in *Proc. EMBC*, 2018, pp. 453–456.

[44] A. Malafeev *et. al.*, "Automatic human sleep stage scoring using deep neural networks," *Front Neurosci.*, vol. 12, no. 781, 2018.

[45] S. Biswal *et al.*, "SLEEPNET: Automated sleep staging system via deep learning," *arXiv preprint arXiv:1707.08262*, 2017.

[46] H. Sun *et al.*, "Large-scale automated sleep staging," *SLEEP*, vol. 40, no. 10, pp. zsx139, 2017.

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

[48] F. Andreotti, H. Phan, N. Cooray, C. Lo, M. T. M. Hu, and M. De Vos, "Multichannel sleep stage classification and transfer learning using convolutional neural networks," in *Proc. EMBC*, 2018, pp. 171–174.

[49] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Joint classification and prediction CNN framework for automatic sleep stage classification," *IEEE Trans. Biomedical Engineering (TBME)*, vol. 66, no. 5, pp. 1285–1296, 2019.

[50] A. Sors *et al.*, "A convolutional neural network for sleep stage scoring from raw single-channel eeg," *Biomed Signal Process Control*, vol. 42, pp. 107–114, 2018.

[51] F. Andreotti, H. Phan, and M. De Vos, "Visualising convolutional neural network decisions in automatic sleep scoring," in *Proc. Joint Workshop on Artificial Intelligence in Health (AIH)*, 2018, pp. 70–81.

[52] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Automatic sleep stage classification using single-channel EEG: learning sequential features with attention-based recurrent neural networks," in *Proc. EMBC*, 2018, pp. 1452–1455.

[53] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computing*, vol. 9, no. 8, pp. 1735–1780, 1997.

[54] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. EMNLP*, 2014, pp. 1724–1734.

[55] E. Hartmann, "The 90-minute sleep-dream cycle," *Archives of General Psychiatry*, vol. 18, no. 3, pp. 280–286, 1968.

[56] I. Feinberg and T. C. Floyd, "Systematic trends across the night in human sleep cycles," *Psychophysiology*, vol. 16, no. 3, pp. 283–291, 1979.

[57] C. Sun *et al.*, "A hierarchical neural network for sleep stage classification based on comprehensive feature learning and multi-flow sequence learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 5, pp. 1351–1366, 2020.

[58] H. Ghimatgar *et al.*, "Neonatal eeg sleep stage classification based on deep learning and hmm," *Journal of Neural Engineering*, vol. 17, no. 3, pp. 036031, 2020.

[59] H. Phan, O. Y. Chén, P. Koch, Z. Lu, I. McLoughlin, A. Mertins, and M. De Vos, "Towards more accurate automatic sleep staging via deep transfer learning," *IEEE Trans. on Biomedical Engineering*, vol. 68, no. 6, pp. 1787–1798, 2021.

[60] H. Seo *et al.*, "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*, vol. 61, pp. 102037, 2020.

[61] M. Perslev *et al.*, "U-time: A fully convolutional network for time series segmentation applied to sleep staging," in *Proc. NeurIPS*, 2019, pp. 4417–4428.

[62] M. Perslev *et al.*, "U-Sleep: resilient high-frequency sleep staging," *npj Digital Medicine*, vol. 4, no. 72, 2021.

[63] Z. Jia, Y. Lin, J. Wang, X. Wang, P. Xie, and Y. Zhang, "SalientSleepNet: Multimodal salient wave detection network for sleep staging," *IJCAI*, 2021.

[64] A. Guillot, F. Sauvet, E. H. During, and V. Thorey, "Dreem open datasets: Multi-scored sleep datasets to compare human and automated sleep staging," *IEEE Transactions on Neural*

Systems and Rehabilitation Engineering, vol. 28, no. 9, pp. 1955–1965, 2020.

[65] W. Neng, J. Lu, and L. Xu, "Ccrrsleepnet: A hybrid relational inductive biases network for automatic sleep stage classification on raw single-channel eeg," *Brain Sciences*, vol. 11, no. 456, 2021.

[66] Z. Jia *et al.*, "GraphSleepNet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification," in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, 2020, pp. 1324–1330.

[67] E. Eldele *et al.*, "An attention-based deep learning approach for sleep stage classification with single-channel eeg," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 809–818, 2021.

