Sleep monitoring using ear-centered setups: Investigating the influence from electrode configurations

Kaare B. Mikkelsen*[1], Huy Phan[2,3] , Mike L. Rank[4], Martin C. Hemmsen[4], Maarten de Vos[5,6], Preben Kidmose[1]

***Abstract*— Modern sleep monitoring development is shifting towards the use of unobtrusive sensors combined with algorithms for automatic sleep scoring. Many different combinations of wet and dry electrodes, ear-centered, forehead-mounted or headband-inspired designs have been proposed, alongside an ever growing variety of machine learning algorithms for automatic sleep scoring.**

Objective: Among candidate positions, those in the facial area and around the ears have the benefit of being relatively hairless, and in our view deserve extra attention. In this paper, we seek to determine the limits to sleep monitoring quality within this spatial constraint.

Methods: We compare 13 different, realistic sensor setups derived from the same data set and analysed with the same pipeline.

Results: All setups which include both a lateral and an EOG derivation show similar, state-of-the-art performance, with average Cohen's kappa values of at least 0.80.

Conclusion: If large electrode distances are used, positioning is not critical for achieving accurate sleep scoring.

Significance: We argue that with the current competitive performance of automated staging approaches, there is a need for establishing an improved benchmark beyond current single human rater scoring.

***Index Terms*— EEG, ear-EEG, Deep Learning, Sleep scoring**

I. Introduction

During an 80 year lifespan, a human spends roughly 27 years asleep. As such, it should not be surprising that sleep has a large impact on virtually every major disease category, from cardiovascular disease over psychiatric disorders to cancer [1]. However, diagnosis of sleep disorders is still largely confined to dedicated sleep laboratories. Laboratory-based polysomnography (PSG) is the main method to gather insight in a patient's sleep, certainly when neurophysiological data is needed. Although sleep is an essential part of several disorders, such as neuropsychiatric disorders, the practical limitations for wide scale use of PSG's hamper the integration of sleep as a vital component in diagnostic and therapeutic trajectories of patients with these disorders. Moreover, the sleep laboratory is a very artificial environment, which has an influence on sleep itself. In order to better understand the impact of healthy and abnormal sleep-wake patterns on various disease conditions, there is an urgent need for sleep monitoring over prolonged periods of time outside traditional sleep clinics.

In the past decade, multiple studies have explored the use of digital wearable (e.g. actigraphy) and bed-side (e.g. radar-based) sensors to quantify various aspects of sleep, but failed to capture the neurophysiological signatures that underpin the quantification of sleep based on the AASM convention [2]. With the introduction of various wearable EEG sensors (ear-EEG, cEEGrid, sleep zeo, Dreem) which capture brain activity from unconventional places (e.g. on the forehead, in or around the ears or using a headband), personalized long-term sleep monitoring on the general population is within reach [3]–[10].

Visually reviewing the large amount of time series sleep data that could be recorded with this new generation of wearable EEG would be time-consuming and costly, in addition to requiring re-training of the human scorers for each new wearable (which would be highly inefficient [10]). Initial approaches to automatic sleep scoring were based on traditional, hand-crafted features, designed using domain knowledge of sleep experts. These features were fed into a suitable machine learning algorithm ([11], [12]). A modern upgrade to this approach is neural network models, in which also the feature extraction is handled by the algorithm. This lead to a variety of promising automated sleep analysis approaches. An important advantage of automated scoring approaches is the absence of intra-scorer variability [13]. Machine learning algorithms for sleep scoring are primarily developed and validated on large, publically available PSG data sets. As was shown recently [14], these same algorithm designs can also produce state of the art results on wearable sleep data. In this paper we also investigate whether such PSG data sets can be used to improve performance on wearable data through pre-training.

Due to the variety of available wearable sensors and the experimental nature of data collection with those devices, 'wearable' datasets are still an order of magnitude smaller compared to PSG data sets, and performance of automated staging approaches requires further investigation.

In this paper, we apply one of the leading analysis pipelines, the SeqSleepNet [15], to multiple different, realistic sensor

1: Department of Electrical and Computer Engineering, Aarhus University. 2: School of Electronic Engineering and Computer Science, Queen Mary University of London. 3: the Alan Turing Institute, London 4: T&W Engineering, Lynge, Denmark. 5: Faculty of Engineering Science, KU Leuven. 6: Faculty of Medicine, KU Leuven.

* mikkelsen.kaare@eng.au.dk

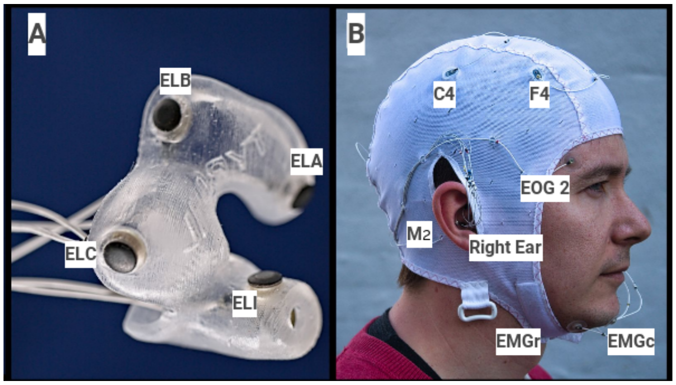

Fig. 1. The recording setup used in this study. A: example of the soft ear-EEG electrode holders, with embedded dry-contact electrodes, placed in each ear. B: cap-mounted PSG setup using 8 scalp EEG electrodes, 2 EOG electrodes and 3 EMG electrodes. See [3] for a detailed description.

configurations. 'Realistic' in this sense means that we only test montages where a limited number of electrode positions are used at a time, and only positions that are reasonably hidden and easy to access (out of the hair line and only on the sides of the head).

The SeqSleepNet pipeline was validated by an independent group on two different datasets, showing that the performance of the network outperformed the average human annotator [16]. In this paper, we show that this network, originally developed and trained for automatically staging PSG, can be directly applied to in-ear EEG data. In addition, we investigate the likely upper limits to mobile sleep scoring accuracy and the variations between different approaches.

