












RESEARCH ARTICLE

Retrospective validation of automatic sleep analysis with grey areas model for human-in-the-loop scoring approach

Matias Rusanen^{1,2,3}  | Gabriel Jouan⁴ | Riku Huttunen^{1,2}  |
Sami Nikkonen^{1,2}  | Sigríður Sigurðardóttir⁴  | Juha Töyräs^{1,5,6}  |
Brett Duce⁷  | Sami Myllymaa^{1,2}  | Erna Sif Arnardóttir^{4,8}  |
Timo Leppänen^{1,2,6}  | Anna Sigridur Islind⁹ | Samu Kainulainen^{1,2}  |
Henri Korkalainen^{1,2} 

¹Department of Technical Physics, University of Eastern Finland, Kuopio, Finland

²Diagnostic Imaging Centre, Kuopio University Hospital, Kuopio, Finland

³HP2 Laboratory, INSERM U1300, Grenoble Alpes University, Grenoble Alpes University Hospital, Grenoble, France

⁴Reykjavik University Sleep Institute, School of Technology, Reykjavik University, Reykjavik, Iceland

⁵Science Service Centre, Kuopio University Hospital, Kuopio, Finland

⁶School of Electrical Engineering and Computer Science, The University of Queensland, Brisbane, Queensland, Australia

⁷Princess Alexandra Hospital, Sleep Disorders Centre, Princess Alexandra Hospital, Brisbane, Queensland, Australia

⁸Landspítali – The National University Hospital of Iceland, Reykjavik, Iceland

⁹Department of Computer Science, Reykjavik University, Reykjavik, Iceland

Correspondence

Matias Rusanen, HP2 Laboratory, INSERM U1300, Grenoble Alpes University, Grenoble Alpes University Hospital, Bd de la Chantourne, 38700 La Tronche, Grenoble, France.

Email: matias.rusanen@uef.fi

Funding information

European Union's Horizon 2020 research and innovation programme, Grant/Award Number: 965417; the State Research Funding for University-Level Health Research, Kuopio University Hospital, Wellbeing Service County of North Savo, Grant/Award Numbers: 5041789, 5041794, 5041797, 504180, 5041803, 5041807, 5041809; Finnish Cultural Foundation; Eemil Aaltonen Foundation; The Research Foundation of the Pulmonary Diseases; Foundation of the Finnish Anti-Tuberculosis Association; Tampereen Tuberkuloosisäätiö; NordForsk, Grant/Award Number: 90458; Business Finland, Grant/Award Number: 5133/31/2018; Agence Nationale de la Recherche, Grant/Award Numbers: ANR-15-IDEX-02, ANR-19-P3IA-0003; Sigrid Jusélius Foundation

Summary

State-of-the-art automatic sleep staging methods have demonstrated comparable reliability and superior time efficiency to manual sleep staging. However, fully automatic black-box solutions are difficult to adapt into clinical workflow due to the lack of transparency in decision-making processes. Transparency would be crucial for interaction between automatic methods and the work of sleep experts, i.e., in human-in-the-loop applications. To address these challenges, we propose an automatic sleep staging model (aSAGA) that effectively utilises both electroencephalography and electro-oculography channels while incorporating transparency of uncertainty in the decision-making process. We validated the model through extensive retrospective testing using a range of datasets, including open-access, clinical, and research-driven sources. Our channel-wise ensemble model, trained on both electroencephalography and electro-oculography signals, demonstrated robustness and the ability to generalise across various types of sleep recordings, including novel self-applied home polysomnography. Additionally, we compared model uncertainty with human uncertainty in sleep staging and studied various uncertainty mapping metrics to identify ambiguous regions, or “grey areas”, that may require manual re-evaluation. The validation of this grey area concept revealed its potential to enhance sleep staging accuracy and to

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *Journal of Sleep Research* published by John Wiley & Sons Ltd on behalf of European Sleep Research Society.

highlight regions in the recordings where sleep experts may struggle to reach a consensus. In conclusion, this study provides a technical basis and understanding of automatic sleep staging uncertainty. Our approach has the potential to improve the integration of automatic sleep staging into clinical practice; however, further studies are needed to test the model prospectively in real-world clinical settings and human-in-the-loop scoring applications.

KEYWORDS

Convolutional neural networks, Human-in-the-loop, Sleep staging, Uncertainty mapping

1 | INTRODUCTION

Manual sleep staging constitutes a fundamental analysis conducted in sleep clinics worldwide (Carskadon and Dement, 2011). This process involves the examination of specific patterns in electroencephalography (EEG), electro-oculography (EOG), and chin electromyography (chin EMG) within 30-second segments i.e., epochs of sleep recordings. Each epoch is subsequently categorised into one of five sleep stages: Wake, non-rapid eye movement sleep stages 1–3 (N1, N2, or N3), or rapid eye movement (REM) sleep. The established scoring rules are applied universally to guide this task (Berry et al., 2020). Despite the diligent efforts invested in manual sleep staging, the presence of ambiguous regions persists, characterised by instances in which the expert scorers are unable to reach a consensus and unintentionally introduce errors (Lee et al., 2021; Van Gorp et al., 2022; Zhang et al., 2015).

Uncertain areas with discrepancies between scorers, i.e. grey areas, of sleep staging exist for multiple reasons. Firstly, the scoring rules employed for sleep staging have faced criticism for insufficient scientific foundation and the potential for interpretive variations (Himanen & Hasan, 2000; Jorgensen et al., 2020; Parrino et al., 2009). Additionally, scoring practices differ between sleep centres, as each institution usually adopts its own set of national or internal practices in addition to the traditionally used American Academy of Sleep Medicine scoring manual (Arnardottir et al., 2016; Norman et al., 2000). Moreover, the simplification of scoring sleep stages into 30-second epochs can lead to confusion, particularly during transitions between stages (Schulz, 2008). Furthermore, sleep recordings are susceptible to artefacts, which can impact the accuracy of sleep stage analysis. In some instances, these artefacts may mimic characteristic patterns observed in EEG, EOG, or EMG traces specific to sleep stages (Kalevo et al., 2019). While skilled sleep technologists can perform sleep staging relatively accurately, even under such challenging circumstances (Nikkonen et al., 2023), the manual sleep staging necessitates substantial effort and an extensive allocation of work hours (Fischer et al., 2012).

To alleviate the challenges associated with manual sleep staging, numerous automatic sleep staging solutions have been devised (Fiorillo et al., 2019; Phan & Mikkelsen, 2022). These solutions have already been integrated into diverse polysomnographic (PSG) analysis software and are sometimes employed for preliminary

scoring, which may speed up and increase the reliability of the manual sleep staging process (Id & Rijsman, 2022). Machine learning (ML)-based methods, representing the cutting-edge in this domain, demonstrate comparable performance to expert scorers in sleep staging, particularly in specific datasets on which the models were trained (Alvarez-Estevéz & Rijsman, 2021; Fiorillo et al., 2019; Fiorillo et al., 2021; Huttunen et al., 2022; Olesen et al., 2021; Perslev et al., 2021; Phan & Mikkelsen, 2022; Phan, Mikkelsen, et al., 2021; Stephansen et al., 2018). Further studies are needed to test the generalisability of the methods across different datasets and recordings conducted on various medical sensor setups. Moreover, based on current clinical recommendations, automatic sleep analysis tools cannot be relied upon solely, and their results necessitate thorough manual review (Goldstein et al., 2020). Only a few solutions have been introduced for highlighting the uncertain regions of automatic sleep analysis (Bakker et al., 2022; Fiorillo et al., 2021; Hong et al., 2021; Kang et al., 2021; Phan, Mikkelsen, et al., 2021), and the concept warrants further validation. In particular, it remains unclear whether these uncertain regions correspond to areas where confusions between manual scorers emerge, and therefore, a careful manual review of these areas could increase the quality of sleep staging. Moreover, conceptualisation and discussion regarding the interaction between automatic and manual scoring remains limited (Goldstein et al., 2020; Van Gorp et al., 2022). This discussion would be crucial for the consideration of human-in-the-loop approaches in sleep analysis where humans play a crucial role in guiding and validating the ML system's decisions, ensuring that they are clinically reasonable and aligned with human based scoring (Monarch, 2021).