[68] J. Cohen, "A coefficient of agreement for nominal scales," *Educ Psychol Meas.*, vol. 20, pp. 37–46, 1960.

[69] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG," *IEEE Trans. on Biomedical Engineering*, vol. 47, no. 9, pp. 1185–1194, 2000.

[70] C. O'Reilly, N. Gosselin, J. Carrier, and T. Nielsen, "Montreal archive of sleep studies: An open-access resource for instrument benchmarking & exploratory research," *Journal of Sleep Research*, pp. 628–635, 2014.

[71] M. M. Ghassemi *et al.*, "You snooze, you win: the physionet/computing in cardiology challenge 2018," in *Proc. 2018 Computing in Cardiology Conference (CinC)*, 2018, vol. 45, pp. 1–4.

[72] S. F. Quan *et al.*, "The sleep heart health study: design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–1085, 1997.

[73] S. Khalighi, T. Sousa, J. M. dos Santos, and U. Nunes, "Isruc-sleep: a comprehensive public dataset for sleep researchers," *Methods Programs Biomed.*, vol. 124, pp. 180–192, 2016.

[74] M. G. Terzano, L. Parrino, A. Smerieri, R. Chervin, S. Chokroverty, C. Guilleminault, M. Hirshkowitz, M. Mahowald, H. Moldofsky, A. Rosa, R. Thomas, and A. Walters, "Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (cap) in human sleep," *Sleep Medicine*, vol. 3, no. 2, pp. 187–199, 2002.

[75] X. Chen *et al.*, "Racial/ethnic differences in sleep disturbances: the multi-ethnic study of atherosclerosis (MESA)," *Sleep*, vol. 38, pp. 877–888, 2015.

[76] T. Blackwell *et al.*, "Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: The osteoporotic fractures in men sleep study," *J. Am. Geriatr. Soc.*, vol. 59, pp. 2217–2225, 2011.

[77] Y. Song *et al.*, "Relationships between sleep stages and changes in cognitive function in older men: the MrOS sleep study," *Sleep*, vol. 38, pp. 411–421, 2015.

[78] C. L. Marcus *et al.*, "Childhood adenotonsillectomy trial (chat). a randomized trial of adenotonsillectomy for childhood sleep apnea," *N Engl J Med.*, vol. 368, no. 25, pp. 2366–2376, 2013.

[79] S. Redline *et al.*, "The childhood adenotonsillectomy trial (chat): rationale, design, and challenges of a randomized controlled trial evaluating a standard surgical procedure in a pediatric population," *Sleep*, vol. 34, no. 11, pp. 1509–1517, 2011.

[80] S. Mousavi, F. Afghah, and R. Acharya, "SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PLoS One*, vol. 14, no. 5, pp. e0216456, 2019.

[81] K. Chen *et al.*, "Sleep staging from single-channel EEG with multi-scale feature and contextual information," in *Proc. 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2019, vol. 23, pp. 1159–1167.

[82] A. Supratak and Y. Guo, "TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG," in *Proc. 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2020.

[83] M. A. Jaoude *et al.*, "Expert-level automated sleep staging of long-term scalp electroencephalography recordings using deep learning," *Sleep*, vol. 43, no. 11, pp. zsaa112, 2020.

[84] H. Korkalainen *et al.*, "Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 7, pp. 2073–2081, 2020.

[85] W. Qu *et al.*, "A residual based attention model for eeg based sleep staging," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2833–2843, 2020.

[86] W. Chen, Y. Yang, and P. Yang, "Hnsleepnet: A novel hybrid neural network for home healthcare automatic sleep staging with raw single-channel eeg," in *2020 IEEE 18th International Conference on Industrial Informatics (INDIN)*, 2020, pp. 555–560.

[87] Y. Li *et al.*, "An automatic sleep staging model combining feature learning and sequence learning," in *Proc. 12th International Conference on Advanced Computational Intelligence (ICACI)*, 2020, pp. 419–425.

[88] H. Nie, S. Tu, and L. Xu, "Recsleepnet: An automatic sleep staging model based on feature reconstruction," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2021, pp. 1458–1461.

[89] W. G. Coon and N. M. Punjabi, "Automatic sleep staging using a small-footprint sensor array and recurrent-convolutional neural networks," in *Proc. 10th International IEEE/EMBS Conference on Neural Engineering (NER)*, 2021, pp. 1144–1147.

