Generalizable Deep Learning-Based Sleep Staging Approach for Ambulatory Textile Electrode Headband Recordings

Matias Rusanen, Riku Huttunen, Henri Korkalainen, Sami Myllymaa, Juha Töyräs, Katja Myllymaa, Sigridur Sigurdardottir, Kristin A. Olafsdottir, Timo Leppänen, Erna S. Arnardottir, and Samu Kainulainen

***Abstract*—Reliable, automated, and user-friendly solutions for the identification of sleep stages in home environment are needed in various clinical and scientific research settings. Previously we have shown that signals recorded with an easily applicable textile electrode headband (FocusBand Technologies, T 2 Green Pty Ltd) contain characteristics similar to the standard electrooculography (EOG, E1–M2). We hypothesize that the electroencephalographic (EEG) signals recorded using the textile electrode headband are similar enough with standard EOG in order to develop an automatic neural network-based sleep staging method that generalizes from diagnostic polysomnographic (PSG) data to ambulatory sleep recordings of textile electrode-based forehead EEG. Standard EOG signals together with manually annotated sleep stages from clinical PSG dataset ($n = 876$) were used to train, validate, and test a fully convolutional neural network (CNN). Furthermore, ambulatory sleep recordings including a standard set of gel-based electrodes and the textile electrode headband were conducted for 10 healthy volunteers at their homes to test the generalizability of the model. In the test set ($n = 88$) of the clinical dataset, the model's accuracy for 5-stage sleep stage classification was 80% ($\kappa = 0.73$) using only the single-channel EOG. The model generalized well for the headband-data, reaching 82% ($\kappa = 0.75$) overall sleep staging accuracy. In comparison, accuracy of the model was 87% ($\kappa = 0.82$) in home recordings using the standard EOG. In conclusion, the CNN model shows potential on automatic sleep staging of healthy individuals using a reusable electrode headband in a home environment.**

***Index Terms*—Deep learning, electrooculography, sleep, textile electrodes, wearables, convolutional neural network.**

Manuscript received 10 May 2022; revised 27 August 2022 and 1 November 2022; accepted 20 January 2023. Date of publication 30 January 2023; date of current version 5 April 2023. This work was supported in part by the NordForsk under Project 90458 through the Business Finland under Grant 5133/31/2018, in part by the Icelandic Centre for Research under Project 90458-06111, in part by the Academy of Finland under Grant 323536, in part by the Research Committee of the Kuopio University Hospital Catchment Area for the State Research Funding under Grants 5041767, 5041768, 5041789, 5041794, 5041803, 5041804, 5041797, and 5041807, in part by the Finnish Cultural Foundation through North Savo Regional Fund and Kainuu Regional Fund, in part by Olvi Foundation, in part by the Respiratory Foundation of Kuopio Region, in part by the Research Foundation of the Pulmonary Diseases, in part by the Finnish Anti-Tuberculosis Foundation, in part by the Tampere Tuberculosis Foundation, in part by the Maud Kuistila Memorial Foundation, in part by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement 965417, and in part by the Emil Aaltonen Foundation. *(Corresponding author: Matias Rusanen.)*

Matias Rusanen, Riku Huttunen, Henri Korkalainen, Sami Myllymaa, and Samu Kainulainen are with the Department of Technical Physics, University of Eastern Finland, FI-70211 Kuopio, Finland, and also with the Diagnostic Imaging Center, Kuopio University Hospital, FI-70211 Kuopio, Finland (e-mail: matias.rusanen@uef.fi; riku.huttunen@uef.fi; henri.korkalainen@uef.fi; sami.myllymaa@uef.fi; samu.kainulainen@uef.fi).

Katja Myllymaa is with the Diagnostic Imaging Center, Kuopio University Hospital, FI-70211 Kuopio, Finland (e-mail: katja.myllymaa@kuh.fi).

Juha Töyräs is with the Department of Technical Physics, University of Eastern Finland, FI-70211 Kuopio, Finland, also with the Science Service Center, Kuopio University Hospital, FI-70211 Kuopio, Finland, and also with the School of Information Technology and Electrical Engineering, The University of Queensland, QLD 4067 Brisbane, Australia (e-mail: juha.toyras@kuh.fi).