This paper presents a novel conceptualisation of the interaction between manual and automatic sleep staging (Figure 1), where we propose a model called aSAGA (Automatic Sleep Analysis with Grey Areas) to realise the human-in-the-loop concept. This model integrates a state-of-the-art automatic sleep staging method that exhibits generalisability across different recording types (Rusanen et al., 2023), along with transparency of decision-making through uncertainty mapping. The uncertainty mapping is based on the hypodensities (Anderer et al., 2023; Stephansen et al., 2018), representing the probabilities of being in each sleep stage for each epoch. We hypothesise that the hypnodensities can be used to highlight areas of the recording where detailed manual revision can lead to an

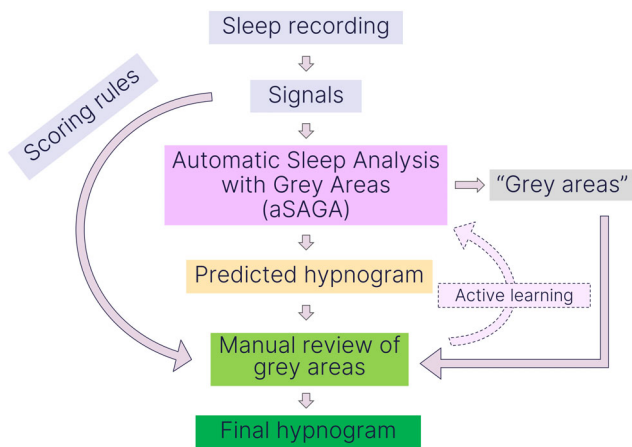


FIGURE 1 The proposed human-in-the-loop sleep staging concept for the interaction of manual and automatic methods. In this approach, the aSAGA model gives a prediction of the hypnogram and highlights uncertain regions, the grey areas, that need manual re-evaluation. In this paper, we aimed to evaluate the aSAGA model with retrospective data. In future work, active learning can be incorporated as a feedback loop to the proposed approach.

increased quality of the sleep staging analysis. The aim is to assess whether the grey areas identified by the model correspond to regions where discrepancies arise between automatic and manual scoring as well as between sleep technologists. Furthermore, we aim to compare the effectiveness of hypnodensity-based uncertainty metrics introduced in the literature and to discuss the possible strategies to define the grey areas.

2 | METHODS

2.1 | Datasets

2.1.1 | Clinical PSG recordings

The clinical in-laboratory PSG recordings used in the training of the automatic sleep staging model are described in detail in a previous work (Huttunen et al., 2022). These recordings form a heterogeneous clinical dataset of suspected obstructive sleep apnea (OSA) patients collected at the Princess Alexandra Hospital (Brisbane, Australia) during the years 2015–2017 (Table S1–S3). These recordings were conducted at the hospital's sleep laboratory with Compumedic (Abbotsford, Australia) Graef 4K PSG system with practices in line with the AASM guidelines (Berry et al., 2020). The Institutional Human Research Ethics Committee of the Princess Alexandra Hospital reviewed and approved a research proposal for the use of these data (HREC/16/QPAH/021 and LNR/2019/QMS/54313). We excluded recordings with less than 1 h of total sleep time and divided the data randomly into training ($n = 710$), validation ($n = 78$), and test sets ($n = 88$). Demographic information of the data is presented in Data S1.

2.1.2 | Dream open data

The Dream Open Data (DOD) is a documented and freely available open-access PSG dataset (Guillot et al., 2020). The data consist of 25 healthy participants (DOD-H) and 56 participants with OSA (DOD-O). The DOD-H data were collected at the French Armed Forces Biomedical Research Institute's (IRBA) Fatigue and Vigilance Unit using Siesta PSG devices (Compumedics, Victoria, Australia). The data included 13 EEG derivations (C3-M2, F4-M1, F3-F4, F3-M2, F4-M2, F4-O2, F3-O1, FP1-F3, FP1-M2, FP1-O1, FP2-F4, FP2-M1, and FP2-O2) as well as derivations for right and left EOG. The DOD-O data were collected at the Stanford Sleep Medicine Centre using a Somno HD PSG device (Somnomedics). These data included eight EEG derivations (C3-M2, C4-M1, F3-F4, F3-M2, F4-O2, F3-O1, FP1-M2, and FP1-F3) along with the left and right EOG. In both datasets, each record was scored by five expert sleep technologists (>5 years of clinical/scoring experience) from three different sleep centres. Further details on DOD data can be found in the original article (Guillot et al., 2020). The DOD data were used as a separate test set in this study, i.e., not in the training of the model. An end-to-end example of how to use these data for training an automatic sleep staging model is available at https://github.com/UEF-SmartSleepLab/sleeplab-format/tree/main/examples/dod_sleep_staging.

2.1.3 | Self-applied PSG data

The self-applied PSG recordings were collected at Akershus University Hospital in Norway and the Reykjavik University Sleep Institute in Iceland as a part of the Sleep Revolution Horizon 2020 project (ethical permissions 21-070 (Iceland) and 2017/2161 (Norway)) (Arnardottir et al., 2022; Sveinbjarnarson et al., 2023). The goal was to get a balanced cohort of healthy subjects and different levels of sleep apnea severity. A total of 50 participants received the self-applied PSG (Self-Applied Somnography (SAS), Nox Medical, Reykjavik, Iceland) equipment to take home for one night, along with written and video instructions on the setup. The SAS data comprised four EEG (AF4-E3E4, AF3-E3E4, AF7-E3E4, and AF8-E3E4) derivations and four EOG (E1-E4, E2-E3, E2-AFZ, and E3-AFZ) derivations that were used in the analysis. Demographic information of the data is presented in the supplementary material (Table S4).

Ten expert sleep technologists from seven different sleep centres scored each of the 50 recordings. Each scorer used Noxturnal Research version 6.1.0 (Nox Medical, Reykjavik, Iceland) software to first run an integrated automatic analysis and then inspected and scored/corrected the scoring following the AASM 2.6 scoring manual (Berry et al., 2018). This scoring protocol was used to correspond to common clinical practice. In the end, we obtained 10 scorings for each recording. A majority score for sleep stages was generated by selecting the most-scored sleep stage for each epoch for the 10 scorers (Nikkonen et al., 2023). A tiebreaker method was used to choose the sleep stage with higher priority in the following order: Wake, N1, N2,

N3, and REM. Data included 3.5% of tied epochs. As a specialty, the scorers were able to mark their scoring as uncertain, if they were unsure about the sleep stage but still labelled with the most likely sleep stage, e.g. uncertain N1. Following this, 3936 epochs (10% of analysed data) were scored with uncertainty by at least one of the 10 scorers. The self-applied PSG recordings were also used as completely separate test data in the validation analysis.

2.2 | Automatic sleep staging

The presented aSAGA model utilises a fully convolutional neural network that has shown state-of-the-art performance and good generalisability between different datasets in previous studies (Huttunen et al., 2022; Rusanen, Huttunen, et al., 2023). The architecture of the model was inspired by previous research conducted by Perslev et al. (Perslev et al., 2019; Perslev et al., 2021) and by Ronneberger et al. (Ronneberger et al., 2015). The model architecture used in aSAGA is the same as in previous studies and described in detail in Huttunen et al. (2022). The codes for setting up the model are available at <https://github.com/rikuhuttunen/psg-simultscoring-models>.

In the aSAGA, we used a single-channel model which was first trained on EEG (C4-M1) and then fine-tuned with an EOG (E1-M2) channel using the clinical PSG recordings. This was done to have generalisability between EEG and EOG channels and to increase the compliance of the model for frontal EEG and EOG channels of self-applicable sensor setups. In the training, we utilised a recently published unified data format, called the *sleeplab format* (Huttunen et al., 2023), to have better reproducibility. The training of the model was conducted over 100 training passes. During each training epoch, each batch input to the model was created by sampling a 1 h segment of recording from eight different subjects. The start times of the segments were randomly sampled from each recording. Model weights were optimised using the AdamW algorithm (Loshchilov & Hutter, 2019). The categorical cross entropy was used as the loss function. Furthermore, instead of a constant learning rate, we used a learning rate scheduler (Smith, 2016). The learning rate division factor was 2.0 after every 20 epochs and the peak learning rate was set to 0.01. After initial training, a constant learning rate of 0.001 was used in the fine-tuning. Otherwise, the finetuning was performed similarly to the initial training. The best model was returned after training according to validation accuracy. Training parameters and model parameters used in the training and finetuning are available in detail at <https://github.com/matias-olavi/aSAGA>.

