

PAPER • OPEN ACCESS

Single-channel EOG sleep staging on a heterogeneous cohort of subjects with sleep disorders

To cite this article: Hans van Gorp *et al* 2024 *Physiol. Meas.* **45** 055007

View the [article online](#) for updates and enhancements.

You may also like

- [BTCRSleep: a boundary temporal context refinement-based fully convolutional network for sleep staging with single-channel EEG](#)
Caihong Zhao, Jinbao Li and Yahong Guo
- [Interbeat interval-based sleep staging: work in progress toward real-time implementation](#)
Gary Garcia-Molina and Jiewei Jiang
- [Automatic sleep staging of EEG signals: recent development, challenges, and future directions](#)
Huy Phan and Kaare Mikkelsen

Breath Biopsy Conference

5th & 6th November
Online

Join the conference to explore the **latest challenges** and advances in **breath research**, you could even **present your latest work!**

Register now for free!



Main talks

Early career sessions

Posters

**OPEN ACCESS****PAPER**

Single-channel EOG sleep staging on a heterogeneous cohort of subjects with sleep disorders

RECEIVED
19 December 2023

REVISED
12 April 2024

ACCEPTED FOR PUBLICATION
23 April 2024

PUBLISHED
15 May 2024

Hans van Gorp^{1,2}, Merel M van Gilst^{1,3}, Sebastiaan Overeem^{1,3}, Sylvie Dujardin³, Angelique Pijpers³, Bregje van Wetten³, Pedro Fonseca^{1,2} and Ruud J G van Sloun¹

¹ Department of Electrical Engineering, Eindhoven University of Technology, The Netherlands

² Philips Sleep and Respiratory Care, Eindhoven, The Netherlands

³ Sleep Medicine Centre Kempenhaeghe, Heeze, The Netherlands

E-mail: h.v.gorp@tue.nl

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.



Keywords: electrooculography, automatic sleep staging, sleep disorders, EOG

Abstract

Objective. Sleep staging based on full polysomnography is the gold standard in the diagnosis of many sleep disorders. It is however costly, complex, and obtrusive due to the use of multiple electrodes. Automatic sleep staging based on single-channel electro-oculography (EOG) is a promising alternative, requiring fewer electrodes which could be self-applied below the hairline. EOG sleep staging algorithms are however yet to be validated in clinical populations with sleep disorders.

Approach. We utilized the SOMNIA dataset, comprising 774 recordings from subjects with various sleep disorders, including insomnia, sleep-disordered breathing, hypersomnolence, circadian rhythm disorders, parasomnias, and movement disorders. The recordings were divided into train (574), validation (100), and test (100) groups. We trained a neural network that integrated transformers within a U-Net backbone. This design facilitated learning of arbitrary-distance temporal relationships within and between the EOG and hypnogram. **Main results.** For 5-class sleep staging, we achieved median accuracies of 85.0% and 85.2% and Cohen's kappas of 0.781 and 0.796 for left and right EOG, respectively. The performance using the right EOG was significantly better than using the left EOG, possibly because in the recommended AASM setup, this electrode is located closer to the scalp. The proposed model is robust to the presence of a variety of sleep disorders, displaying no significant difference in performance for subjects with a certain sleep disorder compared to those without. **Significance.** The results show that accurate sleep staging using single-channel EOG can be done reliably for subjects with a variety of sleep disorders.

1. Introduction

There is a large demand for accurate, inexpensive, and reliable sleep staging methods, as it serves as an essential element in the diagnosis of many prevalent sleep disorders. Sleep staging, as defined by the American Academy of Sleep Medicine manual (Troester *et al* 2023), is the process of classifying 30 s segments of sleep, known as epochs, as belonging to one of five distinct sleep stages: wake (W), rapid eye movement (REM), or non-REM (NREM) stage 1–3. The resulting visualization of the sequence of sleep stages is called a hypnogram. Following the AASM manual, gold-standard sleep staging is carried out by a certified technician after visual analysis of at least the following signals: three scalp electroencephalography (EEG) electrodes, two electro-oculography (EOG) electrodes, and two chin electromyography (EMG) electrodes. Measuring all these signals is costly and causes subject discomfort. For example, the EEG electrodes have to be placed on the scalp above the hairline. Moreover, human inter-rater agreement is limited in this task, averaging around 82.6% (Rosenberg and Van Hout 2013). The need for the full EEG/EOG/EMG montage in the context of sleep staging has recently been called into question (Lambert and Peter-Derex 2023), as only a subset of the EEG/EOG/EMG signals or even surrogate signals might be sufficient for this task.

Automatic sleep stage scoring has been widely researched as an alternative. Unlike human technicians, these automatic scoring algorithms are much more flexible in terms of their input signals and do not require the full EEG/EOG/EMG montage. For example, automatic staging using only EEG electrodes already yields performance on par with the human inter-rater agreement (Phan and Mikkelsen 2022). Alternatively, surrogate measurements can also be leveraged, such as instantaneous heart rate, body movement, respiration, and many more (Bakker *et al* 2021, Imtiaz 2021, Fonseca *et al* 2023, Zhai *et al* 2023). However, these surrogate trackers do not yet reach the same performance as the human inter-rater agreement. Moreover, they typically perform only 4-stage classification, merging the N1 and N2 into a combined N1/N2 ‘light sleep’ stage.

Automatic scoring methods can not only output the ‘hard’ hypnogram but also a hypnodensity graph (Stephansen *et al* 2018). The hypnodensity graph is a visual representation of the probability assigned to each sleep stage by an automatic scoring algorithm, instead of only the most likely sleep stage used to create the ‘hard’ hypnogram. One can interpret the hypnodensity graph as the sleep stage uncertainty or ambiguity of the method (van Gorp *et al* 2022). The hypnodensity concept has gained much attention recently, as it has been shown to match well with the label distribution of a human panel and potentially carries more clinically relevant information than the hypnogram (Stephansen *et al* 2018, Bakker *et al* 2022, Anderer *et al* 2023, Huijben *et al* 2023).

As a middle ground between EEG and surrogate sleep staging, single-channel EOG staging has been proposed in the literature. EOG is advantageous when compared to EEG as it is measured below the hairline, can be self-applied (Virkkala *et al* 2008), and can even be measured with dry electrodes embedded in a sleep mask (Liang *et al* 2015, Hsieh *et al* 2021). At the same time, some amount of desirable EEG interference gets coupled into the EOG signal. Automatic sleep staging algorithms can pick up on and exploit this EEG interference to further enhance their performance (Zhu *et al* 2023). This is in contrast to the AASM rules for human scorers, where the EOG is typically used to identify rapid and slow eye movements, and the EEG interference is typically undesirably. Many examples of sleep staging algorithms based solely on EOG can be found in the literature (Virkkala *et al* 2007, 2008, Kuo *et al* 2014, Liang *et al* 2015, Olesen *et al* 2016, Rahman *et al* 2018, Fan *et al* 2021, Hsieh *et al* 2021, Zhu *et al* 2023). However, the aforementioned studies almost exclusively focus on small cohorts of healthy subjects. Zhu *et al* (2023) recently showed that 5-stage classification with the EOG is possible for subjects with a variety of sleep disorders, but they only used a limited dataset of 26 subjects.

We here implement a novel single-channel EOG staging algorithm using state-of-the-art techniques. Following recent trends in deep learning (Song *et al* 2021, Karras *et al* 2022), we leveraged a network based on transformers (Vaswani *et al* 2017) embedded in a U-Net backbone (Ronneberger *et al* 2015). The U-Net is a convolutional neural network that has shown strong results in a variety of medical segmentation tasks due to its use of a multiscale architecture with skip connections. Its limited field of view is expanded to include the entire night using transformers. The resulting architecture can learn to exploit temporal relations at arbitrarily large time scales.

In this manuscript, we for the first time perform single-lead EOG staging on a relatively large clinical population from the SOMNIA (Sleep and OSA Monitoring with Non-Invasive Applications) dataset (van Gilst *et al* 2019). In total, 774 recordings were used, which were split into 574 train, 100 validation, and 100 hold-out test recordings. We included subjects with a large variety of sleep disorders, including insomnia, sleep-disordered breathing, hypersomnolence, circadian rhythm disorders, parasomnias, and movement disorders.

2. Methods

2.1. Dataset

We made use of recordings from the SOMNIA dataset gathered at the Sleep Medicine Center Kempenhaeghe from subjects with a large variety of sleep disorders (van Gilst *et al* 2019). We included recordings from subjects who underwent a full polysomnography (PSG) between 2017-01-01 and 2021-02-17. The inclusion criteria were: at least 18 years old, presence of the full PSG (which included EOG), and the simultaneous measurement of wrist-worn photoplethysmography (PPG) and actigraphy for future analysis. We excluded recordings with interventions, such as CPAP usage.

In total, 774 overnight recordings were included from 769 subjects, as five subjects underwent two nights of PSG. Each subject underwent an overnight diagnostic sleep study at the Sleep Medicine Center Kempenhaeghe as part of standard clinical care. During the night, a full polysomnography (PSG) was recorded and subsequently manually scored following the AASM guidelines. The bio-electric potentials of EEG, EOG, and EMG were recorded using Ag/AgCl surface electrodes from MFI B.V. (the Netherlands). These signals were then amplified and recorded using a Grael PSG system from Compumedics (USA). For more information about the entire measurement setup, please refer to the original protocol publication (van Gilst *et al* 2019).

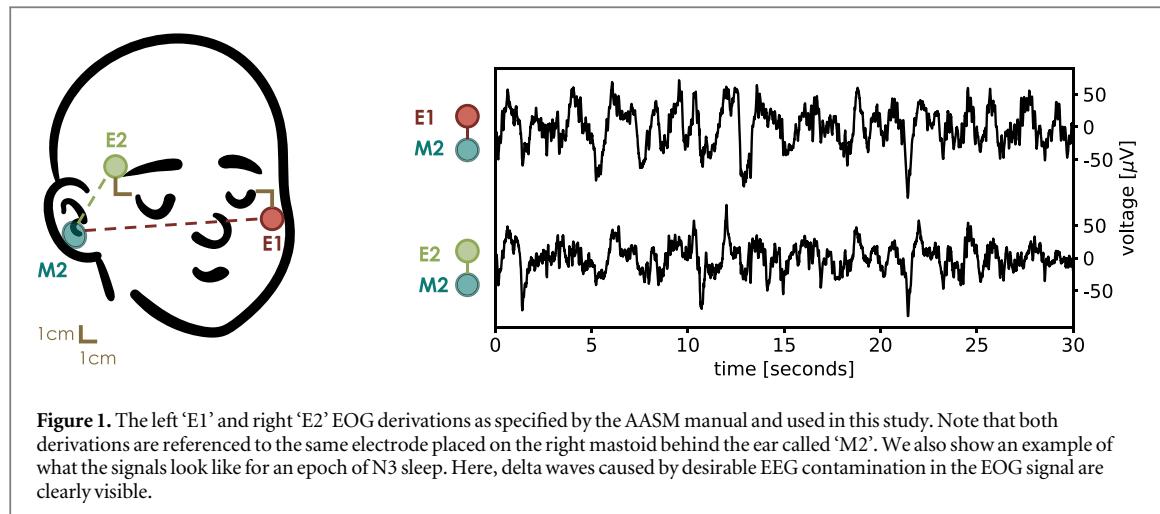


Table 1. Sleep disorders present in different splits of the dataset.^a

Diagnosis	Train	Val.	Test	All
Insomnia	233	40	32	305
Sleep-disordered breathing	344	53	49	446
Hypersomnolence	42	8	11	61
Circadian disorder	13	5	6	24
Parasomnia	74	25	37	136
Movement disorder	104	25	26	155
Other	8	1	2	11
None	17	3	2	22
Number of recordings	574	100	100	774

^a Many subjects had multiple primary diagnoses of sleep disorders, thus the columns do not necessarily add up to the total number of recordings.