II. Methods

A. Data

We used the 80 nights of sleep recordings (4 nights from 20 subjects each) which were presented in Mikkelsen et al 2019[3]. This data set consists of concurrent PSG (13 electrodes) and ear-EEG (6 electrodes in each ear) recordings. See Figure 1 for an example of the setup. Data collection was conducted in accordance with the Good Clinical Practice guide lines and the declaration of Helsinki. Monitoring was performed by the GCP unit at Aarhus University, and the protocol was accepted by the Danish Medicines Agency (ref. nr. 2017111085) and Central Denmark Region Committees on Biomedical Research Ethics (ref. nr. 1-10-72-413-17).

The recordings were sampled with a TMSi Mobita amplifier, with a sampling frequency of 500 Hz. The Mobita amplifier is a mobile EEG amplifier with 24 bit resolution in a 400 mV dynamic range (peak-to-peak), individually shielded inputs, less than 0.4 μV RMS noise in the 0.1–10 Hz band, and greater than 100 dB CMRR.

Rather than using the raw data, we work with the sleep recordings after artefact rejection, as described in Mikkelsen et al [3]. In this artefact rejection pipeline, artefacts are identified on an individual electrode basis, and are removed by changing the relevant sample values to 'NaN' (which enables discarding samples from individual channels). During preparation of the various derivations, NaN-values are ignored when EEG electrodes are averaged (as is the case with ear derivations). If there were any NaN's in a final derivation, the missing samples were linearly interpolated from the nearest non-missing values. For extended missing sections, the interpolated values decayed exponentially towards zero (with time constant 1 second).

The PSG recordings have been scored by two independent and experienced sleep technicians ('scorer 1' and 'scorer 2'), according to the AASM guidelines [2]. We have decided to treat scorer 1 as the ground truth, to which the automatic sleep classifiers will be compared (and trained on). In contrast, scorer 2 is an independent source of labels, which will be used in studying the possible causes of classifier errors.

B. Choice of electrode configurations and epochs

Figure 2 shows all electrode derivations under consideration in this study. As can be seen, we have chosen to rely more on the left than right side of the head. This was done both to reduce the number of derivations at play, and because previous work had shown the left ear electrodes to be slightly more reliable than the right ear electrodes [3]. We will elaborate more on this in the 'Results' section. In designing the 'Scalp' and 'EMG' derivations, we decided to make them as reliable as possible, by combining multiple derivations in one. This was done because we are primarily interested in the performance of a mobile sleep monitoring setup, and we do not consider chin EMG or scalp EEG electrodes to be prime candidates for user friendly mobile setups. Therefore, the primary concern for these data channels is that they are responsible for as little data rejection as possible.

Epoch rejection: To make the comparison of different setups (meaning different combinations of derivations) as unambiguous as possible, **we only use epochs for which all derivations are well defined**. In this regard, a derivation is considered 'ill defined' if all samples in that epoch for that derivation have been rejected (replaced with 'NaN' values). In cases when a derivation is constructed by averaging a set of channels, any 'NaN'-values of an individual channel are ignored.

Using these statistics, we evaluate whether any derivations should be excluded from the analysis. In this regard, the important metric is not the individual reliability of the derivation, but rather to which degree the derivation is well-defined at the same time as others. If it is not, it will be directly responsible for reducing the number of viable epochs. As is shown later, we end up removing the 'right ear' derivation.

C. Derivation distances

In the course of our analysis, we compare classifier performance to the total derivation distance of the electrode setup used. This means that we have measured representative values for the distances between the electrodes of the different derivations under investigation. When multiple derivations are used, we have simply summed the distances of each.

For the in-ear derivations, we have measured the distances between bottom of ear canal and middle of concha, using the ear pieces of the 20 participants in the study, and used the average of these numbers. For all other derivations, we have

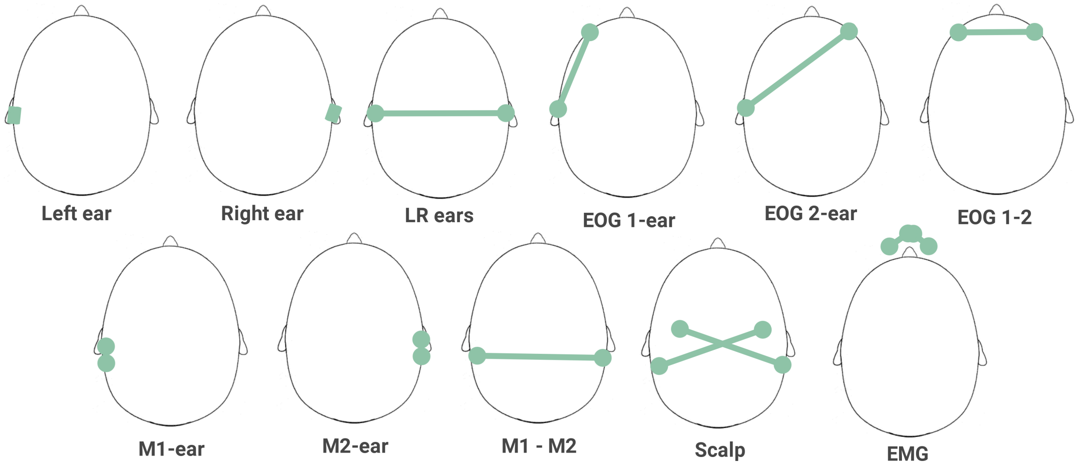

Fig. 2. Overview of different derivations used. 'Left ear' and 'right ear' uses the average over the three innermost ear electrodes versus the average of the three outermost electrodes, in each ear. 'LR ears' uses the average of all electrodes in each ear. 'M1-ear', 'EOG 1-ear' and 'EOG 2-ear' references a single electrode to the average of all left ear electrodes. For 'Scalp', both C3-M2 and C4-M1 are calculated; for each recording we used the derivation with the least rejected or lost samples. For 'Chin EMG', all three derivations between all three EMG electrodes (l, r, c) are calculated. If l-r has a missing sample, r-c is used instead. If r-c is missing as well, l-c is used.