All EEG and EOG signals are pre-processed similarly to training data in the aSAGA, wherein they are bandpass filtered (0.3–32 Hz), re-sampled to 64 Hz, and normalised with an interquartile range. The normalisation with an interquartile range was found to be effective in a previous study with cross-domain predictions where peak-to-peak amplitude of the signal varied significantly (Rusanen,

Huttunen, et al., 2023). Moreover, a previous study reported that EEG signals recorded in self-applied PSGs have significantly lower peak-to-peak amplitudes compared with standard EEG derivations (Rusanen, Korkalainen, et al., 2023). Therefore, normalising each recording is a critical step in the preprocessing, although information about absolute amplitudes will be lost. Standard preprocessing functions of the *sleeplab format* extractor were used (available in GitHub). All signals were run through the single-channel model to create an ensemble prediction of the hypnodensities i.e., the probability of each sleep stage for each 30 s epoch. Following this, the hypnodensities were averaged over the channels and the sleep stage with the highest probability was included in the final hypnogram.

2.3 | Uncertainty mapping – the grey areas

Grey areas were derived from hypnodensities of manual and automatic sleep staging using uncertainty mapping. In the aSAGA, averaged hypnodensities from the ensemble predictions were used as the probability of the sleep stage in the studied epoch. In the manual sleep staging results obtained from 10 scorers, the probability for each sleep stage can be defined as the number of scorers of the overall 10 who scored the respective sleep stage. Let p_i , where $i = 1, 2, \dots, 5$ correspond to each of the five sleep stages, denote the probabilities or the hypnodensities for each 30 s epoch. We can estimate the uncertainty of the model from the hypnodensities with different metrics described in the literature (Monarch, 2021; Kang et al., 2021; Phan, Mikkelsen, et al., 2021). In the present paper, we compared the following five metrics:

1. Least confidence sampling

$$U_L = (1 - \max(p_i)) * \frac{n}{n-1},$$

where $l = \{\text{all sleep stages}\}$ and n is the number of sleep stages.

2. The margin of confidence sampling (between the two most confident labels)

$$U_M = 1 - (\max(p_i) - \max(p_j)),$$

where $l = \{\text{all sleep stages}\}$, and $J = \{\text{all sleep stages except the most probable}\}$.

3. The ratio of confidence (between the two most confident labels)

$$U_R = \max(p_i) / \max(p_j),$$

where $l = \{\text{all sleep stages}\}$, and $J = \{\text{all sleep stages except the most probable}\}$.

4. Coefficient of unalikeability

$$U_U = 1 - \sum_{i=1}^n p_i^2$$

5. Classification entropy

$$U_E = - \frac{\sum_{i=1}^n p_i \log_2 p_i}{\log_2 n},$$

where $0 * \log_2(0) = 0$ was assumed.

2.4 | Validation analysis

To validate the single-channel automatic sleep staging model, we first used the test set of clinical PSG recordings collected at the Princess Alexandra Hospital to show the sleep staging performance from the single EOG and EEG channels. Typical performance metrics such as accuracy, Cohen kappa, and F1-scores were computed in an epoch-wise manner to measure the overall sleep staging performance. Sleep stage-specific classification performance was evaluated by computing precision, recall, and F1-score for each class separately. Furthermore, we visualised the scorings between the automatic and manual sleep staging with a confusion matrix.

To validate the generalisability of the aSAGA, we evaluated the sleep staging performance with the Dreem Open Data. The performance of the automatic method was tested with healthy subjects and patients with OSA separately against the consensus of five scorers. In this validation, we utilised all EEG and EOG channels available in the dataset to create the ensemble prediction described in Section 2.3. Finally, the automatic sleep staging performance of the aSAGA was tested with the self-applied PSGs against the consensus of the 10 scorers. No finetuning of the model was performed i.e., the performance was evaluated in a direct transfer manner with all datasets (sections 2.1.2 and 2.1.3).

Following the sleep staging performance validation, we tested the grey area concept of the aSAGA with two different analyses. First, we tested if excluding the most uncertain epochs of the prediction increased the automatic sleep staging performance when evaluated against the consensus manual scoring. This was done to evaluate if the grey areas corresponded to the areas of the recording where the confusion between manual and automatic sleep staging occurs. This analysis was performed using the uncertainty metrics proposed in Section 2.3 and varying the percentage of epochs considered as grey areas from 1% to 95% of the whole dataset based on the rank order of the uncertainty values. This validation was done using the Dreem Open Data, and the validation protocol is available at <https://github.com/matias-olavi/aSAGA>.

Second, we tested whether the grey areas of the aSAGA capture those areas of the self-applied PSGs where the 10 scorers had the most confusion between each other. The testing was performed using the coefficient of unalikeability, which is suitable for capturing the variability of the manual scorings. A threshold of 0.6 was used for the uncertainty metric to separate grey areas from non-grey areas with both the

10 scorers and the model hypnodensity series. Binary classification between the grey and non-grey areas derived from manual and automatic sleep staging was computed. Confusions in the grey area classification between these two methods were visualised with a confusion matrix. Moreover, the results of this analysis were split into two parts, the epochs that were marked as uncertain by any of the 10 scorers, and the remaining data (not marked as uncertain by any of the scorers). The data consisting of epochs manually marked as uncertain represent grey areas that are known to exist in the sleep stage, whereas grey areas of data consisting of manually scored epochs that were not marked as uncertain represent more the uncertainties unknown for scorers.

3 | RESULTS

3.1 | Sleep staging performance

3.1.1 | Clinical PSG data-test set ($n = 88$)

The single-channel model performed well against manual sleep staging and with similar overall metrics (accuracy = 81% and $\kappa = 0.74$) for both EEG and EOG inputs (Table 1). Sleep stage-specific metrics showed that REM sleep was better detected from the EOG input than from the EEG input, with F1-scores of 0.88 and 0.83, respectively. In contrast, Wake was detected with similar consistency from EEG compared with EOG as an input, with F1-scores of 0.90 and 0.89, respectively. Most of the confusions between automatic and manual sleep staging occurred in the classification of non-REM sleep stages of N1, N2, and N3 (Figure 2).

3.1.2 | Dreem open data ($n = 25 + 56$)

The aSAGA model showed high generalisability of the sleep staging performance to separate open-access datasets consisting of recordings of healthy subjects (DOD-H, $n = 25$) and subjects with OSA (DOD-O, $n = 56$). With DOD-H data, the aSAGA model reached 80% accuracy, and with DOD-O data the accuracy was 82% without any transfer learning approaches. Sleep stages Wake, N2, and REM were detected with a good agreement (F1-scores ≥ 0.82) against the majority scoring (Table 2). Slightly lower performance was reached for the detection of N3 sleep, with F1-scores being 0.78 and 0.79 for DOD-H and DOD-O data, respectively. However, N1 sleep was detected with moderate agreement in both datasets.

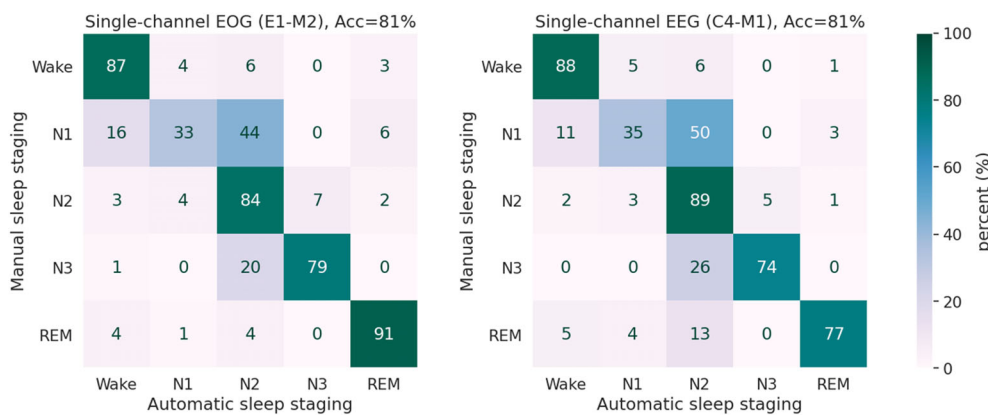
3.1.3 | Self-applied PSGs ($n = 50$)

Manual sleep staging of the self-applied PSGs showed high variation in the agreement of the 10 individual scorings against the majority score with accuracy ranging from 69% to 91% between the 10 scorers (Figure 3). On bar with these manual scorings against the majority score, the aSAGA reached the accuracy of 83% against the majority scoring. High F1-scores (≥ 0.86) were reached for detecting N2, N3, and REM sleep (Table 3, Figure 4). The F1-score for detecting Wake was lower,

TABLE 1 Performance metrics of the automatic sleep staging model in a test set of the clinical PSG data when single-channel inputs were used.