[90] H. Phan, A. Mertins, and M. Baumert, "Pediatric automatic sleep staging: Deep learning ensemble improves accuracy and reduces predictive uncertainty," *arXiv preprint arXiv:2108.10211*, 2022.

[91] S. A. Imtiaz, "A systematic review of sensing technologies for wearable sleep staging," *Sensors*, vol. 21, no. 5, pp. 1562, 2021.

[92] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G. Hauptmann, "Contrastive Adaptation Network for Unsupervised Domain Adaptation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, June 2019, pp. 4888–4897, IEEE.

[93] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A. Efros, "Unsupervised Domain Adaptation through Self-Supervision," *arXiv:1909.11825 [cs, stat]*, Sept. 2019, arXiv: 1909.11825.

[94] M. Padmanabhan, L.R. Bahl, D. Nahamoo, and M.A. Picheny, "Speaker clustering and transformation for speaker adaptation in speech recognition systems," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 71–77, Jan. 1998.

[95] Robert M. French, "Catastrophic forgetting in connectionist networks," *Trends in Cognitive Sciences*, vol. 3, no. 4, pp. 128–135, Apr. 1999.

[96] J. Buckelmüller, H-P P. Landolt, H. H. Stassen, and P. Achermann, "Trait-like individual differences in the human sleep electroencephalogram.," *Neuroscience*, vol. 138, no. 1, pp. 351–356, 2006.

[97] L. A. Finelli, P. Achermann, and A. A. Borbély, "Individual 'fingerprints' in human sleep EEG topography.," *Neuropsychopharmacology*, vol. 25, no. 5 Suppl, Nov. 2001.

[98] Martin Hemmsen, Kaare Mikkelsen, Mike Rank, and Preben Kidmose, "Long-term monitoring of trait-like characteristics of the sleep electroencephalogram using ear-EEG," *Sleep*, vol. 44, no. Supplement_2, pp. A109–A109, May 2021.

[99] Adrienne M. Tucker, David F. Dinges, and Hans P. A. Van Dongen, "Trait interindividual differences in the sleep physiology of healthy young adults," *Journal of Sleep Research*, vol. 16, no. 2, pp. 170–180, June 2007.

[100] Huy Phan, Kaare Mikkelsen, Oliver Y. Chén, Philipp Koch, Alfred Mertins, Preben Kidmose, and Maarten De Vos, "Personalized automatic sleep staging with single-night data: a pilot study with Kullback–Leibler divergence regularization," *Physiological Measurement*, vol. 41, no. 6, pp. 064004, July 2020, Publisher: IOP Publishing.

[101] E. C.-P. Chua, S.-C. Yeo, I. T.-G. Lee, L.-C. Tan, P. Lau, S. S. Tan, I. H. Mien, and J. J. Gooley, "Individual differences in physiologic measures are stable across repeated exposures to total sleep deprivation," *Physiological Reports*, vol. 2, no. 9, pp. e12129, Sept. 2014.

[102] K. B. Mikkelsen, H. Phan, M. L. Rank, M. C. Hemmsen, M. de Vos, and P. Kidmose, "Sleep monitoring using ear-centered setups: Investigating the influence from electrode configurations," *IEEE Transactions on Biomedical Engineering*, 2021.

[103] Kaare B. Mikkelsen, David Bové Villadsen, Marit Otto, and Preben Kidmose, "Automatic sleep staging using ear-EEG," *BioMedical Engineering OnLine*, vol. 16, no. 1, pp. 111, Sept. 2017.

[104] R. Palaniappan and D. P. Mandic, "Biometrics from brain electrical activity: A machine learning approach," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 738–742, 2007.

[105] Kaare B. Mikkelsen, Yousef R. Tabar, Christian B. Christensen, and Preben Kidmose, "EEGs Vary Less Between Lab and Home Locations Than They Do Between People," *Frontiers in Computational Neuroscience*, vol. 15, 2021, Publisher: Frontiers.