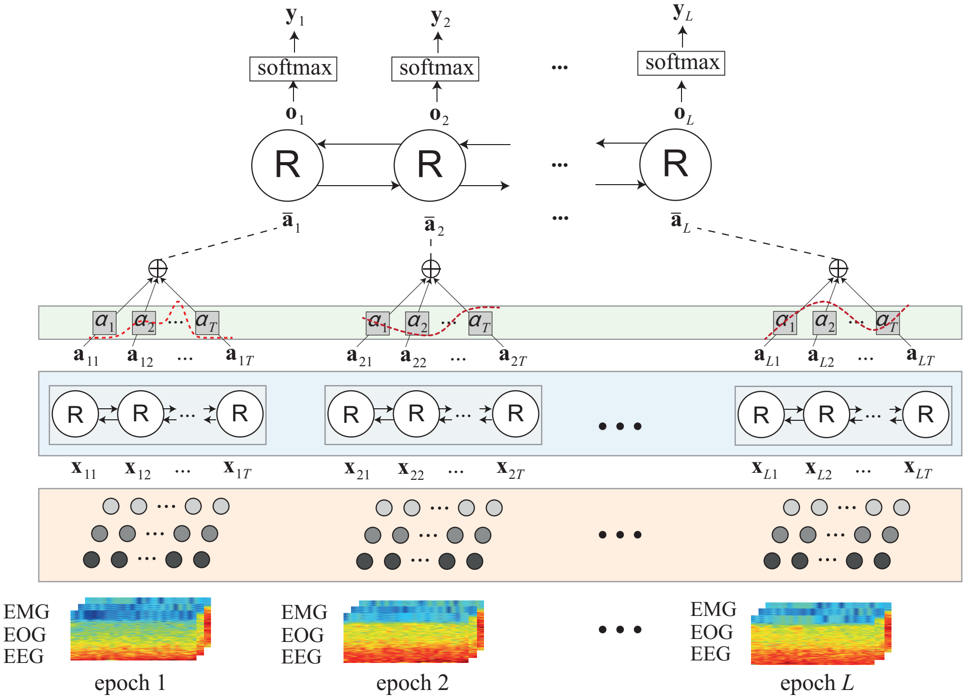
$\mathbf{y}_1$
$\mathbf{y}_2$
$\mathbf{y}_L$
softmax
softmax
softmax
$\mathbf{o}_1$
$\mathbf{o}_2$
$\mathbf{o}_L$
R
$\bar{\mathbf{a}}_1$
$\bar{\mathbf{a}}_2$
$\bar{\mathbf{a}}_L$
α_1 α_2 ... α_T
$\mathbf{a}_{11}$ $\mathbf{a}_{12}$ ... $\mathbf{a}_{1T}$
$\mathbf{a}_{21}$ $\mathbf{a}_{22}$ ... $\mathbf{a}_{2T}$
$\mathbf{a}_{L1}$ $\mathbf{a}_{L2}$ ... $\mathbf{a}_{LT}$
$\mathbf{x}_{11}$ $\mathbf{x}_{12}$ ... $\mathbf{x}_{1T}$
$\mathbf{x}_{21}$ $\mathbf{x}_{22}$ ... $\mathbf{x}_{2T}$
$\mathbf{x}_{L1}$ $\mathbf{x}_{L2}$ ... $\mathbf{x}_{LT}$
EMG
EOG
EEG
epoch 1
epoch 2
epoch L

Fig. 3. Illustration of SeqSleepNet. From the bottom: for L epochs, up to three channels are passed as spectrograms into a filter bank layer. After the filter bank, a recurrent layer processes each epoch (consisting of T steps each). The output form this recurrent layer is passed to an attention layer, which is followed by a second recurring layer, traversing the L epochs. The outputs from this recurrent layer are passed to softmax functions, yielding L epoch labels for L epochs. The figure is adapted from [15], and in this paper, $\boldsymbol{L = 20}$.

used electrode positions on a standard polystyrene model head used in teaching, measured in direct lines. The distances are reported in Table I.

D. The SeqSleepNet classifier

In this study we used SeqSleepNet [15], illustrated in Fig. 3, as the base classifier. SeqSleepNet works by analysing a sequence of L consecutive epochs and classifying them at once into a sequence of L sleep stage labels (i.e., sequence-to-sequence). We set $L = 20$ in this study as recommended in [15]. The data input to the network can be single- or multiple-channel log-scale spectrograms. The data of each channel was normalized to have zero mean and unit variance for each frequency bin using the normalization parameters computed from the training data.

The i-th epoch, $1 \leq i \leq L$, in the input sequence, was encoded into a feature vector $\bar{\mathbf{a}}_i$ via the epoch encoder. The epoch encoder is composed of (1) filter-bank layers, one for

TABLE I

OVERVIEW OF DERIVATION DISTANCES. SEE FIGURE 2 FOR A DESCRIPTION OF THE DERIVATIONS

Derivation	Distance [cm]
Left ear, Right ear	1.8
M1-ear, M2-ear	4.5
LR ears	15.0
EOG1-ear	14
EOG2-ear	16
EMGl-r	3.5
M1-M2	15.5
scalp	15.5

TABLE II

OVERVIEW OF CHANNEL COMBINATIONS USED. '+' MEANS USING MUTIPLE, SEPARATE DATA CHANNELS.

Single, double or triple channel input:		
Single	Double	Tripple
'Left ear'	'M1-ear'+'EOG1-ear'	
'M1-ear'	'M1-ear'+'EOG2-ear'	
'EOG 1-2'	'LR ears'+'EOG1-ear'	'scalp'+'EOG 1-2'+'EMG'
'LR ears'	'LR ears'+'EOG2-ear'	'LR ears'+'M1-ear'+'M2-ear'
'M1-M2'	'LR ears'+'EOG 1-2'	
'scalp'		

each input channel, (2) a bidirectional recurrent layer realized by a long short-term memory (LSTM) cell, and (3) an attention layer. The spectrogram channels first have their frequency dimension smoothed and reduced via the filter-bank layers. The filtered spectrograms are then stacked along the frequency direction and presented to the LSTM, which converts them into a sequence of output vectors. The output vectors in this sequence are eventually combined, using weights learned by the attention layer to form the feature vector $\bar{\mathbf{a}}_i$.