Sleep disorder diagnosis was established by a physician following the international classification of sleep disorders, third edition (American Academy of Sleep Medicine 2023). We merged the different sleep disorders into 8 categories, including an 'other' category, which includes sleep disorders occurring in less than 10 participants in this dataset, for example 'sleep-related headache', and a 'none' class in which the primary diagnosis is not a sleep-related disorder. Note that 41% of subjects had more than one sleep disorder present, which is unsurprising since the Sleep Medicine Center Kempenhaeghe is a tertiary sleep clinic.

The dataset was split into 574 train, 100 validation, and 100 hold-out test recordings. The split was made based on the date of the recording, all recordings before 2019-06-04 were included in the train set, the recordings made between 2019-06-04 and 2020-02-06 were included in the validation set, and the newest recordings were used as the hold-out test set. All of the 100 recordings in the hold-out test set came from unique subjects which were not included in either the train or validation sets. See table 1 for the distribution of sleep disorders over the different sets.

2.2. EOG derivations

From the full PSG, we selected the two EOG channels 'E1-M2' and 'E2-M2', which correspond to the left and right EOG, respectively. Following the AASM specifications, the 'E1' electrode is placed 1 cm to the left and 1 cm below the left outer canthus, and the 'E2' electrode is placed 1 cm to the right and 1 cm above the right outer canthus. The 'M2' electrode is placed on the right mastoid, behind the ear. A visual overview is given in figure 1. Because the EOG derivations are not placed symmetrically, we separately evaluated the usage of the left and right EOG derivation in the hold-out test set.

2.3. Preprocessing

We first resampled the EOG from 512 to 128 Hz. Furthermore, we performed band-pass filtering between 0.3 and 49 Hz using fifth-order Butterworth filters applied both in the forward and backward directions, to ensure a

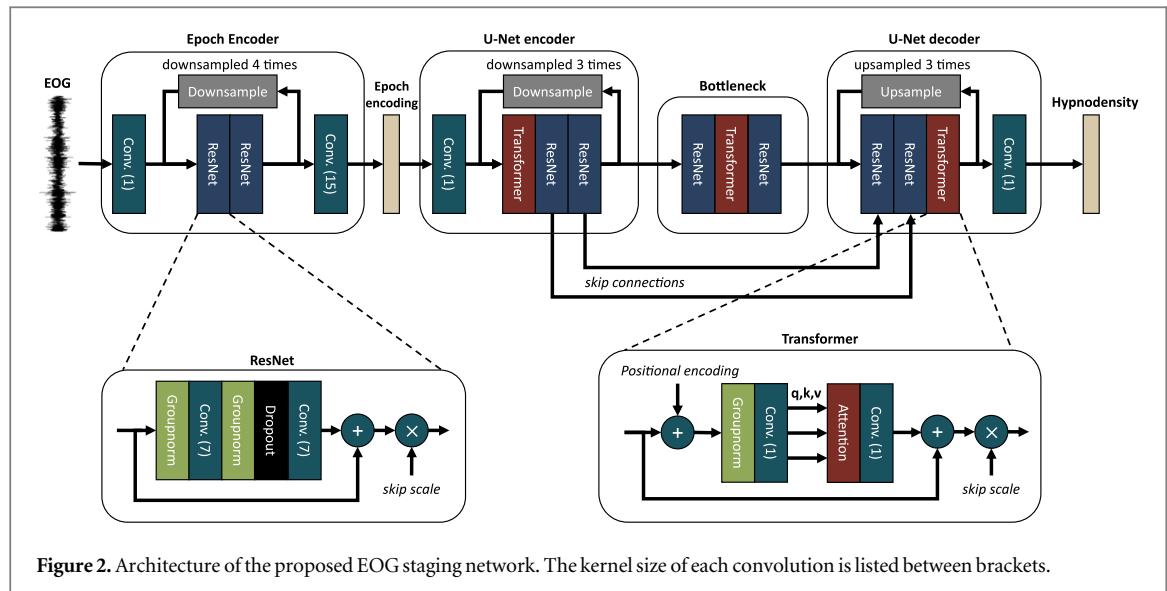


Figure 2. Architecture of the proposed EOG staging network. The kernel size of each convolution is listed between brackets.

zero-phase filtering operation. Additionally, we applied log-normalization per recording to both the EOG signals following (Stephansen *et al* 2018):

$$\mathbf{x} = \text{sign}(\tilde{\mathbf{x}}) \cdot \log\left(\frac{|\tilde{\mathbf{x}}|}{P_{95}(\tilde{\mathbf{x}})} + 1\right), \quad (1)$$

where $\tilde{\mathbf{x}}$ corresponds to an EOG signal of a recording before rescaling, \mathbf{x} is the rescaled EOG signal, and $P_{95}(\tilde{\mathbf{x}})$ is the 95th magnitude percentile. This type of normalization was applied as it is robust against outliers in the data due to the use of the 95th magnitude percentile. Moreover, it pushes up very small values and pushes down very large values due to the use of the logarithm. Lastly, we zero-padded all signals to a size $7 \times 2^8 = 1792$ epochs for implementation purposes. This zero-padding was removed after inference, and before performance evaluation.

2.4. Network architecture

To perform accurate automatic sleep staging using single-channel EOG, we employed transformers (Vaswani *et al* 2017) embedded in a U-Net backbone (Ronneberger *et al* 2015). Specifically, we adapted the DDPM++ architecture from Song *et al* (2021), but with a few changes to make it more suited to the EOG staging task. Firstly, the DDPM++ network was originally proposed for images and as such, it makes use of 2D convolutions. We changed them to 1D convolutions in order to work with time series. Secondly, we added positional encoding to all the self-attention layers, creating transformer-encoder architectures, which are able to learn arbitrary-distance temporal relations within and between the EOG and hypnogram. Thirdly, since we apply the architecture as a discriminative neural network and not in its originally proposed setting of diffusion modeling, we did not make use of ‘noise embedding’. Lastly, we made the ‘U’ asymmetrical by adding an epoch encoder, since the EOG was of much higher dimension than the output hypnogram. See figure 2 for an overview of the network architecture. Figure 2 shows a visual overview of the network architecture, which consists of four distinct stages. The first stage is where the EOG signal is processed by an epoch encoder to extract epoch-level features. To do this, the input signal is compressed from a size of $1792 \text{ epochs} \times 30 \text{ s} \times 128 \text{ samples}$ to $1792 \text{ epochs} \times 16 \text{ features}$. Further details regarding the epoch encoder can be found in appendix A.1.

After the epoch encoding, the output is fed through a U-Net, which is composed of an encoder, bottleneck, and decoder. The U-Net has skip connections added between the encoder and decoder to overcome vanishing gradient problems and allow the network to learn feature embeddings at different time scales in the hypnogram. At the end of the network, a softmax activation function is applied to obtain the probability of each sleep stage at each epoch, which can be interpreted as the hypnodensity. More information about the U-Net encoder, bottleneck, and decoder can be found in appendices A.2, A.3, and A.4, respectively.

The four stages of the network consist of standard building blocks such as convolutional layers, up- and down-sampling layers, dropout layers, and group normalization layers (Wu and He 2018). Additionally, ResNet and transformer layers are used, which are composed of smaller simpler layers, as shown in figure 2. For an in-depth description of the ResNet and transformer layers, please refer to appendix A.5 and appendix A.6, respectively.

2.5. Training

The neural network was trained using both the left and right EOG signal derivations. This was achieved by loading each training recording twice in a dataset iteration during training, once using the left EOG derivation and once using the right EOG derivation. The network was trained using the Adam optimizer (Kingma and Ba 2015) with a learning rate of 10^{-5} . Training continued until convergence, which was monitored using the validation set. Convergence was considered to be reached when the validation loss did not improve for 50 consecutive dataset iterations.

2.6. Testing

After training, we employed the network on the 100 hold-out test recordings using both the left and right EOG derivations. The output hypnograms obtained for each EOG derivation were compared to that of the gold-standard human scoring. For each recording and EOG derivation separately, we computed agreement metrics such as accuracy, Cohen's kappa, and per-class F1 scores. We then calculated both the median and the interquartile range across the 100 test subjects for each metric. Testing was performed in this way in order to ensure that each recording contributed equally, instead of their contribution being based on the total recording time. Furthermore, the median was taken since the metrics were not normally distributed. Additionally, this way of testing allowed for an analysis of each metric per sleep disorder. A confusion matrix was also calculated separately for each EOG derivation based on all aggregated epochs from all hold-out test recordings.

3. Results

3.1. Metrics

The resulting median and interquartile range per metric and diagnosis across the recordings in the hold-out test set are shown in tables 2 and 3 for the network using the left and the right EOG derivation, respectively. In order to aid with comparison to the literature, the mean and standard deviation results are shown in appendix B. As can be seen from tables 2 and 3, the network achieves a slightly better average performance using the right EOG compared to using the left EOG. Using a Wilcoxon signed-rank test to compare the kappa values over the 100 test recordings, this difference was found to be statistically significant ($p = 0.0014$ and test statistic = 1596).

Additionally, we tested within for EOG derivation whether there were significant differences in terms of kappa between subjects with a certain sleep-disorder and those without that sleep disorder. These tests were performed using the Mann-Whitney U rank test. Using a significance level of $p = 0.05$, no significant differences were found.

3.2. Qualitative examples

A qualitative analysis in terms of the predicted hypnodensities and hypnograms was also performed. To that end, we plot the network predictions for the most 'typical' recordings, which were defined as those recordings where the network achieved its median performance in terms of Cohen's kappa for an EOG derivation. Figure 3 shows the most typical recording for the left EOG derivation, where it achieved a kappa of 0.782, and figure 4 shows the most typical recording for the right EOG derivation, with a kappa of 0.796. In both figures, we also show the output using the other EOG derivation. Additionally, appendix C shows qualitative results for random recordings for each diagnosis.

From figures 3 and 4, it can be observed that the predicted hypnograms line up accurately with the ground truth. However, it can be observed that large number of N1 epochs are missed by the network. The difficulty of scoring N1 epochs can also be observed by analyzing the hypnodensities, especially those in figure 4. Uncertainty between the wake and N1 class can be observed at 5 and 9 h into the night.

3.3. Confusion matrices

The confusion matrices calculated over all epochs are shown in figure 5. Overall, there were very few wrongly predicted classes. Only the N1-stage sensitivity was low, with most of the confusion being towards the N2 class. Additionally, there was some confusion with respect to scoring a ground-truth N3 epoch as N2.

4. Discussion

Modern sleep medicine demands accurate and inexpensive sleep staging systems that can reliably be trusted regardless of any underlying sleep condition. Automatic sleep staging using single-channel EOG promises to be a suitable solution to this challenge. However, most EOG staging algorithms described in literature have almost exclusively been tested on healthy subjects, or on small cohorts of patients with sleep disorders. We for the first

Table 2. Results for the network using the left EOG derivation, we show the median*/the interquartile range across the recordings.