[106] Sirin W. Gangstad, Kaare B. Mikkelsen, Preben Kidmose, Yousef R. Tabar, Sigge Weisdorf, Maja H. Lauritzen, Martin C. Hemmsen, Lars K. Hansen, Troels W. Kjaer, and Jonas Duun-Henriksen, "Automatic sleep stage classification based on subcutaneous EEG in patients with epilepsy," *BioMedical Engineering OnLine*, vol. 18, no. 1, pp. 106, Oct. 2019.

[107] B. Koley and D. Dey, "An ensemble system for automatic sleep stage classification using single channel EEG signal," *Computers in Biology and Medicine*, vol. 42, no. 12, pp. 1186–95, 2012.

[108] Alexander Neergaard Olesen, Poul Jennum, Emmanuel Mignot, and Helge B. D. Sorensen, "Deep transfer learning for improving single-EEG arousal detection," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, July 2020, pp. 99–103, ISSN: 2694-0604.

[109] Panteleimon Chriskos, Christos A. Frantzidis, Christiane M. Nday, Polyxeni T. Gkivogkli, Panagiotis D. Bamidis, and Chrysoula Kourtidou-Papadeli, "A review on current trends in automatic sleep staging through bio-signal recordings and future challenges," *Sleep Medicine Reviews*, vol. 55, pp. 101377, Feb. 2021.

[110] E. R. M. Heremans *et al.*, "Feature matching as improved transfer learning technique for wearable EEG," *arXiv preprint arXiv:2201.00644*, 2021.

[111] H. Phan, O. Y. Chén, P. Koch, A. Mertins, and M. De Vos, "Deep transfer learning for single-channel automatic sleep staging with channel mismatch," in *Proc. 27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[112] J. Fan *et al.*, "Eeg data augmentation: towards class imbalance problem in sleep staging tasks," *J. Neural Eng.*, vol. 17, no. 5, pp. 056017, 2020.

[113] O. Al Zoubi, C. K. Wong, R. T. Kuplicki, H.-W. Yeh, A. Mayeli, H. Refai, M. Paulus, and J. Bodurka, "Predicting age from brain eeg signals—a machine learning approach," *Front. Aging Neurosci.*, vol. 10, 2018.

[114] P. Wang and J. Hu, "A hybrid model for eeg-based gender recognition," *Cogn Neurodyn.*, vol. 13, pp. 541–554, 2019.

[115] S. M. Alarcao and M. J. Fonseca, "Emotions recognition using eeg signals: A survey," *IEEE Transactions on Affective Computing*, vol. 10, pp. 374–393, 2019.

[116] S. Sangnark, P. Autthasan, P. Ponglertnapakorn, P. Chalekarn, T. Sudhawiyangkul, M. Trakulruangroj, S. Songsermsawad, R. Assabumrungrat, S. Amplod, K. Ounjai, and T. Wilaiprasitporn, "Revealing preference in popular music through familiarity and brain response," *IEEE Sensors Journal*, vol. 21, pp. 14931–14940, 2021.

[117] G. Zhao, Y. Ge, B. Shen, X. Wei, and H. Wang, "Revealing preference in popular music through familiarity and brain response," *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 362–371, 2018.

[118] "Official journal of the european union. general data protection regulation," https: //eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679, accessed 12/02/2021.

[119] B. M. Gaff, H. E. Sussman, and J. Geetter, "Privacy and big data," *Computer*, vol. 47, no. 6, pp. 7–9, 2014.

[120] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient

learning of deep networks from decentralized data," in *Proc. AISTATS*, 2017, pp. 1273–1282.

[121] "Webank. federated ai technology enabler (fate)," https://github.com/FederatedAI/FATE, accessed 16/02/2021.

[122] "Baidu. federated deep learning in paddlepaddle," https://github.com/PaddlePaddle/PaddleFL, accessed 16/02/2021.

[123] "Tensorflow federated: Machine learning on decentralized data," https://www. tensorflow.org/federated, accessed 16/02/2021.

[124] "Openmined. pysyft," https://github.com/OpenMined/PySyft, accessed 22/02/2021.

[125] W. Xia, H. Yin, and N. K. Jha, "Efficient synthesis of compact deep neural networks," *arXiv preprint arXiv:2004.08704*, 2020.

[126] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations," *Journal of Machine Learning Research*, vol. 18, pp. 1–30, 2018.

[127] C. Zhu, S. Han, H. Mao, and W. J. Dally, "Trained ternary quantization," in *Proc. ICLR*, 2017.