Going through the epoch encoder, the input sequence was transformed into a sequence of feature vectors. An LSTM-based bidirectional recurrent layer was then employed for inter-epoch sequential modelling, converting the sequence of feature vectors into a sequence of output vectors. These output vectors were finally presented to a fully-connected layer, followed by a softmax layer, for classification, producing a sequence of labels, each label corresponding to an epoch in the input sequence. The network was trained end-to-end to minimize the cross-entropy loss averaged over the sequence. See [15] for more details.

E. Classifier training and transfer learning

Training: To test a wide selection of different, relevant electrode combinations, we used different subsets of the electrode derivations as inputs to the network, these are listed in Table II. Here, '+' means that multiple derivations are given as separate inputs. The SeqSleepNet was configured similarly to the original implementation [15], meaning that: each 30-second epoch was transformed into a spectrogram of 129 frequency bins and 29 time bins. We used 20 spectrograms, the filterbanks each had 32 filters (with different banks for different input channels), and the banks had low and high frequencies of 0 and 50 Hz. We used a dropout-rate of 0.25, attention size and hidden sizes of both GRU layers were 64, the learning rate was 1e-4 and the L2-regularization lambda was 1e-3. Batch-size was 32. For a public code repository of seqsleepnet, we suggest https://github.com/pquochuy/SeqSleepNet.

Transfer learning: As an alternative to training directly on the reduced electrode set, we also studied the effect of transfer learning [17]. To this end, we pretrained SeqSleepNet with the Montreal Archive of Sleep Studies (MASS) database, which consists of 200 subjects [18]. For this test, only a single-input version of the network was prepared, using the C4-A1 derivation. The pretrained networks were then used as the starting points and further trained (i.e., the entire network were finetuned) with our data.

Performance evaluation: In the remainder of this paper, we shall refer to a SeqSleepNet trained for a specific set of inputs as a 'classifier'. When discussing both manual sleep scorers and automatic classifiers, we shall refer to all of them as 'scorers'.

Each classifier was trained and tested in a leave-one-subject-out fashion. Of the remaining 19 subjects, 15 were used as training set, and 4 were as validation set for early stopping and to avoid overfitting to the training set. For each subject, all available recordings were used. As each recording is on average 843 epochs long after epoch rejection, and three subjects were only represented by 3 recordings, the average number of epochs used in each training fold was 48679, and the average test set was 3245 epochs.

To quantify classifier performance, we calculate Cohen's kappa [19] between the automatic and manual scoring (from scorer 1) on the test epochs. Since we are performing cross validation, all epochs serve as test epochs at some point, making it possible to calculate kappa values for the full data set (after epoch rejection). This means that all kappa statistics are based on 64905 epochs.

Cohen's kappa was chosen due to its historical use in automatic sleep scoring development, ease of interpretation, and its property of correcting for chance agreement.

III. RESULTS

Figure 4 shows how the set of accepted epochs depends on the chosen set of derivations. On the left is shown rejection statistics for individual derivations. We see that the epoch-wise rejection rate is quite stable across derivations, varying between 3.2% and 6.0%. For comparison, scorer 1 marked 3.5% of the epochs as unclassified. However, more important than the single derivation statistics is the impact on epoch rejection when multiple derivations are considered. On the right, statistics are shown for when all but one derivation are used. Again, we see that excluding a single derivation mostly does not change the overall rejection rate. However, we note that removing the 'right ear' derivation reduces the number of recordings that are completely rejected (from 4 to 3). Because of this, we decided to exclude the 'right ear' derivation from the rest of the analysis, and use the 83% of epochs which are accepted in all other derivations (including being scored by scorer 1). This results in excluding 2 recordings (from two different subjects).

It is here important to point out that the main driver of the left vs. right ear difference is random hardware issues. During