Recording types	Subjects	Signal	Sleep stage	Performance metrics			
				Precision	Recall	F1-score	Number of epochs
Type I PSG	Suspected OSA ($n = 88$)	EOG (E1-M2)	Wake	0.91	0.87	0.89	26,881
			N1	0.51	0.33	0.40	6708
			N2	0.75	0.84	0.80	25,370
			N3	0.80	0.79	0.80	9,695
			REM	0.85	0.91	0.88	9567
			Macro average	0.76	0.75	0.75	78,221
			κ	0.74			78,221
		Accuracy	81%			78,221	
		EEG (C4-M1)	Wake	0.93	0.88	0.90	26,881
			N1	0.50	0.35	0.41	6708
			N2	0.72	0.89	0.80	25,370
			N3	0.85	0.74	0.79	9695
			REM	0.91	0.77	0.83	9567
			Macro average	0.78	0.73	0.75	78,221
κ	0.74				78,221		
Accuracy	81%			78,221			

Note: Metrics were computed against the manual sleep staging. The macro average is the unweighted average over the sleep stage-specific metrics. Abbreviations: epoch, 30-second signal segment; κ , Cohen's kappa; EEG, electroencephalography; EOG, electro-oculography; OSA, obstructive sleep apnea; PSG, polysomnography; REM, rapid eye movement.

**FIGURE 2** Confusion matrixes between manual sleep staging and the single-channel outputs of automatic sleep staging with the aSAGA model. Performance was computed on the test set of the clinical polysomnographic recordings of suspected OSA patients ($n = 88$). Acc, Accuracy; aSAGA, Automatic Sleep Analysis with Grey Area; EEG, electroencephalography; EOG, electro-oculography.

however. This was due to a precision of 0.63, although a high recall of 0.85 was reached. Similarly, as in the manual scorings, stage N1 was detected with only a fair agreement (F1-score = 0.30), and mostly confused with Wake and N2 stages (Figure 4).

3.2 | Grey area validity

3.2.1 | Increase in sleep staging performance

Automatic sleep staging performance increased effectively as a function of the amount of excluded data i.e., the grey areas, based on all of the compared uncertainty metrics in the DOD-O data (Figure 5). Differences in increased sleep staging performance were negligible

between the different uncertainty metrics on less than 10% of excluded data (Figure 5). However, classification entropy (U_E) showed less increase in sleep staging performance between 10% and 75% of excluded data, than the other metrics. Similarly, the coefficient of unlikelihood (U_U) increased the sleep staging performance slightly less between 10% and 30% of excluded data than the more direct measures of hypnodensities (U_L , U_M , and U_R).

Furthermore, confusions in sleep staging between manual and automatic scoring were captured effectively into grey areas (Figure 6). For example, when considering 40% of the data as grey areas, over 80% of the confusion between manual and automatic sleep staging was captured with U_L , U_M , U_R , and U_U metrics. However, entropy-based metrics U_U and U_E showed slightly worse overall performance than the more direct measures of hypnodensities U_L , U_M , and U_R .

TABLE 2 Performance metrics for the automatic sleep staging in Dream Open Data when the aSAGA model was used.

Recording type	Subjects	Signals	Sleep stage	Performance metrics			Number of epochs
				Precision	Recall	F1-score	
Type I PSG	Healthy (n = 25)	All EEG + EOG channels	Wake	0.87	0.77	0.82	3037
			N1	0.71	0.29	0.41	1505
			N2	0.83	0.83	0.83	11,879
			N3	0.65	0.97	0.78	3514
			REM	0.87	0.78	0.82	4727
			Macroaverage	0.79	0.73	0.73	24,662
			κ	0.71			24,662
	Accuracy		80%			24,662	
	OSA (n = 56)		Wake	0.90	0.79	0.84	10,660
			N1	0.50	0.42	0.46	2898
			N2	0.85	0.86	0.86	26,650
			N3	0.71	0.89	0.79	5683
			REM	0.85	0.82	0.83	8306
			Macro average	0.76	0.76	0.76	54,197
κ		0.74			54,197		
Accuracy	82%			54,197			

Note: Metrics were computed against the majority scores of manual sleep staging. Macro average is the unweighted average over the sleep stage-specific metrics.

Abbreviations: aSAGA, Automatic Sleep Analysis with Grey Areas; epoch, 30-second signal segment; κ, Cohen's kappa; EEG, electroencephalography; EOG, electro-oculography; OSA, obstructive sleep apnea; PSG, polysomnography; REM, rapid eye movement.

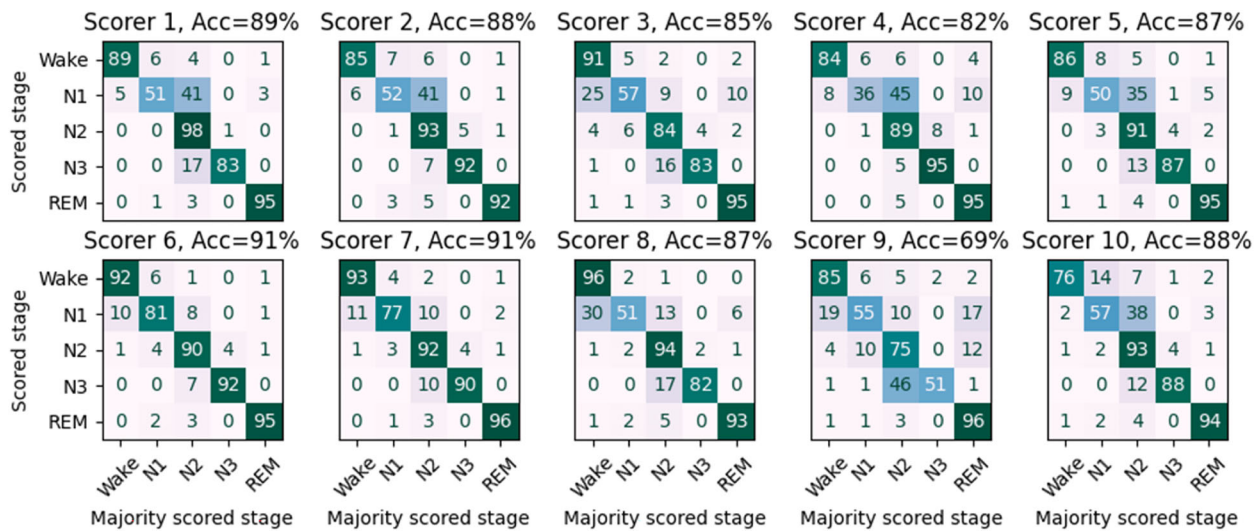


FIGURE 3 Confusions between the 10 individual scorings and the majority score derived from manual sleep staging in self-applied polysomnographic recordings (n = 50). Epochs manually marked as uncertain were excluded from the analysis. Acc, accuracy; REM, rapid eye movement.

3.2.2 | Agreement with manual scoring confusions in self-applied PSG

An example of the hypnodensity series and grey areas derived from manual and automatic sleep staging of one recording is shown in Figure 7. Defining the grey areas similarly from the 10 scorers and the

automatic model's hypnodensity (with threshold $U_U > 0.6$) series resulted in a median of 2.4% and 10.3% of analysed epochs as grey areas in the 50 recordings, respectively.

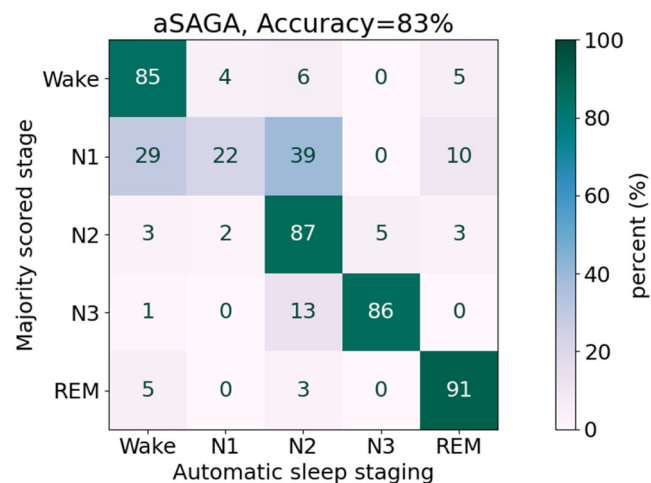
When all epochs were split into marked uncertainties and unmarked uncertainties, the model grey areas agreed better with the manual scoring grey areas within the marked uncertainties than within

TABLE 3 Performance metrics for the automatic sleep staging in self-applied polysomnographic (PSG) recordings when the aSAGA model was used.