Diagnosis	# Test recordings	Accuracy	Cohen's kappa	F1 scores				
				Wake	N1	N2	N3	REM
Insomnia	32	86.7%/7.4%	0.812/0.103	0.906/0.093	0.553/0.087	0.891/0.078	0.850/0.125	0.895/0.095
Sleep-disordered breathing	49	84.3%/7.9%	0.771/0.102	0.888/0.099	0.566/0.132	0.865/0.075	0.849/0.127	0.865/0.072
Hypersomnolence	11	85.8%/4.4%	0.781/0.062	0.898/0.075	0.598/0.195	0.901/0.063	0.856/0.104	0.867/0.082
Circadian disorder	6	85.6%/4.3%	0.782/0.059	0.852/0.095	0.514/0.110	0.908/0.048	0.904/0.091	0.860/0.048
Parasomnia	37	84.7%/9.1%	0.768/0.106	0.874/0.133	0.559/0.137	0.889/0.075	0.842/0.128	0.872/0.053
Movement disorder	26	85.3%/9.4%	0.800/0.130	0.913/0.080	0.560/0.170	0.874/0.071	0.843/0.080	0.895/0.075
Other	2	89.2%/4.5%	0.844/0.051	0.918/0.057	0.535/0.070	0.887/0.023	0.938/0.009	0.903/0.032
None	2	88.2%/1.8%	0.822/0.032	0.876/0.044	0.610/0.086	0.908/0.006	0.945/0.010	0.841/0.021
All	100	85.0%/8.0%	0.781/0.107	0.894/0.109	0.548/0.136	0.885/0.074	0.850/0.120	0.871/0.084

Table 3. Results for the network using the right EOG derivation, we show the median^a/the interquartile range across the recordings.

Diagnosis	# Test recordings	Accuracy	Cohen's kappa	F1 scores				
				Wake	N1	N2	N3	REM
Insomnia	32	86.1%/7.1%	0.812/0.095	0.918/0.091	0.583/0.113	0.892/0.069	0.871/0.106	0.901/0.074
Sleep-disordered breathing	49	84.9%/5.8%	0.795/0.085	0.897/0.079	0.571/0.101	0.878/0.058	0.869/0.120	0.871/0.094
Hypersomnolence	11	84.5%/7.6%	0.766/0.114	0.891/0.054	0.565/0.144	0.888/0.072	0.814/0.128	0.839/0.077
Circadian disorder	6	85.6%/5.7%	0.788/0.073	0.877/0.073	0.513/0.051	0.914/0.040	0.920/0.067	0.849/0.066
Parasomnia	37	84.5%/7.9%	0.788/0.121	0.896/0.123	0.532/0.118	0.886/0.081	0.863/0.150	0.877/0.070
Movement disorder	26	84.7%/8.5%	0.783/0.098	0.912/0.081	0.569/0.124	0.872/0.060	0.860/0.105	0.881/0.074
Other	2	89.1%/4.3%	0.844/0.048	0.910/0.064	0.526/0.074	0.897/0.014	0.958/0.017	0.853/0.018
None	2	88.5%/1.8%	0.828/0.033	0.888/0.032	0.614/0.107	0.910/0.002	0.950/0.013	0.832/0.028
All	100	85.2%/6.9%	0.796/0.103	0.902/0.107	0.551/0.123	0.882/0.064	0.867/0.129	0.874/0.097

^a The median for an even number of recordings was defined as the average of the middle two values.

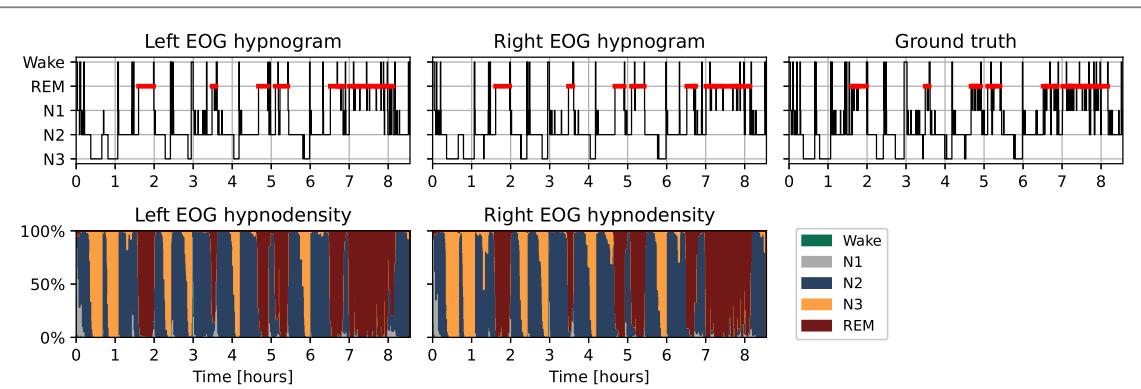


Figure 3. Qualitative results for the recording where the left EOG derivation achieved median performance in terms of Cohen's kappa (0.782). This subject had a parasomnia diagnosis.

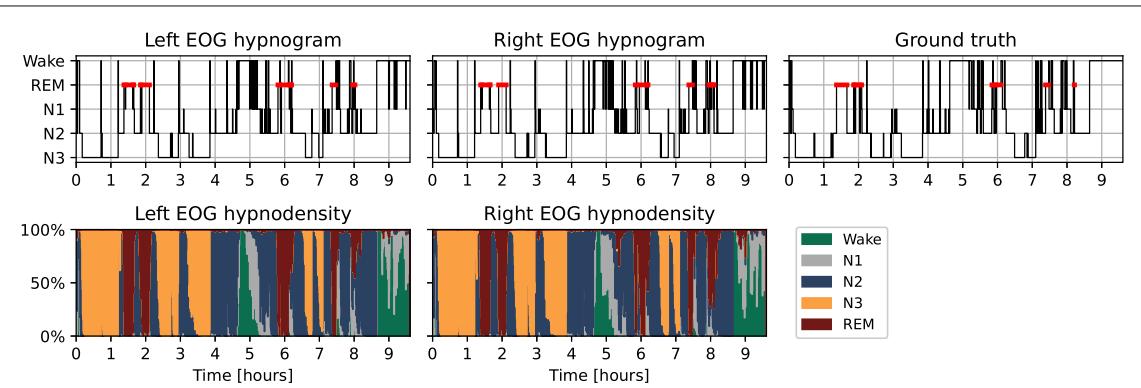


Figure 4. Qualitative results for the recording where the right EOG derivation achieved median performance in terms of Cohen's kappa (0.796). This subject had an 'other' diagnosis.

time we show that automatic sleep staging using single-channel EOG can be done reliably for subjects with a variety of sleep disorders in a relatively large cohort.

4.1. Inter-rater agreement

To put our sleep staging results into context, we can compare it to the human inter-rater agreement which serves as an upper limit on performance (van Gorp *et al* 2022). The human inter-rater agreement has been widely studied in the context of the AASM scoring rules, with Rosenberg and Van Hout (2013) conducting an especially large study comparing the scoring behavior of over 2500 scorers. When comparing a single scorer against a group consensus, Rosenberg and Van Hout (2013) found an average inter-rater agreement of 82.6%. Splitting this agreement into the different stages they found an 84.1% agreement for Wake, 63.0% for N1, 85.2% for N2, 67.4% for N3, and a 90.5% agreement for REM.

The sleep technologists at the SOMNIA data collection site, Sleep Medicine Center Kempenhaeghe, have shown a very high inter-rater agreement, with around 86% agreement on both the AASM interrater agreement program (Rosenberg and Van Hout 2013) and internal institutional inter-rater agreement assessment. Additionally, a second scoring by another technician of the same institution was available for 14 recordings in the dataset used in the present study. This allowed us to estimate the inter-rater agreement for this subset, which was in average 86.1%. Separately evaluating agreement for different stages results yielded a 93.1% agreement for Wake, 64.8% for N1, 85.6% for N2, 88.1% for N3, and 89.1% for REM. Comparing these results with the confusion matrices from figure 5, we can see that our EOG system reaches similar levels of performance.

While our N1 performance is the lowest out of all the classes, 53.5% and 53.3%, this is also the case for the human inter-rater agreement, which was found to be only 63.0% and 64.8% by Rosenberg and Van Hout (2013) and internally at Kempenhaeghe, respectively. This underlines the difficulty of accurately detecting the N1 stage, even for human scorers, and shows that our N1 classification performance is actually close to the expected upper limit, defined by inter-rater human agreement. The high degree of uncertainty about the N1 class is also reflected in the hypnodensity graphs, see for example 4. Nikkinen *et al* (2023) also point out the low agreement of the N1

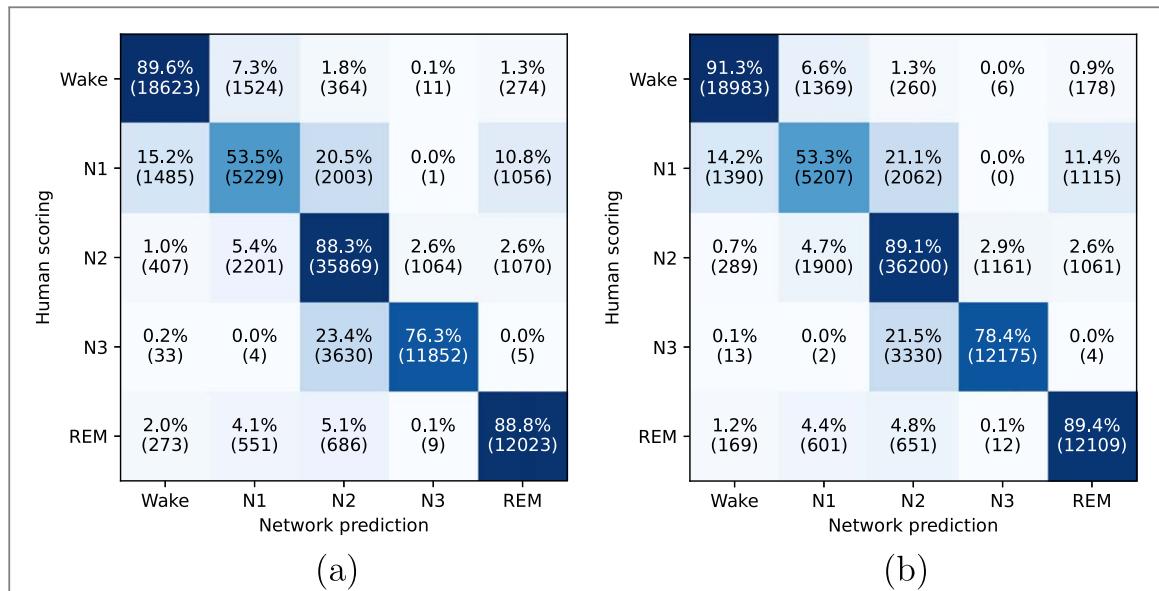


Figure 5. Confusion matrices for the network with the total number of epochs listed between brackets. The EOG derivations used are (a) left EOG, (b) right EOG.

stage between human scorers and highlight different factors that contribute it. Among these are: stage transitions, in particular N1 to N2, the low amount of N1 sleep in general, and the variety in alpha frequencies between individuals, which may even not be present at all.

Most strikingly, our N3 classification performance is substantially higher than the inter-rater agreement found by Rosenberg and Van Hout (2013), 76.3% and 78.4% versus 67.4%. This is similar to the (human) inter-rater agreement for the N3 stage in our dataset, which is 88.1%. This can be explained by the fact that Rosenberg and Van Hout (2013) looked at the differences between scorers coming from a large variety of backgrounds and institutions, while the technicians from the SOMNIA set all came from the same clinic. This results in lower ambiguity about the scoring of N3 for a sleep staging algorithm trained with, and evaluated on data from the same institution (van Gorp *et al* 2022).

4.2. EOG-enabled sleep staging literature

Automatic sleep staging based solely on the EOG dates back to 2007 (Virkkala *et al* 2007). Classical machine learning methods based on feature extraction and classification algorithms such as random forest and support vector machines were explored first (Virkkala *et al* 2007, 2008, Kuo *et al* 2014, Liang *et al* 2015, Olesen *et al* 2016, Rahman *et al* 2018). More recently, end-to-end learning based on deep neural networks has also been described (Fan *et al* 2021, Hsieh *et al* 2021, Zhu *et al* 2023). The reported performance metrics for all of these methods have been promising, with accuracies ranging between 70.8% and 91.7% and Cohen's kappa ranging between 0.60 and 0.806. However, almost all algorithms have only been validated on datasets of healthy participants and of limited size, with the largest set used by Virkkala *et al* numbering 263 subjects. On the other hand, only Zhu *et al* (2023) did a study on subjects with a mix of sleep disorders, but with a limited set of only 26 participants.