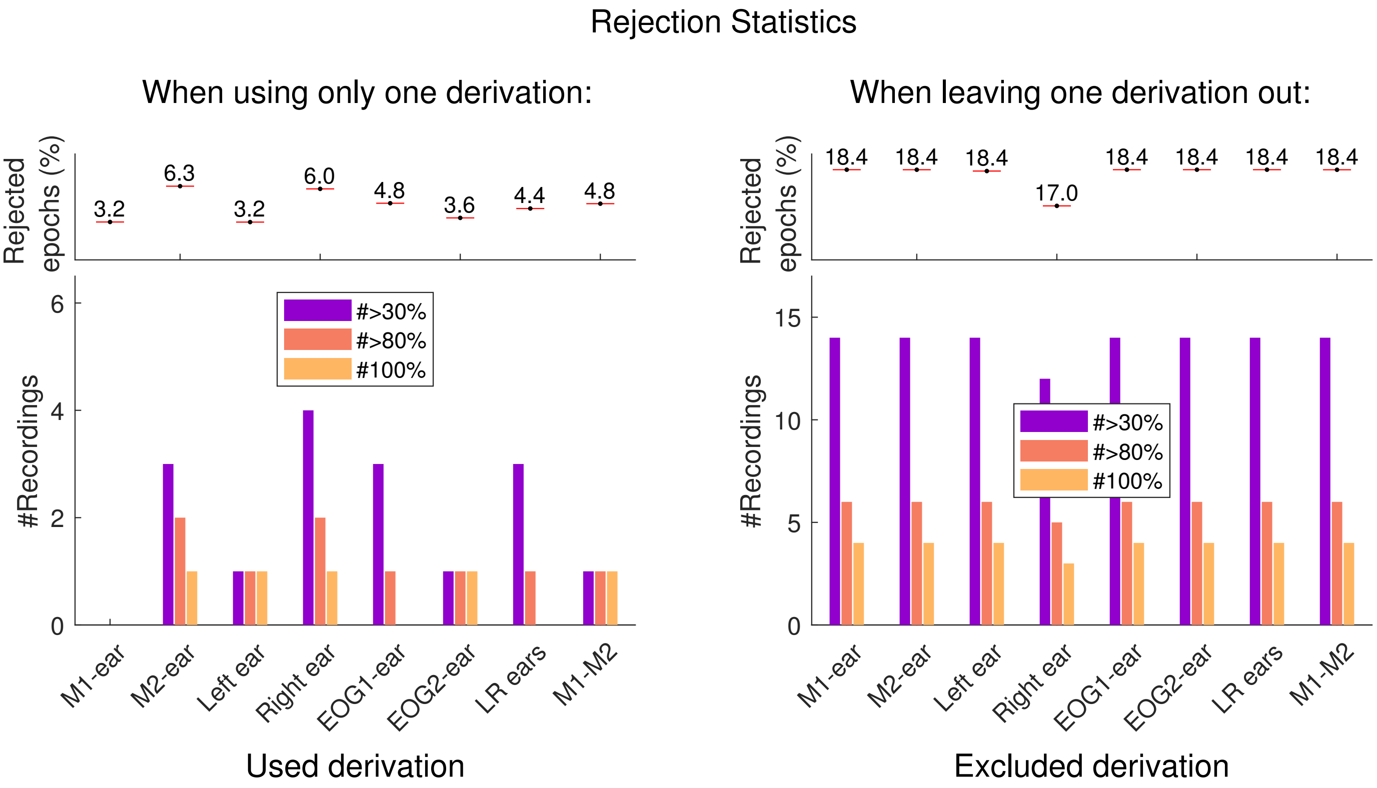

Fig. 4. The relationship between data channels used and amount of accepted epochs. Left, top: percentage of epochs rejected based on a single data channel. Left, bottom: number of recordings for which at least 30, 80 and 100 percent of epochs were rejected based on a single data channel. Right, top: percentage of epochs rejected when all data channels except one were included (the excluded channel is shown on the bottom x-axis). Right, bottom: number of recordings for which at least 30, 80, and 100 percent of epochs were rejected when all data channels except one were included.

a subset of the recordings, some electrodes, in particular the three 'ear canal' electrodes in the right ear, were plagued by instability issues which were caused by faulty shielding. Even though the issue was fixed, the slight imbalance caused by it persists in the data set.

Additionally, please note that EMG, EOG and Scalp derivations have been excluded from the comparison in Figure 4. This is because these derivations are all necessary to perform our analysis (constituting a three-channel PSG classifier), and thus their inclusion is obligatory.

Figure 5 shows boxplots for distributions of Cohen's kappa between the classifier output and the manual labels assigned by scorer 1. It is interesting to note how any classifier which combines both lateral and EOG information reaches kappa values of about 0.8 or above (in particular how well 'EOG 1-2' performs). Also, we see that the 'scalp'+'EOG 1-2'+'EMG'-classifier actually reproduces scorer 1 better than scorer 2 does. This indicates that SeqSleepNet manages to incorporate the special quirks of scorer 1, and that it is probably unwarranted to attempt further improvements in PSG-based scoring (at least when training against a single scorer). In the bottom part of the figure is shown the summed derivation distances for each automatic classifier. We note a quite nice relationship between classifier performance and derivation distance, in particular when focusing on the 'mobile' setups without scalp electrodes.

It is worth noting that we have found no indication that the specific choice of epochs used here (as described above) is particularly easy to score, which would introduce a bias towards artificially high kappa values. We have tested the random forest based classifier presented in Mikkelsen et al. 2019 [3], on the same, reduced epoch set, (used for both testing and training), however it only attains an average kappa coefficient of 0.72 (compared to the 0.73 which was achieved using a larger set of epochs).

In the case of transfer learning, we tested the effect on the 'LR ears', 'LR ears'+'EOG1-ear', 'LR ears'+ 'EOG2-ear' and 'M1-M2'. We found average increases in Cohen's kappa of 0.016, 0.009, 0.005 and 0.029, relative to the scratch-trained classifiers. When performing a two-tailed permutation test of whether these changes are significantly different from 0, we find p-values of 0.0050, 0.0208, 0.0612 and 0.0002, meaning that all but the smallest increase (that for 'LR ears'+'EOG2-ear') are statistically significant.

Figure 6 shows a visual comparison between all scorers, both manual and automatic. For each scorer, all kappa values are calculated relative to all other scorers (by bundling all recordings into one, and calculating one total kappa value), and the two highest values are plotted as edges on the graph. This means that while some nodes (each representing a scorer) have more than two connected edges, all nodes have at least two. The edges are coloured depending on the kappa value, and the nodes are coloured depending on the kappa value between the scorer and scorer 1.

An interesting observation can be made from Figure 6: even

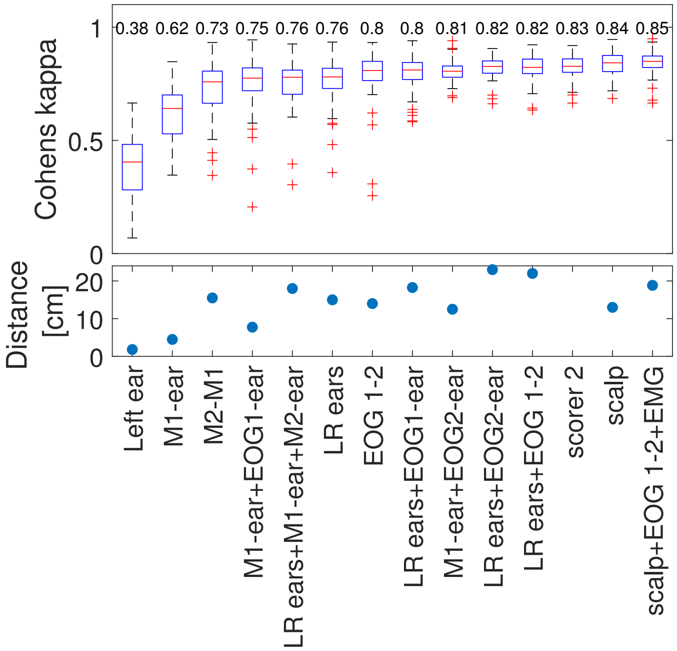

Fig. 5. Top:Distributions of all kappas for all scratch-trained classifiers, relative to the labels from scorer 1. Numbers at the top are averages. Bottom: Total derivation distances for each setup ('scorer 2' excluded).