[128] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," in *Proc. ICLR*, 2016.

[129] P. Molchanov, S. Tyree, T. Karras amd T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient transfer learning," in *Proc. ICLR*, 2017.

[130] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shift: A zero flop, zero parameter alternative to spatial convolutions," in *Proc. ECCV*, 2018.

[131] B. Wu, A. Wan, X. Yue, P. Jin, S. Zhao, N. Golmant, A. Gholaminejad, J. Gonzalez, and K. Keutzer, "Shift: A zero flop, zero parameter alternative to spatial convolutions," in *Proc. CVPR*, 2018.

[132] B. Baker, O. Gupta, N. Naik, and R. Raskar, "Designing neural network architectures using reinforcement learning," in *Proc. ICLR*, 2017.

[133] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," in *Proc. ICLR*, 2017.

[134] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, "MnasNet: Platform-aware neural architecture search for mobile," in *Proc. CVPR*, 2019.

[135] Alex Iranzo, "Sleep in Neurodegenerative Diseases," *Sleep Medicine Clinics*, vol. 11, no. 1, pp. 1–18, Mar. 2016.

[136] Joan Santamaria, Birgit Höogl, Claudia Trenkwalder, and Donald Bliwise, "Scoring Sleep in Neurological Patients: The Need for Specific Considerations," *Sleep*, vol. 34, no. 10, pp. 1283–1284, Oct. 2011.

[137] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N. Chiang, Zhihao Wu, and Xiaowei Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, pp. 101693, July 2020.

[138] Musa H. Asyali, Richard B. Berry, Michael C. K. Khoo, and Ayse Altinok, "Determining a continuous marker for sleep depth," *Computers in Biology and Medicine*, vol. 37, no. 11, pp. 1600–1609, Nov. 2007.

[139] Simona Carrubba, Paul Young Kim, David E. McCarty, Andrew L. Chesson, Clifton Frilot, and Andrew A. Marino, "Continuous EEG-based dynamic markers for sleep depth and phasic events," *Journal of Neuroscience Methods*, vol. 208, no. 1, pp. 1–9, June 2012.

[140] Christian Berthomier, Xavier Drouot, Maria Herman-Stoïca, Pierre Berthomier, Jacques Prado, Djibril Bokar-Thire, Odile Benoit, Jérémie Mattout, and Marie-Pia d'Ortho, "Automatic Analysis of Single-Channel Sleep EEG: Validation in Healthy Individuals," *Sleep*, vol. 30, pp. 1587–1595, Nov. 2007.

[141] Laure Peter-Derex, Christian Berthomier, Jacques Taillard, Pierre Berthomier, Romain Bouet, Jérémie Mattout, Marie Brandewinder, and Hélène Bastuji, "Automatic analysis of single-channel sleep EEG in a large spectrum of sleep disorders," *Journal of Clinical Sleep Medicine*, Mar. 2021, Publisher: American Academy of Sleep Medicine.

[142] Hongyang Li and Yuanfang Guan, "DeepSleep convolutional neural network allows accurate and fast detection of sleep arousal," *Communications Biology*, vol. 4, no. 1, pp. 1–11, 2021.

[143] O. Asan amd A. E. Bayrak and A. Choudhury, "Artificial intelligence and human trust in healthcare: Focus on clinicians," *J Med Internet Res.*, vol. 22, no. 6, pp. e15154, 2020.

[144] D. S. W. Ting, Y. Liu, P. Burlina, X. Xu, N. M. Bressler, and T. Y. Wong, "Ai for medical imaging goes deep," *Nat Med*, vol. 24, no. 5, pp. 539–40, 2018.

[145] A. Rajkomar *et al.*, "Scalable and accurate deep learning with electronic health records," *npj Digital Medicine*, vol. 1, pp. 18, 2018.

[146] A. H. Ribeiro, M. H. Ribeiro1, G. M. M. Paix ao, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. S. Ferreira, C. R. Andersson, P. W. Macfarlane, W. Meira Jr, T. B. Schön, and A. L. P. Ribeiro, "Intelligent assistive technology for alzheimer's disease and other dementias: a systematic review," *Nat Com*, vol. 11, pp. 1760, 2020.