Timo Leppänen is with the School of Information Technology and Electrical Engineering, The University of Queensland, QLD 4067 Brisbane, Australia, also with the Department of Technical Physics, University of Eastern Finland, FI-70211 Kuopio, Finland, and also with the Diagnostic Imaging Center, Kuopio University Hospital, FI-70211 Kuopio, Finland (e-mail: timo.leppanen@uef.fi).

Sigridur Sigurdardottir and Kristin A. Olafsdottir are with the Reykjavik University Sleep Institute, School of Technology, 102 Reykjavik, Iceland (e-mail: sigridursig@ru.is; kristinao@ru.is).

Erna S. Arnardottir is with the Reykjavik University Sleep Institute, School of Technology, 102 Reykjavik, Iceland, and also with the Landspitali–The National University Hospital of Iceland, 101 Reykjavik, Iceland (e-mail: ernasifa@ru.is).

Digital Object Identifier 10.1109/JBHI.2023.3240437

I. Introduction

NEW technological solutions to screen for sleep disorders in the home environment are needed to alleviate the workload of sleep laboratories performing in-laboratory polysomnography (PSG) [1], [2]. Some alternatives, such as unattended and portable Type III sleep studies, are already utilized by many clinical units worldwide [3] and are especially popular options for home sleep apnea testing (HSAT) [1], [4]. However, in comparison to attended in-lab PSG ($\geq$7 channels), type III only records a limited number of channels (4-7) and does not include electroencephalography (EEG) [5]. This is a significant limitation, as the EEG recording is currently required for the accurate determination of sleep structure [6], [7]. However, the clinically used gel-based EEG electrodes are currently not self-applicable and are usually not reusable. As a result, the clinical electrodes are neither suitable for monitoring over multiple nights in a home environment nor for patient self-application [8]. In fact, professional staff is required for accurate positioning and preparation of the EEG montage. Unfortunately, the EEG electrodes are still prone to detach, and especially in an unattended environment, this may lead to failed recording and the need for retesting [9].

Similarly, many research areas are in the need of more reliable ways to quantify sleep structure from unattended multi-night home recordings than what is achieved with the currently available wearable devices without EEG [7]. These needs could be met by supplementing portable recording devices with reusable and wearable EEG electrodes. One possible solution is a textile electrode headband, such as FocusBand (FocusBand Technologies, Windaroo, Australia) which is a neoprene headband with integrated textile electrodes for the recording of forehead EEG [10]. This headband was originally developed for sports training with associated neurofeedback protocols but could also be used alongside home sleep recordings, due to its comfortable and self-applicable design. We have previously shown that the technical performance of the headband is suitable for home-based sleep studies [11]. The headband-recorded forehead EEG signals were found to have significantly lower amplitudes but similar waveforms and frequency content as standard EOG and frontal EEG signals [10], [11]. However, the differences in signal characteristics might make manual sleep staging of the headband recordings unfeasible. That is because the manual process has specific rules related to the interpretation of standard EEG, EOG, and chin-electromyography (chin-EMG) signal features [6]. More adaptive methods, such as deep learning-based automatic sleep staging [11], could be the solution to overcome these issues.

The current deep learning-based automatic sleep staging methods are already as reliable as manual sleep staging when based on standard PSG recordings [12], [13], [14], [15], [16], [17]. However, those methods typically utilize neural networks, that require numerous manually analyzed recordings i.e., the extensive training data, to be optimized for the task. Collecting a big dataset that includes both, the signals recorded using a novel wearable and the standard PSG signals needed for manual sleep staging is a laborious process. Thus, optimizing the automatic sleep staging method directly with the wearable-recorded signals, might not be feasible. Therefore, solutions utilizing large retrospective datasets such as standard clinical PSG recordings and transferring the accurate sleep staging achieved on those to wearable-based data, are of significant interest [18], [19], [20], [21].