though 'scorer 1' is the target that all automatic classifiers are aiming at, they do in fact agree more with each other than with scorer 1. This happens even for classifiers that do not have any input derivations in common (e.g. 'M1-ear' + 'EOG2-ear' and 'LR ears' + 'EOG 1-2' may share electrodes, but not derivations). In particular, we note that even though 'scalp'+'EOG 1-2'+'EMG' attains the highest kappa relative to scorer 1 of any scoring method, it still attains even higher kappa values with other automatic classifiers. We can think of two plausible causes of this: (1) the manual scorer likely also makes some mistakes, meaning that there is an upper limit to how well an entirely rules-based sleep classifier can predict manual scoring. (2) it is possible that the manual sleep scorer uses information not available to the classifiers - either because the manual scorer considers more than the last 10 minutes of recording when scoring a given epoch, or because they consider other aspects of the recording, such as time of night, total duration etc., which are not revealed to the automatic classifier.

When we further analyse the discrepancies between manual and automatic scoring, we find, not surprisingly, that most errors happen close to state transitions. This is shown in Figure 7 where we see that almost 60% of discrepancies between manual and automatic scoring happens immediately before or after a stage transition (as judged by scorer 1). Including three additional epochs to either side of the transition brings the total up to around 80%. For comparison, only 20% of epochs are right next to a transition, and 45% percent are within 4 epochs of a transition.

Given the high agreement between many of the automatic classifiers, we decided to specifically study the level of consensus between some of the most well-performing classifiers. We chose the following 5: 'LR ears'+'EOG2-ear','EOG 1-2','M1-ear'+'EOG2-ear','scalp','scalp'+'EOG 1-2'+'EMG'. When comparing each of the 5 classifier outputs to their own majority vote, we overwhelmingly find that the 6 classifiers mutually agree. Figure 8 shows the average number of votes for the majority (maximum 5) for different sleep stages (as judged by the majority). We see that in 80% of cases there is complete consensus, except for stage N1, which is also considered the least well-defined stage.

IV. Discussion and Conclusion

By applying an advanced sleep scoring algorithm to high quality wearable EEG data, we achieve a high scoring performance relative to previous studies. Relevant comparisons in this regard are Stepnowski et al 2013 [12] which achieved a kappa value of 0.61, Lewendowski et al 2017 [4] which achieved 0.63, Mikkelsen et al 2019 [3] which achieved 0.73, and Arnal et al 2020 [7] which reached a kappa of 0.75. Obviously these different studies used slightly different electrode setups, but except for Mikkelsen et al, they are comparable to either the 'EOG 1-2' setup or 'LR ears'+'EOG 1-2' (Mikkelsen et al. essentially used the 'LR ears' derivation). This means that we are seeing kappa improvements of at least 0.05 relative to previously published work. This is central for the realization of light weight sleep monitoring, and our results here show that this can be reality. Additionally, importantly, we find that a broad selection of electrode placements, all having in common that they have both lateral and EOG components, achieve very similar performances, with a trend towards longer electrode distances leading to better classifiers. This is likely because sleep, and sleep stage-specific oscillations, are global phenomena [20], and as long as the used derivations have a sufficiently large region of sensitivity, it is possible to distinguish between sleep stages. A contributing factor is presumably also that some noise sources (amplifier and thermal electrode noise) are distance independent, meaning a worse signal-to-noise ratio for short electrode distances, and a resulting poor sleep scoring. All of this means that electrode placements should be chosen based on unobtrusiveness, reliability and comfort, and if the recording setup is otherwise sound, we predict that a very large number of different sensor combinations can make a viable sleep monitor. This is interesting from a design perspective of mobile sleep monitors, particularly ear-EEG setups: given different constraints posed by different patient groups and different medical contexts, we are relatively free to choose a number of sensors and placements, and still expect a useable sleep monitor (given adequate training data etc.). We believe that for most home-recording scenarios, a kappa value of 0.76 is sufficient. This means that for 'young healthy' subjects, such as the ones used in this study, simple setups like either a single-sided setup (as in 'M1-ear'+'EOG1-ear') or a horizontal-only setup (like 'LR ears') will be adequate. However, it is clear that by combining both EOG and horizontal derivations, a distinct improvement is possible. We anticipate that for more challenging user groups, the expected decrease in kappa for the 'simple' setups can be counter-acted by upgrading to, for instance, 'LR ears'+'EOG2-ear'. This approach should be