Besides the standard PSG, custom devices to record the EOG have also been proposed in the literature. As stated in the introduction, the EOG can for example be measured with dry electrodes embedded in a sleep mask (Liang *et al* 2015, Hsieh *et al* 2021). On the other hand, portable frontal EEG solutions could also be considered, as the frontal EEG shares many similarities with the EOG. This type of solution dates back even earlier than single channel EOG sleep staging, with devices such as QUISI (Ehlert *et al* 1998) and Biosomnia (Schweitzer *et al* 2004). Modern iterations on this concept can also be found in the literature (Finan *et al* 2016, Bresch *et al* 2018).

To allow for widespread adoption in the clinic, automatic staging based on EOG has to achieve reliable performance not only on healthy subjects but especially on those subjects with (possibly multiple) sleep disorders. A particular strength of this study is that we show that EOG staging is possible by evaluating our method on 100 hold-out test recordings from subjects with a large variety of disorders. The model is robust to the presence of different sleep disorders, exhibiting no significant differences between these.

4.3. Effect of the EOG electrode location

Performance with the right EOG was significantly higher than with the left EOG. This effect might be explained by the fact that the right EOG electrode is placed closer to the top of the scalp in comparison to the left EOG, see

figure 1. Because of this, more desired EEG interference could get coupled into the right EOG, thereby enabling it to perform more accurate sleep staging. Additional research is however needed to establish what the effect of EOG electrode placement is on the performance of automatic sleep staging algorithms. This is particularly important since, for instance, the AASM manual describes both a recommended placement of EOG electrodes—which we used—but also an ‘acceptable’ placement (Troester *et al* 2023). Following the acceptable recommendation, both electrodes are placed 1 cm below the outer canthus and referenced to the Fpz’ electrode, possibly resulting in different signal characteristics. The proposed algorithm could be trained to adapt to these differences through transfer learning, a process that has been proven to work well for differences in frontal EEG electrode placement and acquisition (van der Aar *et al* 2024).

4.4. Model architecture

The architecture of our sleep staging model is based on highly expressive models found in the score-based diffusion literature (Song *et al* 2021, Karras *et al* 2022). These models rely on a U-Net backbone, which was originally proposed for medical image segmentation (Ronneberger *et al* 2015), and has been applied with success to the sleep staging task (Perslev *et al* 2021, van Gorp *et al* 2023). Additionally, transformer layers were added, to allow the network to learn associations at arbitrary time scales. This contrasts with convolutional and recurrent neural networks, which have a strong architectural bias towards learning relationships at smaller time scales and between neighboring samples. The ability to exploit relationships at long time scales can be very useful in the context of sleep staging, for example in the application of the REM continuation rule, which states that REM should be continued to score even in the absence of rapid eye movements under certain conditions such as low muscle tone and the absence of k-complexes and arousals (Troester *et al* 2023).

Other sleep staging algorithms employing transformers have also been proposed in the literature. For example, Phan *et al* (2022) proposed SleepTransformer which uses single-channel EEG data in a convolution- and recurrent neural network-free architecture. The employed sequence length in SleepTransformer is however only 21 epochs, which contrasts with our model, where the entire overnight recording is used. More comparable with our model, sleep staging algorithms that combine attention or transformer layers with convolutional layers have been previously described (Qu *et al* 2020, Zhu *et al* 2020, Eldele *et al* 2021).

To evaluate some of the neural network layers used, a post-hoc ablation study was performed, which can be found in appendix D. Such post-hoc analyses come with the caveat that if one tries to find the parameter settings with highest test set performance, test set leakage may occur (Kapoor and Narayanan 2023). Still, from the ablation experiments it can be concluded that the skip connections, group normalization layers, and dropout layers are essential for good sleep staging performance in this model. Ablating either the large U-Net convolutions or transformer layers did not lead to a significant change in sleep staging performance, indicating that both are valid strategies for learning associations between epochs in their own right.

4.5. Limitations

There are some limitations to this study worth remarking. Firstly, training and testing of the network were done with recordings obtained and scored at the same institution. This could lead to sampling biases in terms of confounders such as recording characteristics, medication use, and patient demographics. Future work should investigate the effects of dataset distribution shift on the final performance of the network, as differences in measurement setup, scoring behavior, and population characteristics between different institutions can impact performance. However, it is important to note that the dataset used in this study comes from a third-line sleep clinic and represents one of the more difficult populations to perform sleep staging on.

Secondly, no extensive hyper-parameter search was performed, for example on kernel size or channel depth. Since the performance of the network was already on par with the inter-rater agreement, we hypothesize that hyper-parameter tuning would not necessarily lead to significant improvements, at least on this dataset.

Thirdly, our study used EOG channel derivations from a full PSG, as applied by medical specialists. To lower healthcare costs and enable ambulatory sleep studies, it would be interesting to study the performance of our network in the context of self-applied EOG electrodes (Virkkala *et al* 2008) or a mask with dry electrodes (Liang *et al* 2015, Hsieh *et al* 2021). Transfer learning could also be used in this case to overcome eventual changes in signal characteristics (van der Aar *et al* 2024).

Fourthly, using a single-channel EOG means that certain measurements can be missed. For example, if the EOG electrode detaches during the night, there is no backup channel available. Moreover, for certain specific disorders, single-channel EOG might not be sufficient, for example in the case of sleep-related epilepsy, where the concurrent measurement of EEG would be beneficial.

Lastly, single-channel EOG may not be the end-all solution, as single-channel EEG or other surrogate measures can also be leveraged by automatic sleep staging algorithms. Especially the signal measured by a very frontal EEG electrode would be similar to the current setup. The EOG is however already part of the standard

AASM PSG setup, while pre-frontal EEG electrodes such as Fp1 and Fp2 are not. When choosing a measurement setup for sleep staging, a trade-off between signal quality, ease of measurement, cost, and subject (dis)comfort should be made, depending on the subject and suspected sleep disorder.

5. Conclusion

In summary, we developed an automatic sleep staging algorithm for single-channel EOG leveraging transformers embedded in a U-Net backbone. We used a relatively large dataset of 774 recordings with a variety of sleep disorders, split into 574 train, 100 validation, and 100 hold-out test recordings. We verified that the performance of our network was on par with the human inter-rater agreement. Furthermore, we found no significant differences in performance between subjects with different sleep disorders. The main findings of this study are as follows. Firstly, the proposed architecture and training mechanism are effective in EOG-based sleep staging. Secondly, the performance of automatic sleep staging based solely on a single channel EOG comes very close to the human inter-rater agreement. Thirdly, the differences in location between the left and right EOG electrodes, as recommended by AASM, affect the sleep staging performance. Lastly, the use of single-channel EOG in automatic sleep staging has shown similar performance regardless of the underlying of sleep disorders. These results pave the way for the adoption of automatic sleep staging using single-channel EOG in clinical settings where subjects with complex disorders can be encountered.

Data availability statement

The SOMNIA data used in this study is available from the Sleep Medicine Centre Kempenhaeghe upon reasonable request. The data can be requested by presenting a scientific research question and by fulfilling all the regulations concerning the sharing of the human data. The details of the agreement will depend on the purpose of the data request and the entity that is requesting the data (e.g. research institute or corporate). Each request will be evaluated by the Kempenhaeghe Research Board and, depending on the request, approval from independent medical ethical committee might be required. Access to data from outside the European Union will further depend on the expected duration of the activity; due to the work required from a regulatory point of view, the data is less suitable for activities that are time critical, or require access in short notice. Specific restrictions apply to the availability of the data collected with sensors not comprised in the standard PSG set-up, since these sensors are used under license and are not publicly available. These data may however be available from the authors with permission of the licensors. For inquiries regarding availability, please contact Merel van Gilst (M.M.v.Gilst@tue.nl).

Ethical statement

The SOMNIA study (van Gilst *et al* 2019) was reviewed by the medical ethical committee of the Maxima Medical Center (Veldhoven, the Netherlands. File no: N16.074 and W17.128). The protocol for data analysis was approved by the Institutional Review Board of the Kempenhaeghe hospital. The data collection as part of the SOMNIA study was conducted in accordance with the declaration of Helsinki and in accordance with local statutory requirements.

Conflicts of interest statement

This work was performed within the IMPULSE framework of the Eindhoven MedTech Innovation Center (e/MTIC, incorporating Eindhoven University of Technology, Philips Research, and Sleep Medicine Center, Kempenhaeghe Foundation), including a PPS supplement from the Dutch Ministry of Economic Affairs and Climate Policy. At the time of writing, HG and PF were employed and/or affiliated with Royal Philips, a commercial company and manufacturer of consumer and medical electronic devices, commercializing products in the area of sleep diagnostics and sleep therapy. Philips had no role in the study design, decision to publish, or preparation of the manuscript. PF reports personal fees from Philips during the conduct of the study; personal fees from Philips, outside the submitted work. SO received an unrestricted research grant from UCB Pharma and participated in advisory boards for Jazz Pharmaceuticals, Bioprojet, and Abbvie. All unrelated to present work.

Appendix A. Details regarding the network architecture

In this appendix, we elaborate on the architectural details of the neural network, as shown in figure A1.

A.1. Epoch encoder

Because the EOG signals and the hypnograms had different sampling frequencies (128 Hz versus 1/30 Hz), we first needed to downsample the EOG before we could use the U-Net structure of our model. To that end, we employed a context encoder, which downsampled the EOG signal from $\mathbb{R}^{1792 \times 30 \times 128 \times 1}$ to $\mathbb{R}^{1792 \times 16}$, i.e. a context encoding of length number of epochs with 16 channels.

The context encoder worked as follows. First, a convolution of kernel size 1 expanded the number of channels from 1 to 16. Then, a series of two ResNets was employed to extract meaningful features from the EOG signal (see A.5 for further details regarding the ResNet). This pattern was repeated 5 times with 4 downsampling operations between the 5 blocks. Each downsampling operation used a kernel of [1,1,1,1] and a stride of 4, to effectively downsample the input by a factor of 4. At the end of the epoch encoder, another convolution of kernel and stride 15 was used, thus compressing the EOG signal to a feature map of size $\mathbb{R}^{1792 \times 16}$ which was used as input to the U-Net encoder.

A.2. U-Net encoder

The U-Net encoder first employed a convolution of kernel size 1 to increase the channel size from 16 to 32. Then, a Transformer layer together with two ResNet blocks was employed (see A.6 for further details regarding the Transformer layer). After each ResNet block, a skip connection was added to the U-Net decoder at the same resolution. This pattern of a transformer with two ResNets was repeated 4 times with 3 downsampling operations in between. Again, a kernel of [1,1,1,1] and a stride of 4 was used in the downsampling operations. Note that in the original DDPM++ implementation (Song *et al* 2021), an attention layer was added after each ResNet in the encoder. However, to bring down the computational complexity of our method and to make the encoder symmetric with the decoder, we employed only a single transformer layer at the start of each resolution level in the U-Net encoder.

A.3. Bottleneck

In the bottleneck, the feature map was of its smallest size, namely $\mathbb{R}^{28 \times 16}$. Here, one transformer layer sandwiched between two ResNet blocks was used to learn the highest-level features of the hypnogram.

A.4. U-Net decoder

The decoder followed a mirrored structure to the encoder. The skip connections from the corresponding resolution levels were concatenated to the inputs of each ResNet block. These connections allowed the feature maps to skip the downward path of the ‘U’ and enabled the model to learn both high-and-low level features of the hypnogram. The upsampling operation of the decoder was implemented using a transposed convolution with the same filter of [1,1,1,1].