Recording type	Subjects	Signals	Sleep stage	Performance metrics			
				Precision	Recall	F1-score	Number of epochs
Self-applied PSG	Healthy to severe OSA ($n = 50$)	All EEG + EOG channels	Wake	0.63	0.85	0.73	3183
			N1	0.50	0.22	0.30	2303
			N2	0.88	0.87	0.87	17,443
			N3	0.87	0.86	0.86	6136
			REM	0.85	0.91	0.88	6168
			Macro average	0.75	0.74	0.73	35,233
			κ	0.75			35,233
			Accuracy	83%			35,233

Metrics were computed against the majority scores of manual sleep staging. Epochs manually scored with uncertainty were excluded from the analysis. Macro average is the unweighted average over the sleep stage-specific metrics.

Abbreviations: aSAGA, Automatic Sleep Analysis with Grey Areas; EEG, electroencephalography; EOG, electro-oculography; epoch, 30-second signal segment; κ , Cohen's kappa; SAS, Self-Applied Somnography (Nox Medical).

**FIGURE 4** Confusions between automatic sleep staging of aSAGA model and the majority score derived from manual sleep staging in self-applied home sleep recordings. Epochs manually marked as uncertain were excluded from the analysis. aSAGA, Automatic Sleep Analysis with Grey Areas; REM, rapid eye movement.

the unmarked uncertainties (Figure 8). In both cases, the model captured most of the manual scoring grey areas, i.e., 54% within unmarked uncertainties and 61% within the marked uncertainties.

4 | DISCUSSION

This paper presents a novel human-in-the-loop concept for sleep staging that incorporates the integration of automatic analysis and manual review of the automatic output. The efficacy of various hypnodensity-based uncertainty metrics was evaluated in delineating the grey areas within automatic sleep staging and these grey areas compared with

those observed in manual sleep staging. Through preclinical validation, we observed that the performance of the automatic sleep staging model was comparable to that of the manual scorers in different recording types. Hypnodensity-based metrics of uncertainty show potential in defining the grey areas that capture confusions in automatic versus manual sleep staging. Additionally, the grey areas identified from the automatic sleep staging captured most of the regions of the recording where significant discrepancies between multiple manual scorers occurred.

Consequently, focussing on the manual revision of the grey areas alone could enhance the accuracy of sleep staging, while simultaneously minimising the workload associated with fully manual sleep staging. By leveraging the concept of grey areas, the proposed approach offers a valuable user-assistive solution for improving the reliability of sleep staging, especially with unconventional sensor setups. At the same time, transparency of uncertainty can reduce the scepticism related to the application of machine learning methods in clinical medicine (Linardatos et al., 2021; Rudin, 2019), which is one of the greatest challenges in implementing automatic PSG scoring at scale (Alvarez-estevez, 2023). Nevertheless, further research is warranted to evaluate the clinical feasibility, incorporate active learning (Monarch, 2021), and to quantify the reduction in working hours achievable with this approach.

A few previous studies have investigated the uncertainty of automatic sleep staging, utilising various approaches that should be discussed (Bakker et al., 2022; Fiorillo et al., 2021; Hong et al., 2021; Kang et al., 2021; Phan, Mikkelsen, et al., 2021). Phan et al. focussed on a single entropy metric for hypnodensities and employed a fixed threshold to assess sleep staging accuracy in uncertain and certain epochs (Phan, Mikkelsen, et al., 2021). Kang et al. conducted a study involving the use of model uncertainty to enhance sleep staging accuracy, examining 20 recordings (Kang et al., 2021). However, their findings revealed statistically significant improvement in Cohen's kappa

FIGURE 5 Performance metrics of the aSAGA model in automatic sleep staging as a function of the amount of data excluded as grey areas. Grey areas were defined based on different hypnodensity-based uncertainty metrics (U) to compare the effectiveness of these measures. The compared metrics were the least confidence U_L , the margin of confidence U_M , the ratio of confidence U_R , the coefficient of unalikeability U_U , and the classification entropy U_E . F1-score is the weighted average of class-wise F1-scores. Dreem Open Data with obstructive sleep apnea subjects was used in the analysis ($n = 56$ full-night polysomnographies). aSAGA, Automatic Sleep Analysis with Grey Areas.

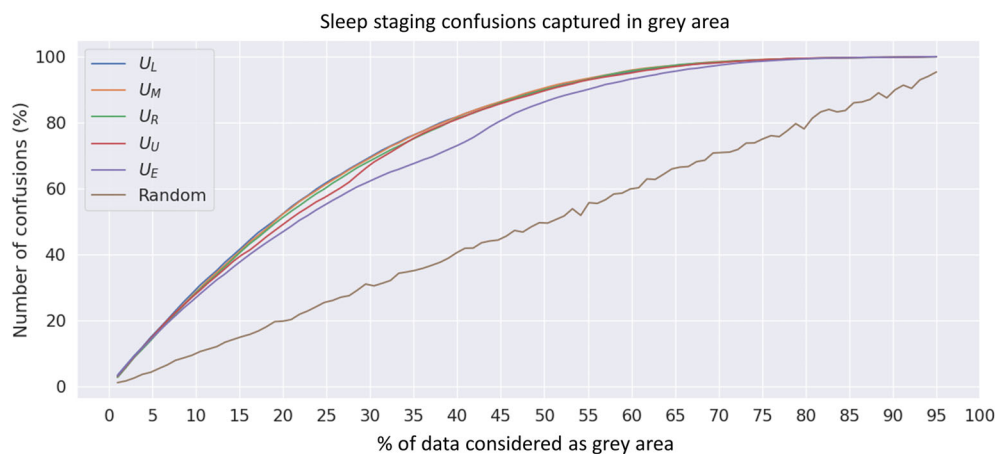
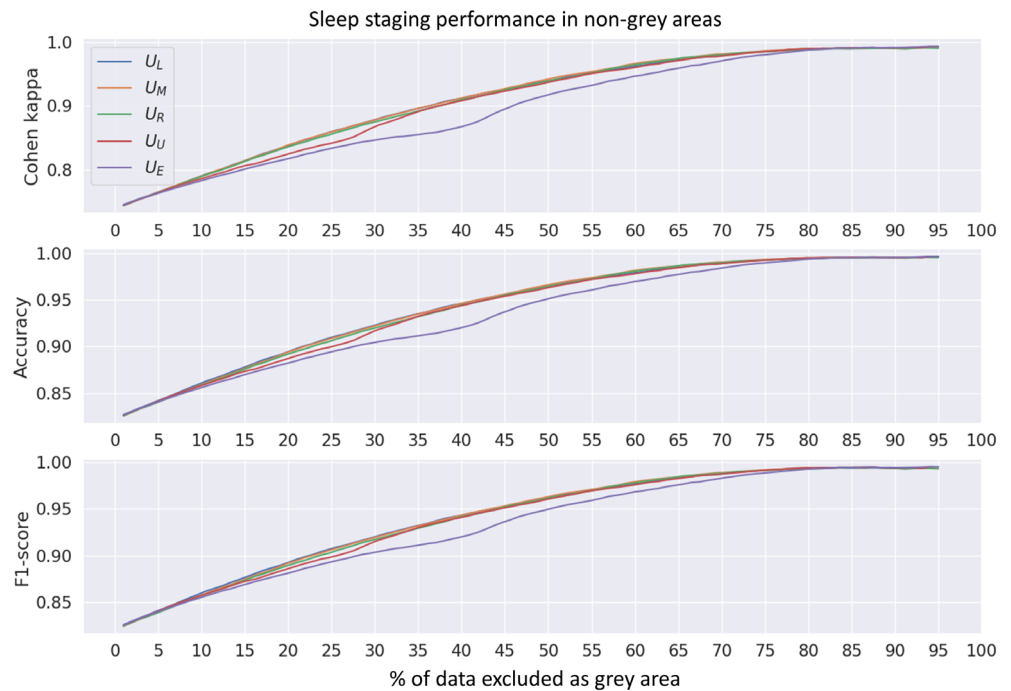


FIGURE 6 Percentage of sleep staging confusion between automatic and manual majority scoring captured in the model grey areas. Grey areas were defined based on different hypnodensity-based uncertainty metrics (U) to compare the effectiveness of these measures. The compared metrics were the least confidence U_L , the margin of confidence U_M , the ratio of confidence U_R , the coefficient of unalikeability U_U , and the classification entropy U_E . Dreem Open Data with obstructive sleep apnea subjects was used in the analysis ($n = 56$ full-night polysomnographies). aSAGA, Automatic Sleep Analysis with Grey Areas.

only for REM vs. N3, rather than for agreement across all five stages. Fiorillo et al. and Hong et al. introduced sophisticated models capable of estimating the uncertainty of automatic sleep staging, aiming to capture discrepancies between manual and automatic approaches (Fiorillo et al., 2021; Hong et al., 2021). Previous validation analyses have primarily focussed on evaluating the extent of discrepancies between manual and automatic scoring within uncertain epochs, without considering the true grey areas associated with manual sleep staging. Only one study compared hypnodensities derived from multiple manual scorings against automatic scoring, revealing a strong correlation between the two (Bakker et al., 2022). However, none of these

works have demonstrated the generalisability of their methods across different recording types. In this paper, we contribute to the field by considering both validation aspects and demonstrating the generalisability of our approach across diverse recording types. Additionally, as a novel contribution, we compare different uncertainty metrics and aim to stimulate discussion surrounding various strategies for defining the grey areas of automatic sleep analysis.