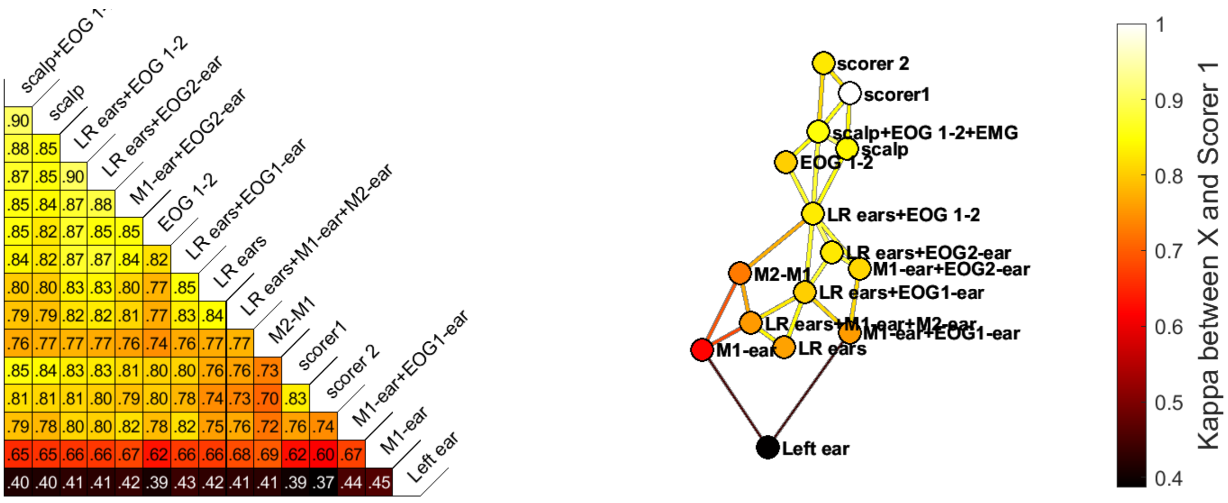

Fig. 6. Left: Matrix showing all pairwise Cohens kappa coefficients, ordered to maximise nearest neighbour values. Right: Graph ordering scorers based pairwise kappas. Each node represents a scorer, and the edge weights represent the kappa value between the two scorer outputs. The node color shows the kappa value between classifier output and scorer 1 labels. For clarity, only the two strongest edges for each node have been included.

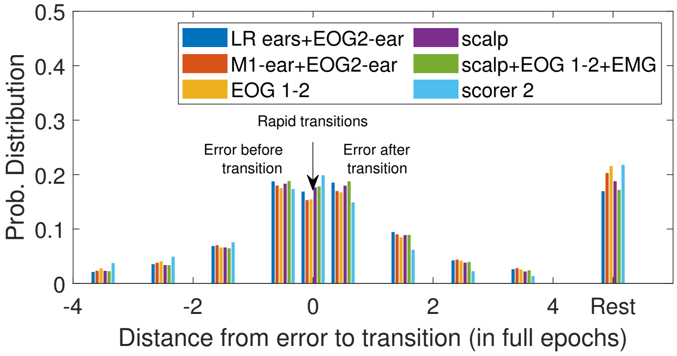

Fig. 7. Distribution of automatic classifier errors as a function of distance to nearest stage transition (as defined by the scoring from manual scorer 1). In the figure, 'Rapid transitions' refers to the scenario where scorer 1 only spends a single epoch in the given stage, meaning that possible error is both immediately before and after the transition.

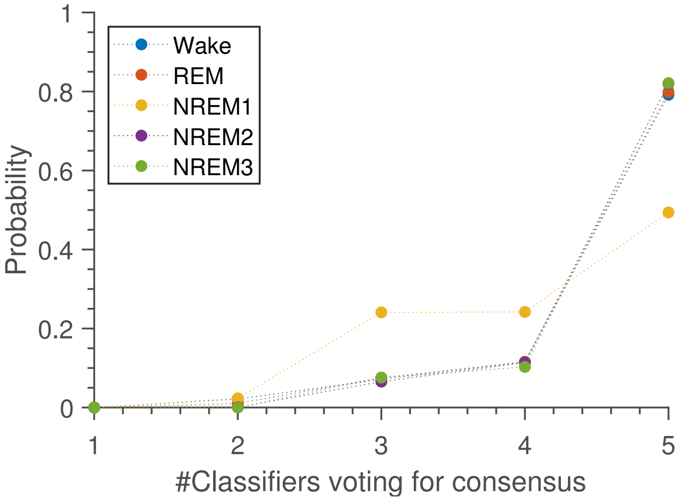

Fig. 8. Among the 6 best automatic classifiers, we count how many of them agree with their own consensus label, for each epoch. We see that in the overwhelming majority of epochs, all 6 classifiers agree. The one clear outlier is NREM1, which it much less well defined than the others. Please note that only integer numbers of votes are considered (1-5) on the x-axis, and the connecting lines are only included as visual guides.

tested in future studies. We note that these considerations are similar to what is discussed in the recent study by da Silva et al [21] focusing only on single-sided around-the-ear EEG, where an ear-EEG setup very similar to the 'M1-ear'+'EOG1-ear' combination is specifically discussed.

In particular, we found that a PSG-based automatic scorer, which performed very well in reproducing scorer 1, still had a higher Kappa value with at least two other automatic classifiers. This indicates that the high internal consistency among the automatic classifiers is not entirely due to limited sleep information in the non-PSG derivations, but is likely also related to the human peculiarities in the scoring by scorer 1. Apparently, the automatic classifiers all manage to define certain special cases more consistently than scorer 1, leading to the 'scalp+EOG 1-2+EMG' classifier attaining both the highest kappa value with scorer 1, while at the same time having higher agreement with other automatic classifiers.

Based on this observation, it would be very interesting to compare the output of the classifiers presented here with output from consensus-trained PSG-based classifiers such as the one presented in Stephansen et al 2018 [22], which the authors believe could be more consistent than the gold standard manual sleep scoring.

In future work, it will be interesting to see how these results change when a more challenging cohort is used - it is possible that as sleep and its associated biomarkers change with age or infirmity, the optimal electrode locations will change accordingly.

On the topic of future directions, we feel that this work highlights the need for a change in focus regarding how machine learning is used to improve clinical sleep analysis. Given the apparent high reliability of automatic sleep scoring shown in this paper and others, we believe that the goal of reproducing manual scoring for 'regular sleep' has been largely reached. Rather than marking any kind of end to the project of updating clinical sleep analysis, we believe this points to the beginning of a new phase. To anyone following this field, it should be clear that cost-effective, long term sleep monitoring is becoming a reality. The question now is, how can automatic scoring be transformed into a trusted, clinical tool (as was recently suggested by the American Academy of Sleep Science [23]), and how can we use this tool to actually update the framework within which sleep is analyzed? For instance, the work presented in this paper would likely have benefitted from a more finegrained definition of sleep stages, such as the 'hypnodensities' that some researchers have been advocating [24]. This concept also includes higher temporal resolution, which would presumably have changed the results discussed in Figure 7. We believe that 'hypnodensity' is an example of how the existence of accurate, automatic sleep scoring, suitable for long-term monitoring, can motivate and support development of new approaches. We hope that much more of such developments are on the way.