[147] C. Su, Z. Xu, J. Pathak, and F. Wang, "Deep learning in mental health outcome research: a scoping review," *Translational Psychiatry*, vol. 10, pp. 116, 2020.

[148] M. Ienca, F. Jotterand, B. Elger, M. Caon, A. S. Pappagallo, and R. W. Kressig *et al.*, "Intelligent assistive technology for alzheimer's disease and other dementias: a systematic review," *J Alzheimers Dis*, vol. 56, no. 4, pp. 1301–40, 2017.

[149] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[150] H. Danker-Hopfe *et al.*, "Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard," *J. Sleep Res.*, vol. 18, pp. 74–84, 2009.

[151] D. Gunning and D. W. Aha, "DARPA's explainable artificial intelligence (XAI) program," *AI Magazine*, vol. 40, pp. 44–58, 2019.

[152] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," *arXiv preprint arXiv:1506.06579*, 2015.

[153] P. M. Rasmussen, K. H. Madsen, T. E. Lund, and L. K. Hansen, "Visualization of nonlinear kernel models in neuroimaging by sensitivity maps," *NeuroImage*, vol. 55, no. 3, pp. 1120–1131, 2011.

[154] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[155] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U^2-net: Going deeper with nested u-structure for salient object detection," *Pattern Recognition*, vol. 106, pp. 107404, 2020.

[156] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NIPS*, 2017, p. 5998–6008.

[157] T. Lee, J. Hwang, and H. Lee, "Trier: Template-guided neural networks for robust and interpretable sleep stage identification from eeg recordings," *arXiv Preprint arXiv:2009.05407*, 2020.

[158] I. Al-Hussaini, C. Xiao, M. B. Westover, and J. Sun, "SLEEPER: interpretable sleep staging via prototypes from expert rules," *Proceedings of Machine Learning Research*, vol. 106, pp. 721–739, 2019.

[159] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS ONE*, vol. 10, no. 7, pp. e0130140, 2015.

[160] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, 2018.

[161] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. ICML*, 2017, p. 3145–3153.

[162] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. ICML*, 2017, p. 3319–3328.

[163] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc.*

ECCV, 2014, p. 818–833.

[164] J. Li, W. Monroe, and D. Jurafsky, "Understanding neural networks through representation erasure," *arXiv preprint arXiv:1612.08220*, 2016.

[165] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proc. ICCV*, 2017, p. 3429–3437.

[166] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," in *Proc. ICLR*, 2017.

[167] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proc. ACM SIGKDD*, 2016, p. 1135–1144.

[168] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv:1503.02531*, 2015.

[169] A. Sterr, J. K. Ebajemito, K. B. Mikkelsen, M. A. Bonmati-Carrion, N. Santhi, C. della Monica, L. Grainger, G. Atzori, V. Revell, S. Debener, D.-J. Dijk, and M. De Vos, "Sleep eeg derived from behind-the-ear electrodes (ceegrid) compared to standard polysomnography: A proof of concept study," *Frontiers in Human Neuroscience*, vol. 12, no. 452, 2018.

[170] T. Penzel, X. Zhang, and I. Fietze, "Inter-scorer reliability between sleep centers can teach us what to improve in the scoring rules," *Clin. Sleep Med.*, vol. 9, no. 1, pp. 89–91, 2013.

[171] S. Myllymaa *et al.*, "Assessment of the suitability of using a forehead eeg electrode set and chin emg electrodes for sleep staging in polysomnography," *J. Sleep Res.*, vol. 25, no. 6, pp. 636–645, 2016.

[172] E. Eldele, M. Ragab, Z. Chen, M. Wu, C.-K. Kwoh, X. Li, and C. Guan, "Adversarial domain adaptation with self-training for eeg-based sleep stage classification," *arXiv preprint arXiv:2107.04470*, 2021.

[173] C. Yoo, H. W. Lee, and J. Kang, "Transferring structured knowledge in unsupervised domain adaptation of a sleep staging network," *IEEE Journal of Biomedical and Health Informatics*, 2021.

[174] R. Zhao, Y. Xia, and Y. Zhang, "Unsupervised sleep staging system based on domain adaptation," *Biomedical Signal Processing and Control*, vol. 69, pp. 102937, 2021.

[175] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, pp. 373–440, 2020.

[176] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 4037–4058, 2021.