In this study, we defined grey areas using two different approaches. Firstly, when comparing uncertainty metrics, grey areas were defined as a percentage of the most uncertain epochs within the recording. Secondly, when comparing uncertainties between the

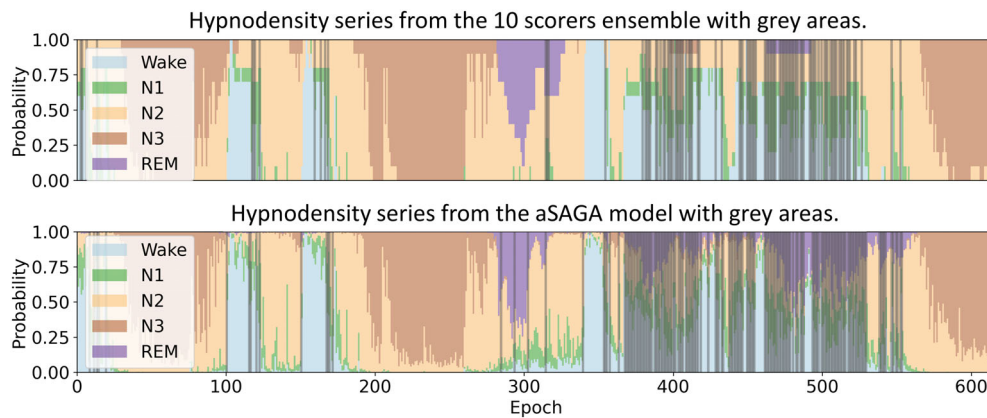


FIGURE 7 An example of hypnodensity series and grey areas derived from manual sleep staging confusions and aSAGA model outputs for one of the self-applied home sleep recordings. Hypnodensities are shown as stacked bars with the respective colours indicated in the legend. Grey areas are shown as shaded vertical bars of a grey colour corresponding to epochs where the coefficient of unlikelihood was higher than 0.6. aSAGA, Automatic Sleep Analysis with Grey Areas.

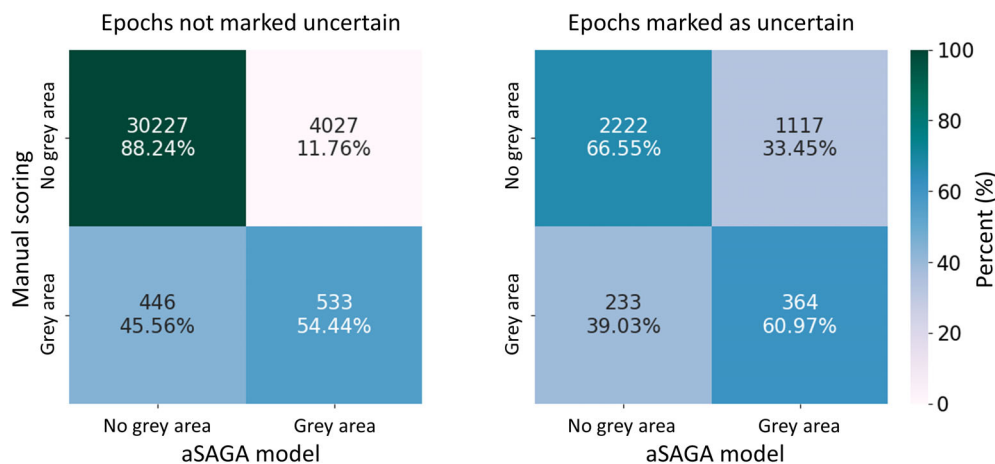


FIGURE 8 Confusions in the grey areas derived from 10 manual scorings and the aSAGA model output. The left confusion matrix represents those areas of the recording that were not annotated as uncertain by any of the 10 scorers while scoring the sleep stages. The right confusion matrix is for epochs that were marked as uncertainty at least by one of the 10 scorers while still scoring the most likely sleep stage (10% of all data). aSAGA = Automatic Sleep Analysis with Grey Areas.

automatic and multiple manual scorings in self-applied PSGs, grey areas were determined using a fixed threshold for the applied uncertainty metric.

Using the threshold method, the number of grey areas varied across recordings, reflecting the inherent variations in the difficulty of scoring different recordings, e.g., healthy vs. OSA subjects (Boostani et al., 2017). It is natural for some recordings to present greater challenges in analysis than others, and theoretically, the threshold method should capture this. However, when employing a fixed value for uncertainty, the calibration of the model influences the number of grey areas. If the model is applied to data differing from the training data, higher uncertainties may be obtained, resulting in an increased number of grey areas. Now the model was trained on standard PSG signals and applied on the forehead EEG of the self-applied PSGs. This could explain why the threshold method yielded more grey areas based on model hypnodensities compared with 10 scorer

hypnodensities in self-applied PSGs (with a median percentage of 2.4% for manual scoring and 10.3% for the model).

Conversely, a fixed percentage of the most uncertain epochs as grey areas could be considered, and then the number of grey areas would remain constant for each recording. However, this approach also has its limitations. Recordings that pose greater challenges for the model's analysis would be treated the same as recordings that are easier to analyse. Consequently, some true grey areas might be overlooked in difficult-to-analyse recordings, or already correctly analysed epochs might be highlighted in easier-to-analyse recordings, which would not be a good use of the manual scorer's time.

This study was limited to comparing the grey areas derived from hypnodensities averaged over predictions from multiple EEG and EOG channels. Although this averaged ensemble prediction appears suitable for generalising over different recording types with various EEG and EOG derivations, some information about the source of the

uncertainty might be lost. This information may differentiate the uncertainty associated with data quality from the uncertainty emerging from manual scoring discrepancies in the training data, although they are likely not fully distinct. These two sources of uncertainty are not currently separated by the proposed model.

The most effective method to define the grey areas for the interaction between manual and automatic sleep staging remains uncertain. Ideally, the sensitivity for grey areas, such as the threshold for uncertainties or the percentage of most uncertain epochs, could be interactively adjusted for each recording, based on the expertise of the sleep technologist. This personalised approach could optimise the identification of grey areas and enhance the overall performance of the sleep staging process. Another approach would be to use, for example, automatic domain adaptation methods and to calibrate the model for each recording type separately (Phan, Chen, et al., 2021). Furthermore, model uncertainty is an active area of machine learning research, and novel ways to quantify it, such as Monte Carlo dropout, should be tested further in the field of sleep medicine (Monarch, 2021; Abdar et al., 2021; Fiorillo et al., 2021).

5 | CONCLUSION

Through extensive evaluation and comparison, we have herein demonstrated that the proposed automatic sleep staging model, along with the identification of grey areas, can achieve similar performance to manual scorers on various types of sleep recordings. Different hypnodynamic-based uncertainty metrics are shown to effectively capture areas of sleep recordings that should be re-analysed for increased sleep staging accuracy. A significant portion of the automatically identified grey areas overlap with those areas of the recordings where confusions between expert-based scorings appear. The two approaches employed for defining grey areas, namely the percentage-based method and the fixed threshold method, offer distinct advantages and trade-offs. However, the optimal approach for determining grey areas should be personalised, considering the expertise of sleep technologists and the specific characteristics of each recording. In conclusion, this study introduces a way to combine manual and automatic sleep staging and provides an investigation of the uncertainty in automatic sleep staging that brings us closer to integrating human-in-the-loop scoring into clinical workflows.

ACKNOWLEDGEMENTS

The authors thank Sleep Technologists at the Charite-Universitaetsmedizin Berlin, Berlin, Germany; Istituto Clinici Scientifico Maugeri Spa Societa Benefit, Pavia, Italy; Princess Alexandra Hospital, Brisbane, Australia; Reykjavik University Sleep Institute, Reykjavik, Iceland; Turku University Central Hospital, Turku, Finland; University of Gothenburg, Gothenburg, Sweden; and University of Lisbon, Lisbon, Portugal for manual scoring of the self-applied polysomnography recordings.