As a final step toward creating a hypnogram, the U-Net decoder employed a convolution of kernel size 1 to map the input to 5 channels, where each channel corresponded to one of the five sleep stages. A softmax activation function was then used to map each channel to a class probability. This creates a ‘hypnodensity’, a soft version of the hypnogram where each epoch is partially associated with each sleep stage according to some probability (Stephansen *et al* 2018). If instead a ‘hard’ hypnogram is desired, the argmax of the hypnodensity can be taken.

A.5. ResNet

The ResNet, or Residual Network, was repeated throughout the architecture. It consists of two group normalization layers and two convolutions in an alternating pattern. Group normalization, as described by Wu and He (2018), applies a learned normalization across groups of channels, enabling faster training. In our case, each group consisted of 4 channels. The 1D convolutions of the ResNet each used a kernel of size 7 and zero-padding set to ‘same’. Each convolution was followed by SiLU (Sigmoid Linear Unit) activation (Hendrycks and Gimpel 2016). Additionally, a spatial dropout layer was added before the second convolution, which drops out entire channels during training with a probability of 10%. Spatial dropout is a better regularizer for convolutional neural networks, since neighbouring samples are often highly correlated (Tompson *et al* 2015). Finally, a residual connection was added to help combat vanishing gradient problems. To limit the magnitude of the signals, scaling with a factor of skip scale = $\sqrt{0.5}$ was applied.

A.6. Transformer layer

The original transformer architecture is a sequence-to-sequence model composed of both an encoder and a decoder (Vaswani *et al* 2017). Where each element consists of a scaled dot-product attention layer and an element-wise feed-forward network. Additionally, positional encoding is added at the start of the encoding and decoding stacks. We adapt the transformer architecture to be suited for our network. Firstly, we did not use the decoder, since it is used to generate new sequence in an auto-regressive manner. Secondly, since we embedded the layers within a larger convolutional neural network, there was no need for separate element-wise feed-

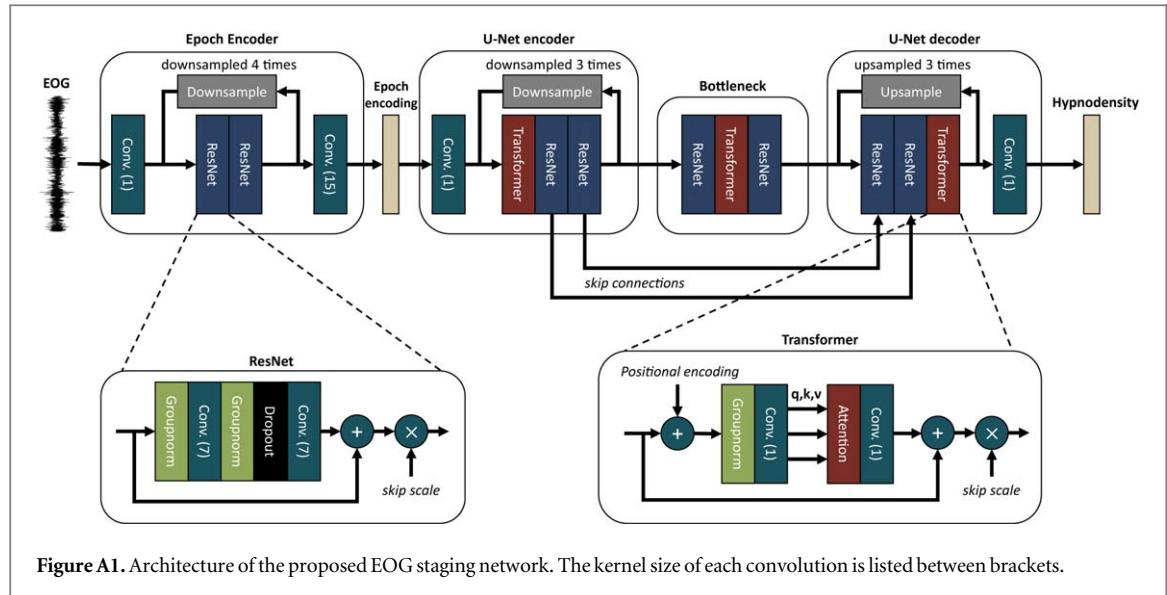


Figure A1. Architecture of the proposed EOG staging network. The kernel size of each convolution is listed between brackets.

forward networks. lastly, because the attention layers operated at different time scales, we added positional encoding to each of them.

The positional encoding was implemented using sine-cosine embedding. In this scheme, a positional encoding matrix is added element-wise to the input sequence of the transformer. To that end, the input sequence \mathbf{S} and positional encoding matrix \mathbf{P} should be of the same size: $\mathbf{S}, \mathbf{P} \in \mathbb{R}^{L \times C}$, where L is the length of the input sequence and C is the number of channels. The positional encoding matrix is given by:

$$\begin{aligned} \mathbf{P}_{(l,2d)} &= \sin(l \cdot 1000^{-2c/C}) \\ \mathbf{P}_{(l,2d+1)} &= \cos(l \cdot 1000^{-2c/C}), \end{aligned} \quad (\text{A.1})$$

with $l \in [0, 1, \dots, L - 1]$ and $c \in [0, 1, \dots, C - 1]$. This type of encoding enables the transformer to exploit information about both the absolute and relative positions of samples along the night.

Each of the transformer layers used scaled dot-product self-attention. While the attention mechanism can be implemented using multiple attention-heads for added complexity, we here only made use of a single head. In scaled dot-product self-attention, three linear projections are applied to transform the sequence to a query, key, and value matrix:

$$\mathbf{Q} = \mathbf{S}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{S}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{S}\mathbf{W}_V, \quad (\text{A.2})$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{C \times C}$ are learned linear projection weights and $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{L \times C}$ are the query, key, and value matrices, respectively. These linear projections can be implemented efficiently by a single convolutional layer of kernel size 1 and output channel size of $3C$, as its output can be split along the channel dimension into the three separate components.

Following a database analogy, the queries are going to look for matching keys and propagate the associated values to the output, where each individual query, key, and value are found along the rows of their respective matrices. This process is defined by the scaled dot-product self-attention mechanism:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{C}}\right)\mathbf{V}, \quad (\text{A.3})$$

where \mathbf{K}^T denotes the transpose of the key matrix. Moreover, $\mathbf{Q}\mathbf{K}^T \in \mathbb{R}^{L \times L}$ denotes the attention map. To ensure that the magnitudes in the attention map do not grow too large, it is scaled down by a factor of $1/\sqrt{C}$. Additionally, a softmax activation is applied along the rows of the attention map in order to ensure that the attention sums to 1.

After the scaled dot-product attention layer, another linear projection using a 1D convolution was applied. Similar to the ResNet, a residual connection was applied with a scaling of skip scale = $\sqrt{0.5}$.

Appendix B. Additional quantitative results per diagnosis

In this appendix, we provide additional quantitative results in terms of the mean and standard deviation across the recordings, see tables B1 and B2.

Table B1. Results for the network using the left EOG derivation, we show the mean \pm the standard deviation across the recordings.

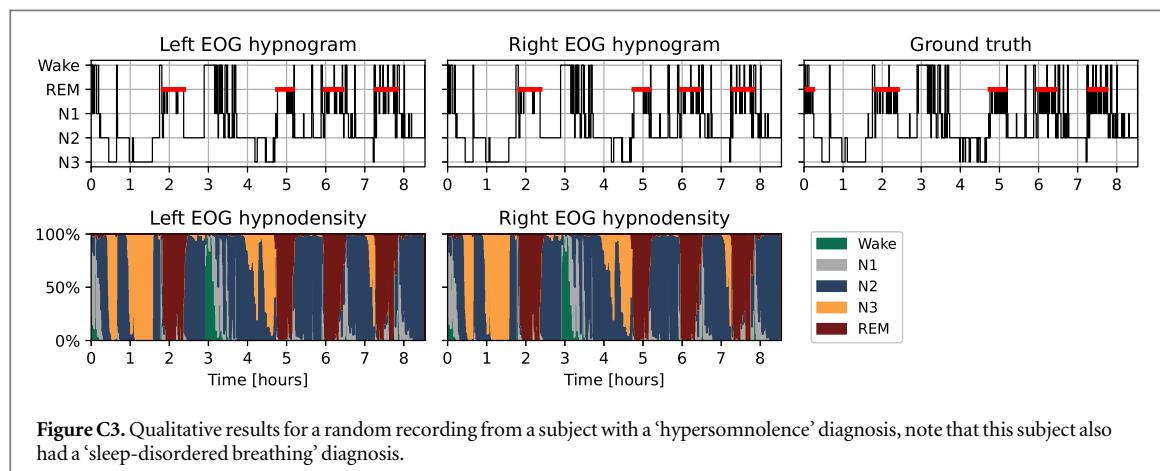
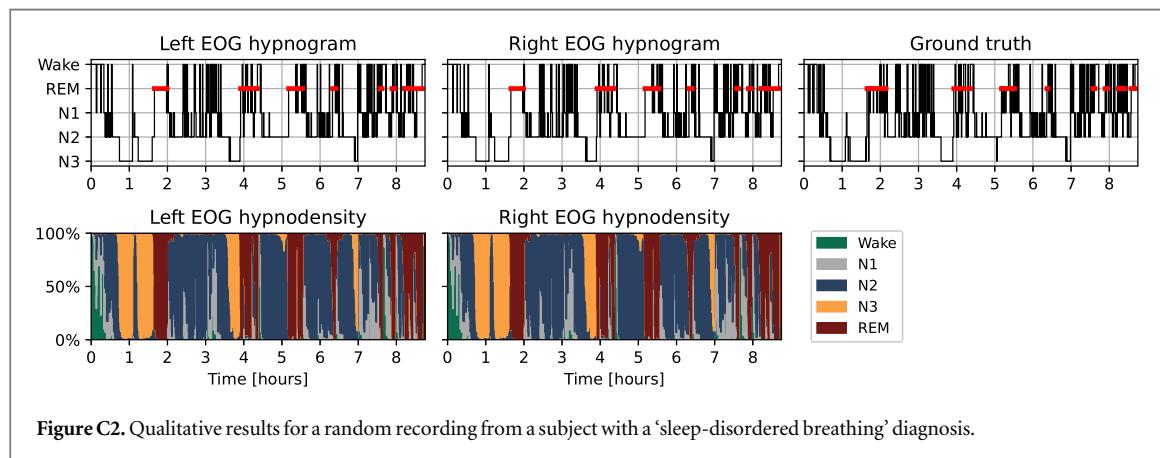
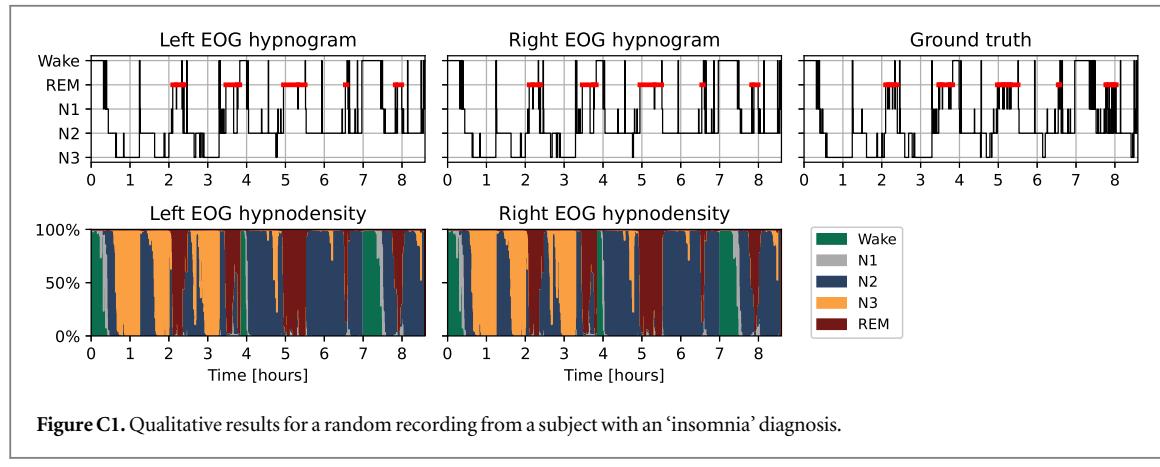
Diagnosis	# Test recordings	Accuracy	Cohen's kappa	F1 scores				
				Wake	N1	N2	N3	REM
Insomnia	32	83.7% \pm 9.8%	0.767 \pm 0.136	0.888 \pm 0.081	0.550 \pm 0.126	0.843 \pm 0.140	0.764 \pm 0.256	0.834 \pm 0.195
Sleep-disordered breathing	49	83.1% \pm 6.5%	0.763 \pm 0.090	0.872 \pm 0.098	0.546 \pm 0.117	0.858 \pm 0.065	0.788 \pm 0.218	0.828 \pm 0.156
Hypersomnolence	11	83.6% \pm 6.4%	0.764 \pm 0.085	0.869 \pm 0.103	0.515 \pm 0.169	0.881 \pm 0.055	0.809 \pm 0.128	0.855 \pm 0.067
Circadian disorder	6	85.5% \pm 3.6%	0.789 \pm 0.051	0.864 \pm 0.074	0.546 \pm 0.079	0.894 \pm 0.037	0.834 \pm 0.182	0.844 \pm 0.051
Parasomnia	37	82.8% \pm 7.6%	0.757 \pm 0.106	0.858 \pm 0.110	0.546 \pm 0.118	0.847 \pm 0.096	0.760 \pm 0.248	0.816 \pm 0.189
Movement disorder	26	84.0% \pm 6.0%	0.777 \pm 0.081	0.888 \pm 0.092	0.539 \pm 0.135	0.850 \pm 0.076	0.801 \pm 0.169	0.877 \pm 0.072
Other	2	89.2% \pm 6.4%	0.844 \pm 0.072	0.918 \pm 0.081	0.535 \pm 0.099	0.887 \pm 0.032	0.938 \pm 0.013	0.903 \pm 0.045
None	2	88.2% \pm 2.5%	0.822 \pm 0.045	0.876 \pm 0.063	0.610 \pm 0.121	0.908 \pm 0.008	0.945 \pm 0.015	0.841 \pm 0.030
All	100	83.4% \pm 7.7%	0.763 \pm 0.106	0.873 \pm 0.093	0.538 \pm 0.124	0.850 \pm 0.101	0.800 \pm 0.192	0.830 \pm 0.165