V. Competing Interests

Authors MLR and MCH are employed by T & W Engineering, which develops equipment for long term EEG monitoring.

VI. Acknowledgements

This work was sponsored by the Innovation Fund Denmark, grant 7050-00007 and supported by funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

VII. Author Contribution

KBM performed the recordings and the analysis and wrote the manuscript. HP and MdV designed the sleep scoring algorithm, PK designed and build the recording equipment, KBM, MLR, MCH and PK designed the experiment, all authors participated in presenting and formulating the results.

VIII. Data and code availability

The authors are happy to share both the aggregate statistics presented in this manuscript, as well as the entire code base.

References

[1] I. Perez-Pozuelo *et al.*, "The future of sleep health: A data-driven revolution in sleep science and medicine," en, *npj Digital Medicine*, vol. 3, no. 1, pp. 1–15, Mar. 2020, ISSN: 2398-6352.

[2] R. B. Berry *et al.*, "AASM scoring manual updates for 2017 (version 2.4)," *Journal of Clinical Sleep Medicine*, vol. 13, no. 5, pp. 665–666, 2017.

[3] K. B. Mikkelsen *et al.*, "Accurate whole-night sleep monitoring with dry-contact ear-EEG," en, *Scientific Reports*, vol. 9, no. 1, pp. 1–12, Nov. 2019, ISSN: 2045-2322.

[4] D. J. Levendowski *et al.*, "The Accuracy, Night-to-Night Variability, and Stability of Frontopolar Sleep Electroencephalography Biomarkers," eng, *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine*, vol. 13, no. 6, pp. 791–803, Jun. 2017, ISSN: 1550-9397.

[5] B. P. Lucey *et al.*, "Comparison of a single-channel EEG sleep study to polysomnography.," *Journal of sleep research*, vol. 25, no. 6, pp. 625–635, Dec. 2016, ISSN: 1365-2869.

[6] D. Looney *et al.*, "A Wearable In-Ear Encephalography Sensor for Monitoring Sleep: Preliminary Observations from Nap Studies," *Annals of the American Thoracic Society*, Sep. 2016, ISSN: 2329-6933.

[7] P. J. Arnal *et al.*, "The Dreem Headband compared to polysomnography for electroencephalographic signal acquisition and sleep staging," *Sleep*, vol. 43, no. zsaa097, Nov. 2020, ISSN: 0161-8105.

[8] K. B. Mikkelsen *et al.*, "Automatic sleep staging using ear-EEG," *Biomedical Engineering Online*, vol. 16, no. 111, Sep. 2017.

[9] D. Popovic, M. Khoo, and P. Westbrook, "Automatic scoring of sleep stages and cortical arousals using two electrodes on the forehead: Validation in healthy adults.," *Journal of sleep research*, vol. 23, no. 2, pp. 211–221, Apr. 2014, ISSN: 1365-2869.

[10] K. B. Mikkelsen *et al.*, "Machine-learning-derived sleep-wake staging from around-the-ear electroencephalogram outperforms manual scoring and actigraphy," en, *Journal of Sleep Research*, vol. 0, no. 0, e12786, Nov. 2018, ISSN: 1365-2869.

[11] 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. 1186–1195, Dec. 2012, ISSN: 1879-0534.

[12] C. Stepnowsky *et al.*, "Scoring accuracy of automated sleep staging from a bipolar electroocular recording compared to manual scoring by multiple raters," *Sleep Medicine*, vol. 14, no. 11, pp. 1199–1207, Nov. 2013, ISSN: 13899457.

[13] C. Berthomier *et al.*, "Exploring scoring methods for research studies: Accuracy and variability of visual and automated sleep scoring," eng, *Journal of Sleep Research*, vol. 29, no. 5, e12994, Oct. 2020, ISSN: 1365-2869.

[14] H. Phan *et al.*, "Deep Transfer Learning for Single-Channel Automatic Sleep Staging with Channel Mismatch," in *2019 27th European Signal Processing Conference (EUSIPCO)*, Sep. 2019, pp. 1–5.

[15] H. Phan *et al.*, "SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 3, pp. 400–410, Mar. 2019, ISSN: 1558-0210.

[16] A. Guillot *et al.*, "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, Sep. 2020, ISSN: 1558-0210.

[17] H. Phan *et al.*, "Towards More Accurate Automatic Sleep Staging via Deep Transfer Learning," *IEEE Transactions on Biomedical Engineering*, pp. 1–1, 2020, ISSN: 1558-2531.

[18] C. O'Reilly *et al.*, "Montreal Archive of Sleep Studies: An open-access resource for instrument benchmarking and exploratory research," en, *Journal of Sleep Research*, vol. 23, no. 6, pp. 628–635, 2014, ISSN: 1365-2869.

[19] J. Cohen, "A Coefficient of Agreement for Nominal Scales," en, *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, Apr. 1960, ISSN: 0013-1644.

[20] A. Brancaccio *et al.*, "Cortical source localization of sleep-stage specific oscillatory activity," en, *Scientific Reports*, vol. 10, no. 1, p. 6976, Apr. 2020, ISSN: 2045-2322.

[21] C. F. da Silva Souto *et al.*, "Flex-Printed Ear-EEG Sensors for Adequate Sleep Staging at Home," *Frontiers in Digital Health*, vol. 3, p. 66, 2021, ISSN: 2673-253X.

[22] J. B. Stephansen *et al.*, "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy," en, *Nature Communications*, vol. 9, no. 1, p. 5229, Dec. 2018, ISSN: 2041-1723.

[23] Goldstein Cathy A. *et al.*, "Artificial intelligence in sleep medicine: An American Academy of Sleep Medicine position statement," *Journal of Clinical Sleep Medicine*, vol. 16, no. 4, pp. 605–607,

[24] J. B. Stephansen *et al.*, "The use of neural networks in the analysis of sleep stages and the diagnosis of narcolepsy," *arXiv:1710.02094 [cs]*, Oct. 2017.