FUNDING INFORMATION

This work was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 965417; by the State Research Funding for university-level health research, Kuopio University Hospital, Wellbeing service county of North Savo (Grants 5041807, 5041789, 5041794, 5041797, 5041803, 5041809, 5041804); by the Finnish Cultural Foundation through Kainuu Regional Fund and Central Fund; by the Emil Aaltonen Foundation; by the Research Foundation of the Pulmonary Diseases, by the Foundation of the Finnish Anti-Tuberculosis Association; by Tampere Tuberculosis Foundation; by the NordForsk (NordSleep Project 90458) through Business Finland (Grant 5133/31/2018); by the French National Research Agency, framework of the Investissements MIAI Artificial Intelligence chairs of excellence from the Grenoble Alpes (grants ANR-15-IDEX-02 and ANR-19-P3IA-0003) and the Icelandic Research Fund.

CONFLICT OF INTEREST STATEMENT

ESA reports honoraria from ResMed, Nox Medical, Jazz Pharmaceuticals, Linde Healthcare, Alcoa-Fjardaral, Wink Sleep, Apnimed, and Vistor. She is also a member of the Philips Medical Advisory Board. Other authors declare no conflict of interest relevant to this study.

DATA AVAILABILITY STATEMENT

The private clinical data from Princess Alexandra Hospital include sensitive medical information and can therefore not be made publicly available. The self-applied PSGs are stored in a safe medical data environment for research purposes, but not made publicly available. DREAM Open Data is publicly available.

ORCID

Matias Rusanen  <https://orcid.org/0000-0001-9633-6016>

Riku Huttunen  <https://orcid.org/0000-0001-8290-4764>

Sami Nikkonen  <https://orcid.org/0000-0003-0615-4118>

Sigríður Sigurðardóttir  <https://orcid.org/0000-0001-6327-8956>

Juha Töyräs  <https://orcid.org/0000-0002-8035-1606>

Brett Duce  <https://orcid.org/0000-0002-3134-5138>

Sami Myllymaa  <https://orcid.org/0000-0002-9779-4358>

Erna Sif Arnardóttir  <https://orcid.org/0000-0003-0877-3529>

Timo Leppänen  <https://orcid.org/0000-0003-4017-821X>

Samu Kainulainen  <https://orcid.org/0000-0003-2798-2115>

Henri Korkalainen  <https://orcid.org/0000-0002-4686-6112>

REFERENCES

- Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76, 243–297. <https://doi.org/10.1016/j.inffus.2021.05.008>
- Alvarez-Estevéz, D. (2023). Challenges of applying automated polysomnography scoring at scale. *Sleep Medicine Clinics*, 18(3), 277–292. <https://doi.org/10.1016/j.jsmc.2023.05.002>

- Alvarez-Estevez, D., & Rijsman, R. M. (2021). Inter-database validation of a deep learning approach for automatic sleep scoring. *PLoS One*, 16(8), 1–27. <https://doi.org/10.1371/journal.pone.0256111>
- Anderer, P., Ross, M., Cerny, A., Vasko, R., Shaw, E., & Fonseca, P. (2023). Overview of the hypnodensity approach to scoring sleep for polysomnography and home sleep testing. *Frontiers in Sleep*, 2, 1–24. <https://doi.org/10.3389/frsle.2023.1163477>
- Arnardóttir, E. S., Islind, A. S., Óskarsdóttir, M., Ólafsdóttir, K. A., August, E., Jónasdóttir, L., Hrubos-Strøm, H., Saavedra, J. M., Grote, L., Hedner, J., Höskuldsson, S., Ágústsson, J. S., Jóhannsdóttir, K. R., McNicholas, W. T., Pevernagie, D., Sund, R., Töyräs, J., Leppänen, T., & Sleep Revolution. (2022). The sleep revolution project: The concept and objectives. *Journal of Sleep Research*, 31(4), 1–8. <https://doi.org/10.1111/jsr.13630>
- Arnardóttir, E. S., Verbraecken, J., Gonçalves, M., Gjerstad, M. D., Grote, L., Puertas, F. J., Mihaicuta, S., McNicholas, W. T., Parrino, L., & The National Representatives as Study Collaborators. (2016). Variability in recording and scoring of respiratory events during sleep in Europe: A need for uniform standards. *Journal of Sleep Research*, 25(2), 144–157. <https://doi.org/10.1111/jsr.12353>
- Bakker, J. P., Ross, M., Cerny, A., Vasko, R., Shaw, E., Kuna, S., Magalang, U. J., Punjabi, N. M., & Anderer, P. (2022). Scoring sleep with artificial intelligence enables quantification of sleep stage ambiguity: Hypnodensity based on multiple expert scorers and auto-scoring. *Sleep*, no. July, 1–12. <https://doi.org/10.1093/sleep/zsac154>
- Berry, R. B., Brooks, R., Gamaldo, C. E., Harding, S. M., Lloyd, R. M., Quan, S. F., Troester, M. M., Vaughn, B. V. (2018). *AASM manual for the scoring of sleep and associated events*. American Academy of Sleep Medicine.
- Berry, R. B., Brooks, R., Gamaldo, C. E., Harding, S. M., Lloyd, R. M., Quan, S. F., Troester, M. M., Vaughn, B. V. (2020). *The AASM manual for the scoring of sleep and associated events*. American Academy of Sleep Medicine.
- Boostani, R., Karimzadeh, F., & Nami, M. (2017). A comparative review on sleep stage classification methods in patients and healthy individuals. *Computer Methods and Programs in Biomedicine*, 140, 77–91. <https://doi.org/10.1016/j.cmpb.2016.12.004>
- Carskadon, M. A., & Dement, W. (2011). Monitoring and staging human sleep. In S. A. Keenan, M. Hirshkowitz, & H. Casseres (Eds.), *Principles and practice of sleep medicine* (5th ed., pp. 16–26). Elsevier Saunders.
- Fiorillo, L., Favaro, P., & Faraci, F. D. (2021). DeepSleepNet-lite: A simplified automatic sleep stage scoring model with uncertainty estimates. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29, 2076–2085. <https://doi.org/10.1109/TNSRE.2021.3117970>
- Fiorillo, L., Puiatti, A., Papandrea, M., Ratti, P. L., Favaro, P., Roth, C., Bargiotas, P., Bassetti, C. L., & Faraci, F. D. (2019). Automated sleep scoring: A review of the latest approaches. *Sleep Medicine Reviews*, 48, 101204. <https://doi.org/10.1016/j.smrv.2019.07.007>
- Fischer, J., Dogas, Z., Bassetti, C. L., Berg, S., Grote, L., Jennum, P., Levy, P., Mihaicuta, S., Nobili, L., Riemann, D., Puertas Cuesta, F. J., Raschke, F., Skene, D. J., Stanley, N., Pevernagie, D., Executive Committee (EC) of the Assembly of the National Sleep Societies (ANSS), & Board of the European Sleep Research Society (ESRS), Regensburg, Germany. (2012). Standard procedures for adults in accredited sleep medicine centres in Europe. *Journal of Sleep Research*, 21(4), 357–368. <https://doi.org/10.1111/j.1365-2869.2011.00987.x>
- Goldstein, C. A., Berry, R. B., Kent, D. T., Kristo, D. A., Seixas, A. A., Susan Redline, M., Westover, B., Fariha Abbasi-Feinberg, R., Aurora, N., Carden, K. A., Kirsch, D. B., Malhotra, R. K., Martin, J. L., Olson, E. J., Ramar, K., Rosen, C. L., Rowley, J. A., & Shelgikar, A. V. (2020). Artificial intelligence in sleep medicine: An American Academy of sleep medicine position statement. *Journal of Clinical Sleep Medicine*, 16(4), 15–17.
- Guillot, A., Sauvet, F., During, E. H., & Thorey, V. (2020). Dreem open datasets: Multi-scored sleep datasets to compare human and automated sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(9), 1955–1965. <https://doi.org/10.1109/TNSRE.2020.3011181>
- Himanan, S. L., & Hasan, J. (2000). Limitations of Rechtschaffen and Kales. *Sleep Medicine Reviews*, 4(2), 149–167. <https://doi.org/10.1053/smr.1999.0086>
- Hong, J. K., Lee, T., Delos Reyes, R. D., Hong, J., Tran, H. H., Lee, D., Jung, J., & Yoon, I. Y. (2021). Confidence-based framework using deep learning for automated sleep stage scoring. *Nature and Science of Sleep*, 13(December), 2239–2250. <https://doi.org/10.2147/NSS.S333566>
- Huttunen, R., Kainulainen, S., Nikkonen, S., Rusanen, M., & Korkalainen, H. (2023). UEF-SmartSleepLab/sleeplab-format. <https://doi.org/10.5281/zenodo.7861887>
- Huttunen, R., Leppänen, T., Duce, B., Arnardóttir, E. S., Nikkonen, S., Myllymaa, S., Töyräs, J., & Korkalainen, H. (2022). A comparison of signal combinations for deep learning-based simultaneous sleep staging and respiratory event detection. *IEEE Transactions on Biomedical Engineering*, 70(5), 1704–1714. <https://doi.org/10.1109/TBME.2022.3225268>
- Id, D. A., & Rijsman, R. M. (2022). Computer-assisted analysis of polysomnographic recordings improves inter-scorer associated agreement and scoring times. *PLoS One*, 17(9), e0275530. <https://doi.org/10.1371/journal.pone.0275530>
- Jorgensen, G., Downey, C., Goldin, J., Melehan, K., Rochford, P., & Ruehland, W. (2020). An Australasian commentary on the AASM manual for the scoring of sleep and associated events. *Sleep and Biological Rhythms*, 18(3), 163–185. <https://doi.org/10.1007/s41105-020-00259-9>
- Kalevo, L., Miettinen, T., Leino, A., Kainulainen, S., Myllymaa, K., Toyras, J., Leppänen, T., & Myllymaa, S. (2019). Improved sweat artifact tolerance of screen-printed EEG electrodes by material selection-comparison of electrochemical properties in artificial sweat. *IEEE Access*, 7, 133237–133247. <https://doi.org/10.1109/ACCESS.2019.2941052>
- Kang, D. Y., DeYoung, P. N., Tantiogloc, J., Coleman, T. P., & Owens, R. L. (2021). Statistical uncertainty quantification to augment clinical decision support: A first implementation in sleep medicine. *npj Digital Medicine*, 4(1), 1–9. <https://doi.org/10.1038/s41746-021-00515-3>
- Lee, Y. J., Lee, J. Y., Cho, J. H., & Choi, J. H. (2021). Inter-rater reliability of sleep stage scoring: A meta-analysis. *Journal of Clinical Sleep Medicine*, 18(1), 193–202. <https://doi.org/10.5664/jcsm.9538>
- Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1), 1–45. <https://doi.org/10.3390/e23010018>
- Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. Available: <https://arxiv.org/abs/1711.05101>
- Monarch, R. (2021). Human-in-the-loop machine learning: Active learning and annotation for human-centered AI. *Shelter Island: Manning*.
- Nikkonen, S., Somaskandhan, P., Korkalainen, H., Kainulainen, S., Terrill, P. I., Gretarsdóttir, H., Sigurdardóttir, S., Ólafsdóttir, K. A., Islind, A. S., Óskarsdóttir, M., Arnardóttir, E. S., & Leppänen, T. (2023). Multicentre sleep-stage scoring agreement in the sleep revolution project. *Journal of Sleep Research*, no. March, 1–13. <https://doi.org/10.1111/jsr.13956>
- Norman, R. G., Pal, I., Stewart, C., Walsleben, J. A., & Rapoport, D. M. (2000). Interobserver agreement among sleep scorers from different centers in a large dataset. *Sleep*, 23(7), 901–908. <https://doi.org/10.1093/sleep/23.7.1e>
- Olesen, A. N., Jørgen Jennum, P., Mignot, E., & Sorensen, H. B. D. (2021). Automatic sleep stage classification with deep residual networks in a mixed-cohort setting. *Sleep*, 44(1), 1–12. <https://doi.org/10.1093/sleep/zsaa161>
- Parrino, L., Ferri, R., Zucconi, M., & Fanfulla, F. (2009). Commentary from the Italian Association of Sleep Medicine on the AASM manual for the scoring of sleep and associated events: For debate and discussion. *Sleep Medicine*, 10(7), 799–808. <https://doi.org/10.1016/j.sleep.2009.05.009>