Table B2. Results for the network using the left EOG derivation, we show the mean \pm the standard deviation across the recordings.

Diagnosis	# Test recordings	Accuracy	Cohen's kappa	F1 scores				
				Wake	N1	N2	N3	REM
Insomnia	32	86.0% \pm 5.0%	0.800 \pm 0.068	0.904 \pm 0.063	0.572 \pm 0.094	0.872 \pm 0.069	0.811 \pm 0.213	0.865 \pm 0.129
Sleep-disordered breathing	49	84.0% \pm 6.1%	0.775 \pm 0.085	0.881 \pm 0.103	0.553 \pm 0.113	0.868 \pm 0.056	0.807 \pm 0.207	0.820 \pm 0.185
Hypersomnolence	11	82.7% \pm 6.2%	0.752 \pm 0.084	0.877 \pm 0.085	0.507 \pm 0.153	0.871 \pm 0.059	0.785 \pm 0.146	0.848 \pm 0.061
Circadian disorder	6	86.1% \pm 3.5%	0.797 \pm 0.050	0.871 \pm 0.071	0.537 \pm 0.044	0.898 \pm 0.042	0.838 \pm 0.206	0.848 \pm 0.055
Parasomnia	37	83.8% \pm 7.2%	0.769 \pm 0.102	0.867 \pm 0.122	0.547 \pm 0.114	0.855 \pm 0.102	0.765 \pm 0.255	0.808 \pm 0.219
Movement disorder	26	84.7% \pm 5.6%	0.786 \pm 0.075	0.897 \pm 0.083	0.543 \pm 0.127	0.858 \pm 0.075	0.816 \pm 0.165	0.881 \pm 0.058
Other	2	89.1% \pm 6.1%	0.844 \pm 0.068	0.910 \pm 0.090	0.526 \pm 0.104	0.897 \pm 0.020	0.958 \pm 0.024	0.853 \pm 0.026
None	2	88.5% \pm 2.5%	0.828 \pm 0.046	0.888 \pm 0.046	0.614 \pm 0.151	0.910 \pm 0.004	0.950 \pm 0.018	0.832 \pm 0.039
All	100	84.5% \pm 6.0%	0.779 \pm 0.083	0.884 \pm 0.091	0.547 \pm 0.110	0.862 \pm 0.079	0.813 \pm 0.187	0.837 \pm 0.158

Appendix C. Qualitative results per diagnosis

In this appendix, we provide additional qualitative results for a random recording from each diagnostic group, see figures C1–C8.



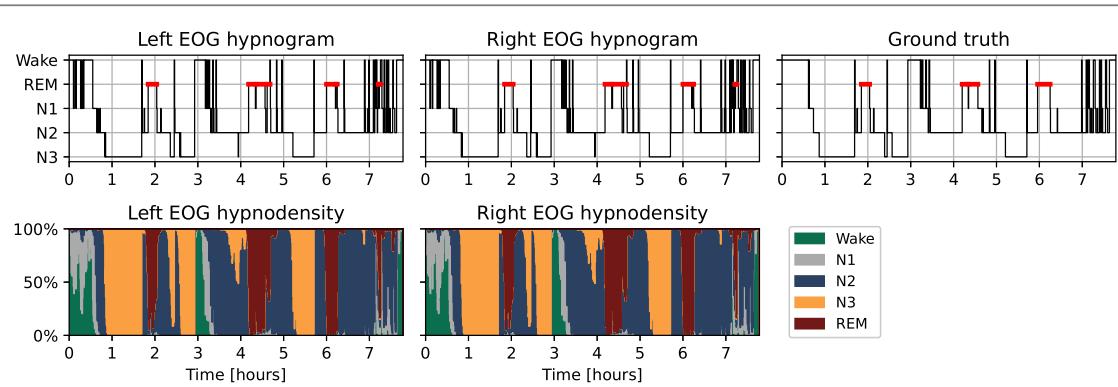


Figure C4. Qualitative results for a random recording from a subject with a ‘circadian disorder’ diagnosis, note that this subject also had a ‘sleep-disordered breathing’ and a ‘parasomnia’ diagnosis.

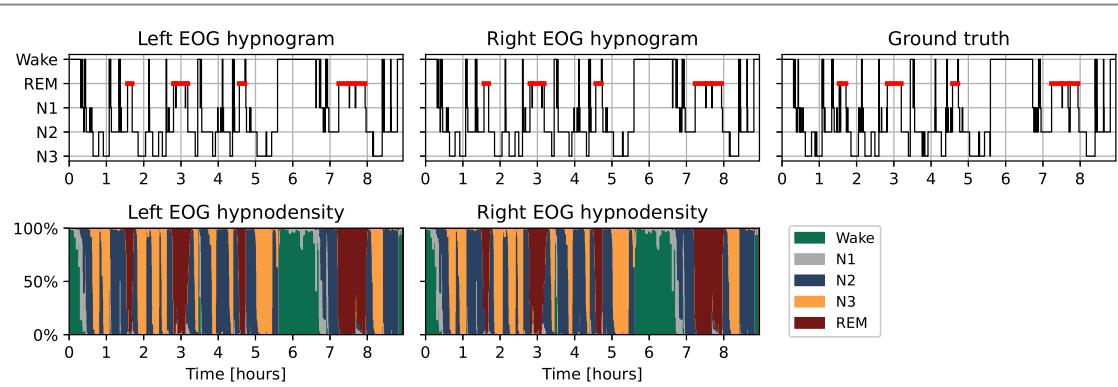


Figure C5. Qualitative results for a random recording from a subject with a ‘parasomnia’ diagnosis.

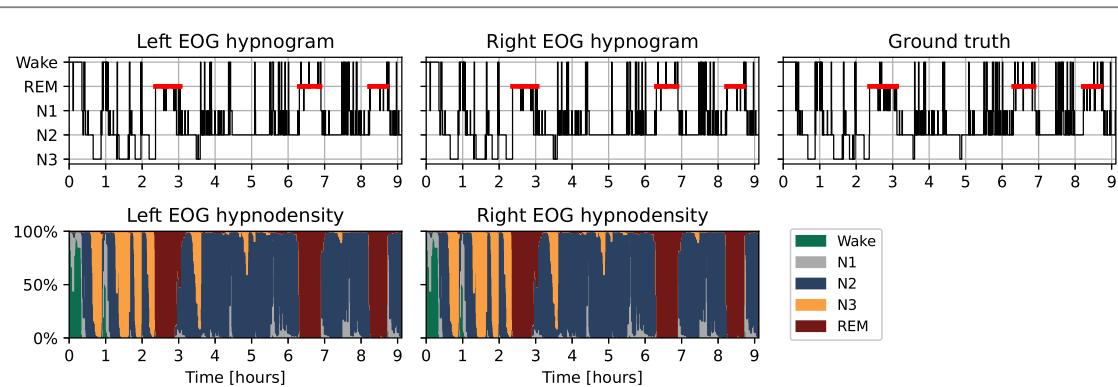


Figure C6. Qualitative results for a random recording from a subject with a ‘movement disorder’ diagnosis.

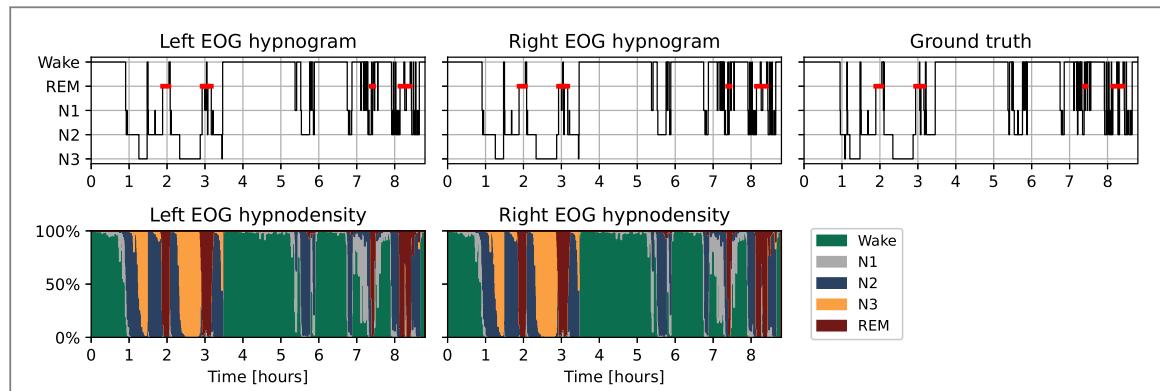


Figure C7. Qualitative results for a random recording from a subject with an ‘other’ diagnosis, note that this subject also had a ‘sleep-disordered breathing’ diagnosis.