[177] X. Jiang, J. Zhao, B. Du, and Z. Yuan, "Self-supervised contrastive learning for eeg-based sleep staging," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021.

[178] C. Yang, D. Xiao, M. B. Westover, and J. Sun, "Self-supervised electroencephalogram representation learning for automatic sleep staging," *arXiv preprint arXiv:2110.15278*, 2021.

[179] M. N. Mohsenvand, M. R. Izadi, and P. Maes, "Self-supervised contrastive learning for eeg-based sleep staging," in *Proc. of Machine Learning Research*, 2020, vol. 136, pp. 238–253.

[180] H. Banville *et al.*, "Uncovering the structure of clinical eeg signals with self-supervised learning," *Journal of Neural Engineering*, vol. 18, no. 4, pp. 046020, 2021.

[181] K. B. Mikkelsen, Y. R. Tabar, , and P. Kidmose, "Predicting sleep classification performance without labels," in *Proc. EMBC*, 2020, pp. 645–648.

[182] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Proc. Artificial Neural Networks and Machine Learning (ICANN)*, 2018.

[183] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Proc. NIPS*, 2018, pp. 268–283.

[184] G. Wilso and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Transactions on Intelligent Systems and Technology*, vol. 11, no. 5, pp. 1–46, 2020.

[185] S. Nasiri and G. D. Clifford, "Attentive adversarial network for large-scale sleep staging," in *Proceedings of Machine Learning Research*, 2020, vol. 126, pp. 1–21.

[186] J. Fan *et al.*, "Unsupervised domain adaptation by statistics alignment for deep sleep staging networks," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2022.

[187] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.

[188] H. Zhao, R. T. Des Combes, K. Zhang, and G. Gordon, "On learning invariant representations for domain adaptation," in *Proc. ICML*, 2019, pp. 7523–7532.

[189] S. Chambon, M. N. Galtier, and A. Gramfort, "Domain adaptation with optimal transport improves eeg sleep stage classifiers," in *Proc. 2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, 2018.

[190] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Proc. NIPS*, 2016, pp. 136–144.

[191] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstructionclassification networks for unsupervised domain adaptation," in *Proc. ECCV*, 2016, pp. 597–613.

[192] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Proc. NIPS*, 2016, pp. 343–351.

[193] M. Lin, Q. Chen, and S. Yan:, "Network in network," in *Proc. ICLR*, 2014.

[194] R. S. Rosenberg and S. Van Hout, "The american academy of sleep medicine inter-scorer reliability program: Respiratory events," *Journal of Clinical Sleep Medicine*, vol. 10, pp. 447–454, 2014.

[195] N. Nigam, T. Dutta, and H. P. Gupta, "Impact of noisy labels in learning techniques: A survey," in *Proc. ICDIS*, 2020, pp. 403–411.

[196] B. Han, Q. Yao, T. Liu, G. Niu, I. W. Tsang, J. T. Kwok, and M. Sugiyama, "A survey of label-noise representation learning: Past, present and future," *arXiv preprint arXiv:2011.04406*, 2020.

[197] S.-F. Liang, Y.-H. Shih, P.-Y. Chen, and C.-E. Kuo, "Development of a human-computer collaborative sleep scoring system for polysomnography recordings," *PLoS ONE*, vol. 14, no. 7, pp. e0218948, 2020.

[198] L. Fiorillo, P. Favaro, and F. D. Faraci, "Deepsleepnet-lite: A simplified automatic sleep stage scoring model with uncertainty estimates," *arXiv preprint arXiv:2108.10600*, 2021.

[199] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proc. CVPR*, 2015, pp. 427–436.

[200] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, 2019.

[201] N. Banluesombatkul, P. Ouppaphan, P. Leelaarporn, P. Lakhan, B. Chaitusaney, N. Jaimchariyatam, E. Chuangsuwanich, W. Chen, H. Phan, N. Dilokthanakul, and T. Wilaiprasitporn, "MetaSleepLearner: A pilot study on fast adaptation of bio-signals-based sleep stage classifier to new individual subject using meta-learning," *IEEE Journal of Biomedical and Health Informatics (JBHI)*, vol. 25, no. 6, pp. 1949–1963, 2021.

[202] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *PNAS*, vol. 114, pp. 3521–3526, 2017.