- Perslev, M., Darkner, S., Kempfner, L., Nikolic, M., Jennum, P. J., & Igel, C. (2021). U-sleep: Resilient high-frequency sleep staging. *npj Digital Medicine*, 4(1), 1–12. <https://doi.org/10.1038/s41746-021-00440-5>
- Perslev, M., Jensen, M. H., Darkner, S., Jennum, P. J., & Igel, C. (2019). U-time: A fully convolutional network for time series segmentation applied to sleep staging. *arXiv*, 1–19. Available: <https://arxiv.org/abs/1910.11162>
- Phan, H., Chen, O. Y., Koch, P., Lu, Z., McLoughlin, I., Mertins, A., & De Vos, M. (2021). Towards more accurate automatic sleep staging via deep transfer learning. *IEEE Transactions on Biomedical Engineering*, 68(6), 1787–1798. <https://doi.org/10.1109/TBME.2020.3020381>
- Phan, H., & Mikkelsen, K. (2022). Automatic sleep staging of EEG signals: Recent development, challenges, and future directions. *Physiological Measurement*, 43(4), 4TR01. <https://doi.org/10.1088/1361-6579/ac6049>
- Phan, H., Mikkelsen, K., Chén, O. Y., Koch, P., Mertins, A., & De Vos, M. (2021). *SleepTransformer: Automatic sleep staging with interpretability and uncertainty quantification*. *IEEE Transactions on Biomedical Engineering*, 2456–2467. <http://arxiv.org/abs/2105.11043>
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. N. Navab, J. Hornegger, W. Wells, A. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. MICCAI (Vol. 935). Springer. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Rusanen, M., Huttunen, R., Korkalainen, H., Myllymaa, S., Toyras, J., Myllymaa, K., Sigurdardottir, S., Olafsdottir, K. A., Leppänen, T., Arnardottir, E. S., & Kainulainen, S. (2023). Generalizable deep learning-based sleep staging approach for ambulatory textile electrode headband recordings. *IEEE Journal of Biomedical and Health Informatics*, 27(8), 1869–1880. <https://doi.org/10.1109/JBHI.2023.3240437>
- Rusanen, M., Korkalainen, H., Gretarsdottir, H., Siilak, T., Olafsdottir, K. A., Töyräs, J., Myllymaa, S., Arnardottir, E. S., Leppänen, T., & Kainulainen, S. (2023). Self-applied somnography: Technical feasibility of electroencephalography and electro-oculography signal characteristics in sleep staging of suspected sleep-disordered adults. *Journal of Sleep Research*, no. June, 1–13. <https://doi.org/10.1111/jsr.13977>
- Schulz, H. (2008). Rethinking sleep analysis. *Journal of Clinical Sleep Medicine*, 4(2), 99–103.
- Smith, L. N. (2016). A disciplined approach to neural network hyperparameters: Part 1 –learning rate, batch size, momentum, and weight decay. Available: <https://arxiv.org/abs/1803.09820>.
- Stephansen, J. B., Olesen, A. N., Olsen, M., Ambati, A., Leary, E. B., Moore, H. E., Carrillo, O., Lin, L., Han, F., Yan, H., Sun, Y. L., Dauvilliers, Y., Scholz, S., Barateau, L., Hogl, B., Stefani, A., Hong, S. C., Kim, T. W., Pizza, F., ... Mignot, E. (2018). Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature Communications*, 9(1), 1–16. <https://doi.org/10.1038/s41467-018-07229-3>
- Sveinbjarnarson, B. F., Schmitz, L., Arnardottir, E. S., & Islind, A. S. (2023). The sleep revolution platform: A dynamic data source pipeline and digital platform architecture for complex sleep data. *Current Sleep Medicine Reports*, 9(2), 91–100. <https://doi.org/10.1007/s40675-023-00252-x>
- Van Gorp, H., Huijben, I. A. M., Fonseca, P., Van Sloun, R. J. G., Overeem, S., & Van Gilst, M. M. (2022). Certainty about uncertainty in sleep staging: A theoretical framework. *Sleep*, 45(8), 1–8. <https://doi.org/10.1093/sleep/zsac134>
- Zhang, X., Dong, X., Kantelhardt, J. W., Li, J., Zhao, L., Garcia, C., Glos, M., Penzel, T., & Han, F. (2015). Process and outcome for international reliability in sleep scoring. *Sleep & Breathing*, 19(1), 191–195. <https://doi.org/10.1007/s11325-014-0990-0>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Rusanen, M., Jouan, G., Huttunen, R., Nikkonen, S., Sigurðardóttir, S., Töyräs, J., Duce, B., Myllymaa, S., Arnardottir, E. S., Leppänen, T., Islind, A. S., Kainulainen, S., & Korkalainen, H. (2024). Retrospective validation of automatic sleep analysis with grey areas model for human-in-the-loop scoring approach. *Journal of Sleep Research*, e14362. <https://doi.org/10.1111/jsr.14362>