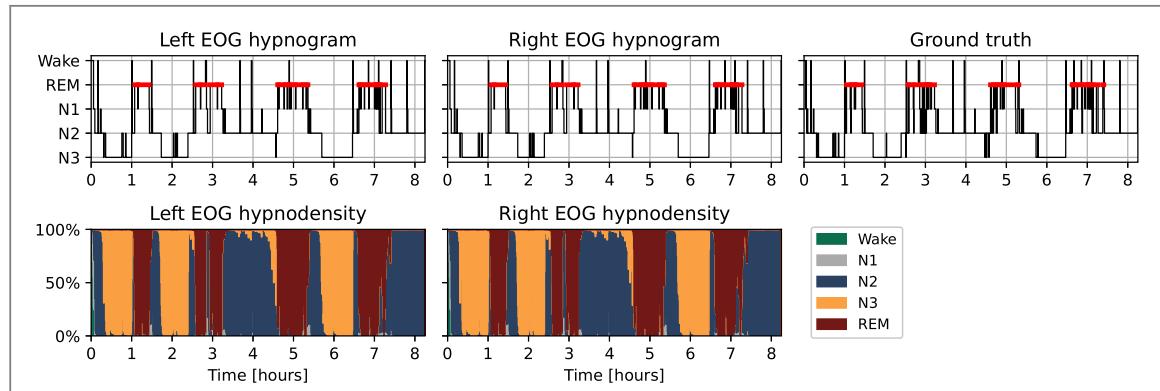


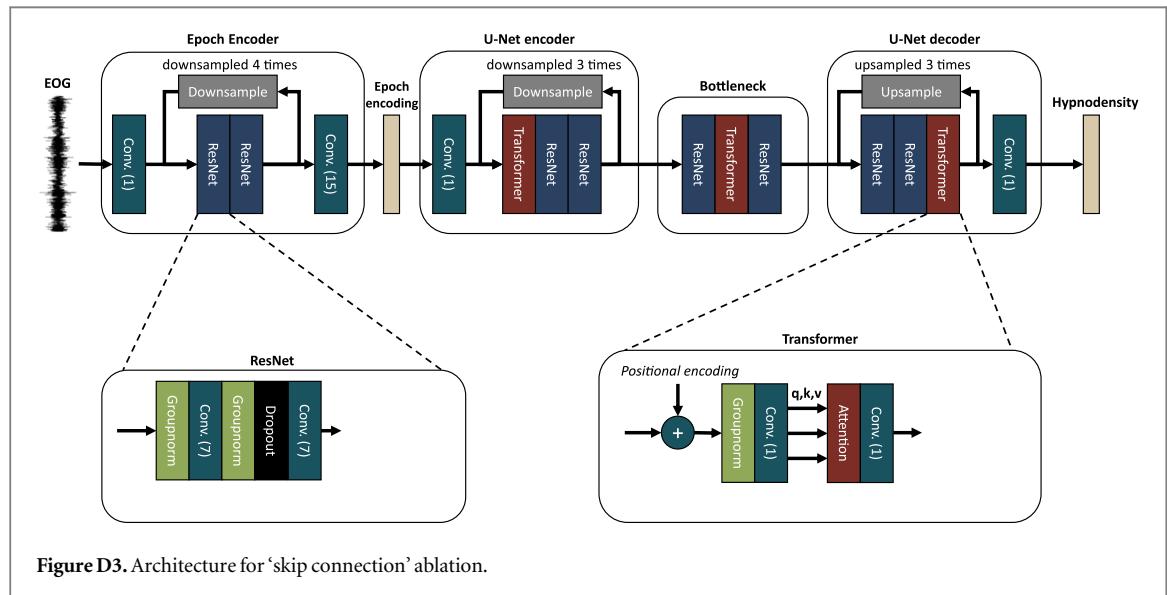
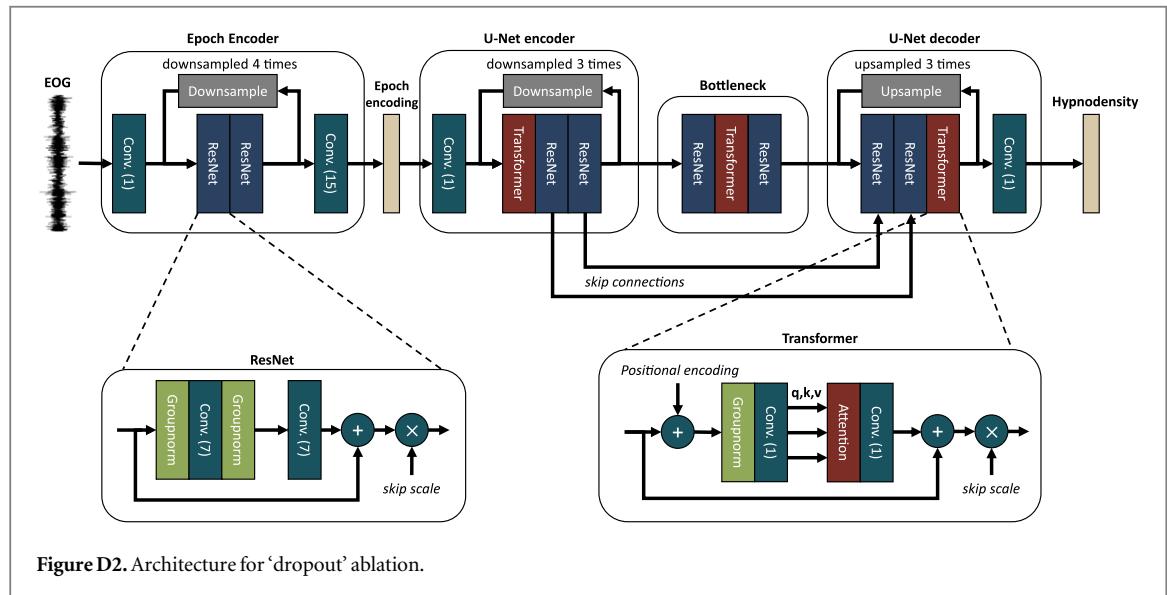
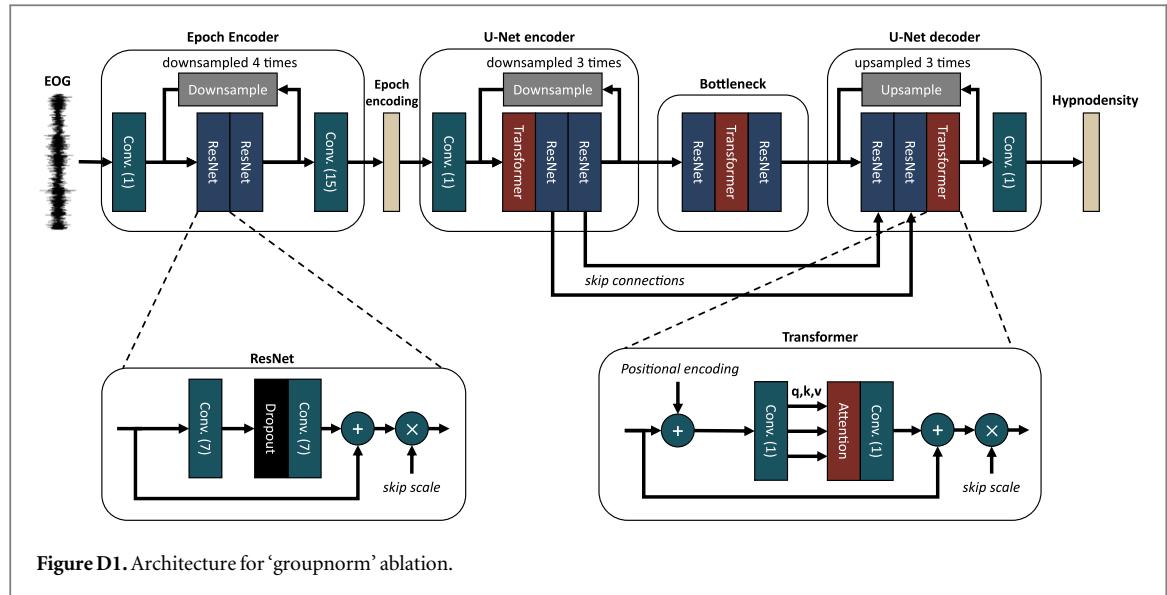
Figure C8. Qualitative results for a random recording from a subject with no primary sleep diagnosis.

Appendix D. Ablation experiment

A post-hoc ablation study was performed to evaluate some of the neural network layers used. Five different ablations were performed and the resulting networks were trained and evaluated in a similar way as the base network proposed in the manuscript. The ablations were as follows. Firstly, the group normalization layers were ablated, as shown in figure D1. Secondly, the dropout layers were ablated, as shown in figure D2. Thirdly, all the skip connection were ablated, as shown in figure D3. Fourthly, all transformer layers were ablated, as shown in figure D4. Lastly, we ablated the convolutions in the U-net by setting their kernel sizes to ‘1’. This effectively changes them to linear layers, without any ability to aggregate information between neighbouring epochs. This is shown in figure D5.

The resulting test set performance in terms of median and interquartile range is shown in table D1. We tested if the resulting metrics were significantly different from the base performance using Wilcoxon signed-rank tests, with a significance value of $p = 0.05$.

From table D1, it can be observed that the group normalization layers, dropout layers, and skip connections are essential for the network to have good sleep staging performance. As the performance metrics for these ablations are significantly lower than those of the base network. Additionally, we can see that the U-Net, which is responsible for aggregating information between neighbouring epochs, can retain its performance using either only convolutions or only transformers. Both are valid strategies for enabling the network to learn associations between epochs.



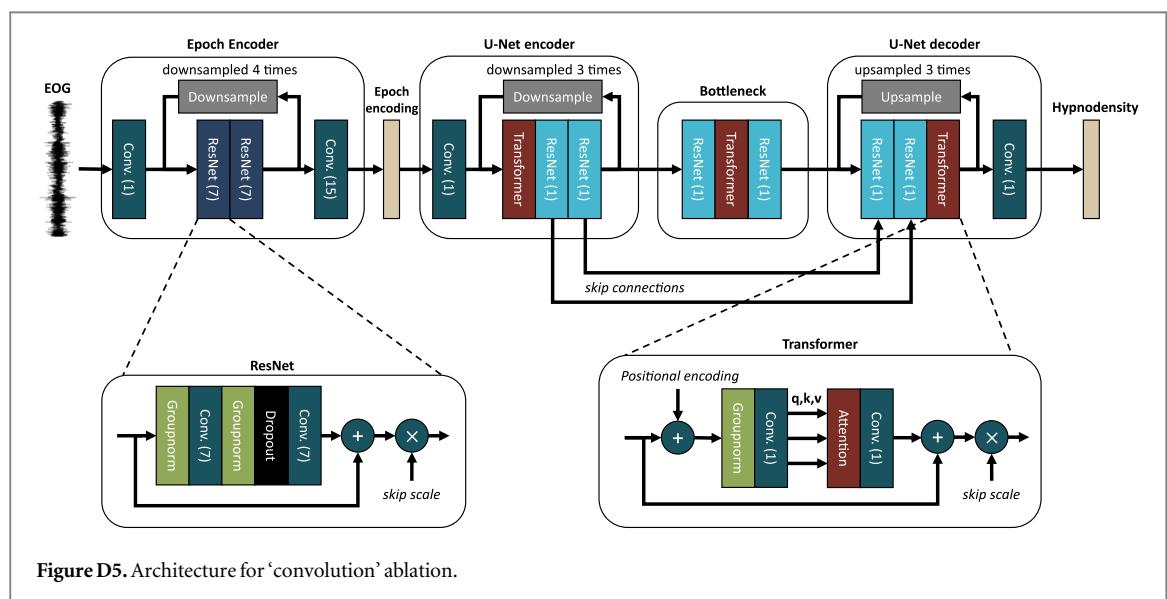
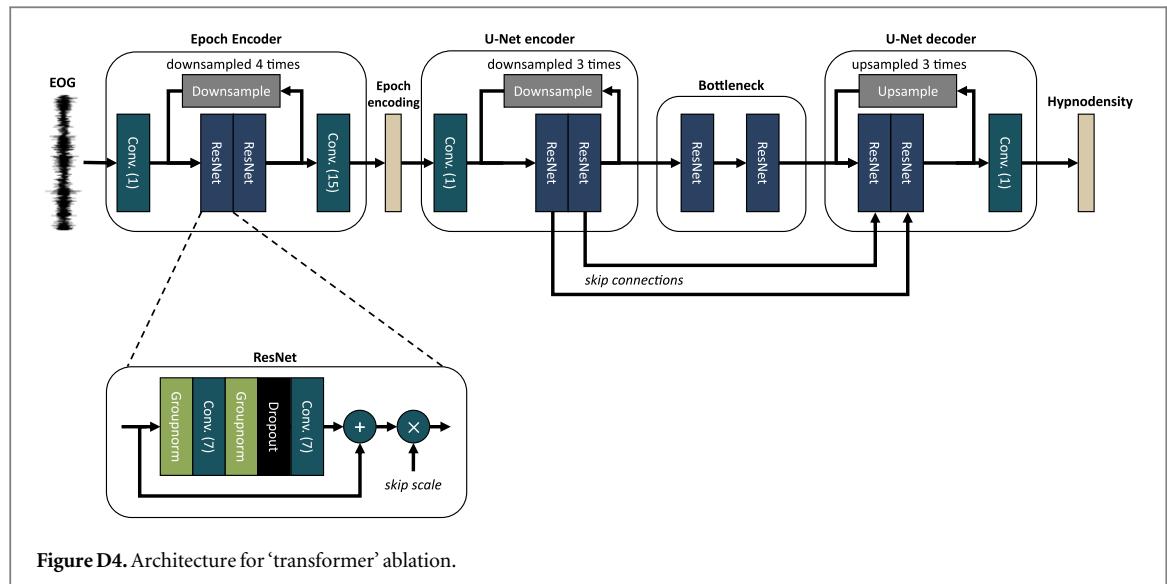


Table D1. Results for the different ablation experiments. We show the median/the interquartile range across the recordings. If an ablation result was significantly different from the base result, an asterisk is displayed.

Ablation	Left EOG		Right EOG	
	Accuracy	Kappa	Accuracy	Kappa
Base	85.0%/8.0%	0.781/0.107	85.2%/6.9%	0.796/0.103
groupnorm	82.2%/9.9%*	0.745/0.126*	83.5%/10.0%*	0.759/0.136*
dropout	81.6%/7.2%*	0.738/0.097*	83.0%/6.9%*	0.754/0.088*
skip connections	70.8%/10.3%*	0.570/0.151*	71.5%/9.7%*	0.576/0.137*
transformer	84.3%/7.0%	0.776/0.101	84.9%/6.5%	0.790/0.097
convolution	85.1%/8.0%	0.787/0.113	85.3%/7.0%	0.797/0.100

ORCID iDs

- Hans van Gorp  <https://orcid.org/0000-0003-4823-2874>
Merel M van Gilst  <https://orcid.org/0000-0003-2138-5686>
Sebastiaan Overeem  <https://orcid.org/0000-0002-6445-9836>
Pedro Fonseca  <https://orcid.org/0000-0003-2932-6402>
Ruud J G van Sloun  <https://orcid.org/0000-0003-2845-0495>

References

- American Academy of Sleep Medicine 2023 *International Classification of Sleep Disorders* 3rd edn (American Academy of Sleep Medicine)
Anderer P, Ross M, Cerny A, Vasko R, Shaw E and Fonseca P 2023 Overview of the hypnodensity approach to scoring sleep for polysomnography and home sleep testing *Front. Sleep* **2** 1163477
Bakker J P, Ross M, Cerny A, Vasko R, Shaw E, Kuna S, Magalang U J, Punjabi N M and Anderer P 2022 Scoring sleep with artificial intelligence enables quantification of sleep stage ambiguity: hypnodensity based on multiple expert scorers and auto-scoring *Sleep* **46** zsac154
Bakker J P et al 2021 Estimating sleep stages using cardiorespiratory signals: validation of a novel algorithm across a wide range of sleep-disordered breathing severity *J. Clin. Sleep Med.* **17** 1343–54
Bresch E, Großkathöfer U and Garcia-Molina G 2018 Recurrent deep neural networks for real-time sleep stage classification from single channel eeg *Front. Comput. Neurosci.* **12** 85
Ehlert I, Danker-Hopfe H, Höller L, Von Rickenbach P, Baumgart-Schmitt R and Herrmann W 1998 A comparison between eeg-recording and scoring by quasi version 1.0 and standard psg with visual scoring *Somnologie* **2** 104–16
Eldele E, Chen Z, Liu C, Wu M, Kwok C-K, Li X and Guan C 2021 An attention-based deep learning approach for sleep stage classification with single-channel eeg *IEEE Trans. Neural Syst. Rehabil. Eng.* **29** 809–18
Fan J, Sun C, Long M, Chen C and Chen W 2021 Eognet: a novel deep learning model for sleep stage classification based on single-channel eog signal *Front. Neurosci.* **15** 573194
Finan P H, Richards J M, Gamaldo C E, Han D, Leoutsakos J M, Salas R, Irwin M R and Smith M T 2016 Validation of a wireless, self-application, ambulatory electroencephalographic sleep monitoring device in healthy volunteers *J. Clin. Sleep Med.* **12** 1443–51
Fonseca P et al 2023 A computationally efficient algorithm for wearable sleep staging in clinical populations *Sci. Rep.* **13** 9182
Hendrycks D and Gimpel K 2016 Gaussian error linear units (gelus) arXiv:1606.08415
Hsieh T-H, Liu M-H, Kuo C-E, Wang Y-H and Liang S-F 2021 Home-use and real-time sleep-staging system based on eye masks and mobile devices with a deep learning model *J. Med. Biol. Eng.* **41** 659–68
Huijben I A M, Hermans L W A, Rossi A C, Overeem S, van Gilst M M and van Sloun R J G 2023 Interpretation and further development of the hypnodensity representation of sleep structure *Physiol. Meas.* **44** 015002
Imtiaz S 2021 A systematic review of sensing technologies for wearable sleep staging *Sensors* **21** 1562
Kapoor S and Narayanan A 2023 Leakage and the reproducibility crisis in machine-learning-based science *Patterns* **4** 1–13
Karras T, Aittala M, Aila T and Laine S 2022 Elucidating the design space of diffusion-based generative models *Adv. Neural Inf. Process. Syst.* **35** 26565–77
Kingma D P and Ba J 2015 Adam: a method for stochastic optimization *Int. Conf. on Learning Representations (ICLR)*
Kuo C-E, Liang S-F, Lee Y-C, Cherng F-Y, Lin W-C, Chen P-Y, Liu Y-C and Shaw F-Z 2014 An eog-based automatic sleep scoring system and its related application in sleep environmental control *Physiological Computing Systems: 1st Int. Conf., PhyCS 2014, Revised Selected Papers 1 (Lisbon, Portugal)* (Springer) pp 71–88
Lambert I and Peter-Derex L 2023 Spotlight on sleep stage classification based on eeg *Nat. Sci. Sleep* **15** 479–490
Liang S-F, Kuo C-E, Lee Y-C, Lin W-C, Liu Y-C, Chen P-Y, Cherng F-Y and Shaw F-Z 2015 Development of an eog-based automatic sleep-monitoring eye mask *IEEE Trans. Instrum. Meas.* **64** 2977–85
Nikkonen S et al 2023 Multicentre sleep-stage scoring agreement in the sleep revolution project *J. Sleep Res.* **33** e13956
Olesen A N, Christensen J A E, Sorensen H B D and Jennum P J 2016 A noise-assisted data analysis method for automatic EOG-based sleep stage classification using ensemble learning *2016 38th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)* pp 3769–72
Perslev M, Darkner S, Kempfner L, Nikolic M, Jennum P J and Igel C 2021 U-sleep: resilient high-frequency sleep staging *Npj Digit. Med.* **4** 72
Phan H and Mikkelsen K 2022 Automatic sleep staging of EEG signals: recent development, challenges, and future directions *Physiol. Meas.* **43** 04TR01
Phan H, Mikkelsen K B, Chen O, Koch P, Mertins A and de Vos M 2022 Sleeptransformer: automatic sleep staging with interpretability and uncertainty quantification *IEEE Trans. Biomed. Eng.* **69** 2456–67
Qu W, Wang Z, Hong H, Chi Z, Feng D D, Grunstein R and Gordon C 2020 A residual based attention model for EEG based sleep staging *IEEE J. Biomed. Health Inform.* **24** 2833–43
Rahman M M, Bhuiyan M I H and Hassan A R 2018 Sleep stage classification using single-channel eog *Comput. Biol. Med.* **102** 211–20
Ronneberger O, Fischer P and Brox T 2015 U-net: convolutional networks for biomedical image segmentation *Medical Image Computing and Computer-assisted Intervention—MICCAI 2015* (Springer) pp 234–41
Rosenberg R S and Van Hout S 2013 The american academy of inter-scorer reliability program: sleep stage scoring *J. Clin. Sleep Med.* **9** 81–7
Schweitzer M, Mohammad A, Binder R, Steinberg R, Schreiber W H and Weeß H-G 2004 Biosomnia-validity of a mobile system to detect sleep and sleep quality *Somnologie* **8** 131–8
Song Y, Sohl-Dickstein J, Kingma D P, Kumar A, Ermon S and Poole B 2021 Score-based generative modeling through stochastic differential equations *Int. Conf. on Learning Representations* (<https://doi.org/10.48550/arXiv.2011.13456>)
Stephansen J B et al 2018 Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy *Nat. Commun.* **9** 5229
Tompson J, Goroshin R, Jain A, LeCun Y and Bregler C 2015 Efficient object localization using convolutional networks *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* pp 648–56
Troester M M, Quan S F and Berry R B 2023 *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications* (American Academy of Sleep Medicine) Version 3

- van der Aar J F, van den Ende D A, Fonseca P, van Meulen F B and van Gilst M M 2024 Deep transfer learning for automated single-lead eeg sleep staging with channel and population mismatches *Front. Physiol.* **14** 1287342
- van Gilst M M et al 2019 Protocol of the somnia project: an observational study to create a neurophysiological database for advanced clinical sleep monitoring *BMJ Open* **9** e030996
- van Gorp H, Huijben I A M, Fonseca P, van Sloun R J G, Overeem S and van Gilst M M 2022 Certainty about uncertainty in sleep staging: a theoretical framework *Sleep* **45** zsac134
- van Gorp H, van Gilst M M, Fonseca P, Overeem S and van Sloun R J G 2023 Modeling the impact of inter-rater disagreement on sleep statistics using deep generative learning *IEEE J. Biomed. Health Inform.* **27** 5599–609
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I 2017 Attention is all you need *Adv. Neural Inf. Process. Syst.* **30**
- Virkkala J, Hasan J, Värrí A, Himanen S-L and Müller K 2007 Automatic sleep stage classification using two-channel electro-oculography *J. Neurosci. Methods* **166** 109–15
- Virkkala J, Velin R, Himanen S-L, Varri A, Muller K and Hasan J 2008 Automatic sleep stage classification using two facial electrodes 2008 *30th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* pp 1643–6
- Wu Y and He K 2018 Group normalization *Proc. of the European Conference on Computer Vision (ECCV)* pp 3–19
- Zhai H, Yan Y, He S, Zhao P and Zhang B 2023 Evaluation of the accuracy of contactless consumer sleep-tracking devices application in human experiment: a systematic review and meta-analysis *Sensors* **23** 4842
- Zhu H, Fu C, Shu F, Yu H, Chen C and Chen W 2023 The effect of coupled electroencephalography signals in electrooculography signals on sleep staging based on deep learning methods *Bioengineering* **10** 573
- Zhu T, Luo W and Yu F 2020 Convolution-and attention-based neural network for automated sleep stage classification *Int. J. Environ. Res. Public Health* **17** 4152