Transfer learning is an effective way to generalize the learning achieved in one dataset to another by fine-tuning the network with a limited number of recordings [18], [22], [23]. However, extensive regularization and freezing parts of the network are required to avoid overfitting, making it difficult to perform optimal fine-tuning [17]. Fortunately, some of the state-of-the-art neural networks used in sleep staging are shown to be relatively well generalizable over different datasets even without further fine-tuning [14], [15], [22]. This could enable an interesting opportunity to use signals recorded with wearables as an input for such models. The requirement is that the wearable-based signals should, at least on some level, correspond to those used in the training of the network. The level of correspondence that is required for this method to work accurately in automatic sleep staging remains to be studied.

Based on our previous study [11], we hypothesize that the forehead EEG signals recorded using the textile electrode headband are similar enough with standard EOG to enable reliable automatic sleep staging with a deep neural network that is optimized using an extensive PSG dataset including standard EOG signals and manually annotated sleep stages. Therefore, the aim of this study was to develop and test an automatic sleep staging method that could generalize from in-laboratory PSG recordings to forehead EEG recorded using the textile electrode headband (Fig. 1).

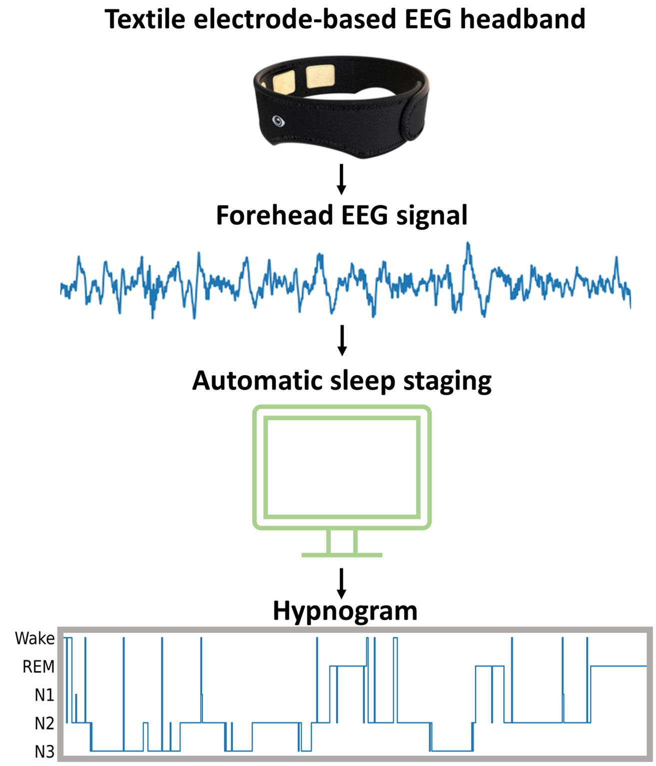

Fig. 1. A wearable electroencephalography (EEG) electrode headband is proposed for home sleep recordings for its comfortable design and full reusability. In this study, we aimed to develop, test, and validate an automatic sleep staging method suitable for the visualized protocol. Figure of the headband courtesy of FocusBand Technologies, Windaroo, Australia.

II. Methods

A. Datasets

This study utilized two different datasets: a large retrospective clinical dataset of in-lab PSGs ($n = 876$) and a smaller set ($n = 10$) of home-based sleep recordings on healthy volunteers. The clinical dataset was used for training, validation, and testing the automatic sleep staging model. The dataset consisting of home sleep recordings was used as a separate test set to investigate how the automatic sleep staging model performs on the textile electrode headband recordings (Fig. 2). Ethical clearance for research use of the collected clinical data was applied from and granted by The Institutional Human Research Ethics Committee of the Princess Alexandra Hospital (HREC/16/QPAH/021 and LNR/2019/QMS/54313). The favorable statement for the protocol for home sleep recordings was given by The Research Ethics Committee of The Northern Savo Hospital District (849/2018) and all research participants were informed with oral as well as written instructions and they gave written informed consent.

1) Clinical Dataset: Diagnostic PSG recordings were conducted for 933 individuals with suspicion of obstructive sleep apnea (OSA) at the Princess Alexandra Hospital, Brisbane, Australia between the years 2015 and 2017. Only successful

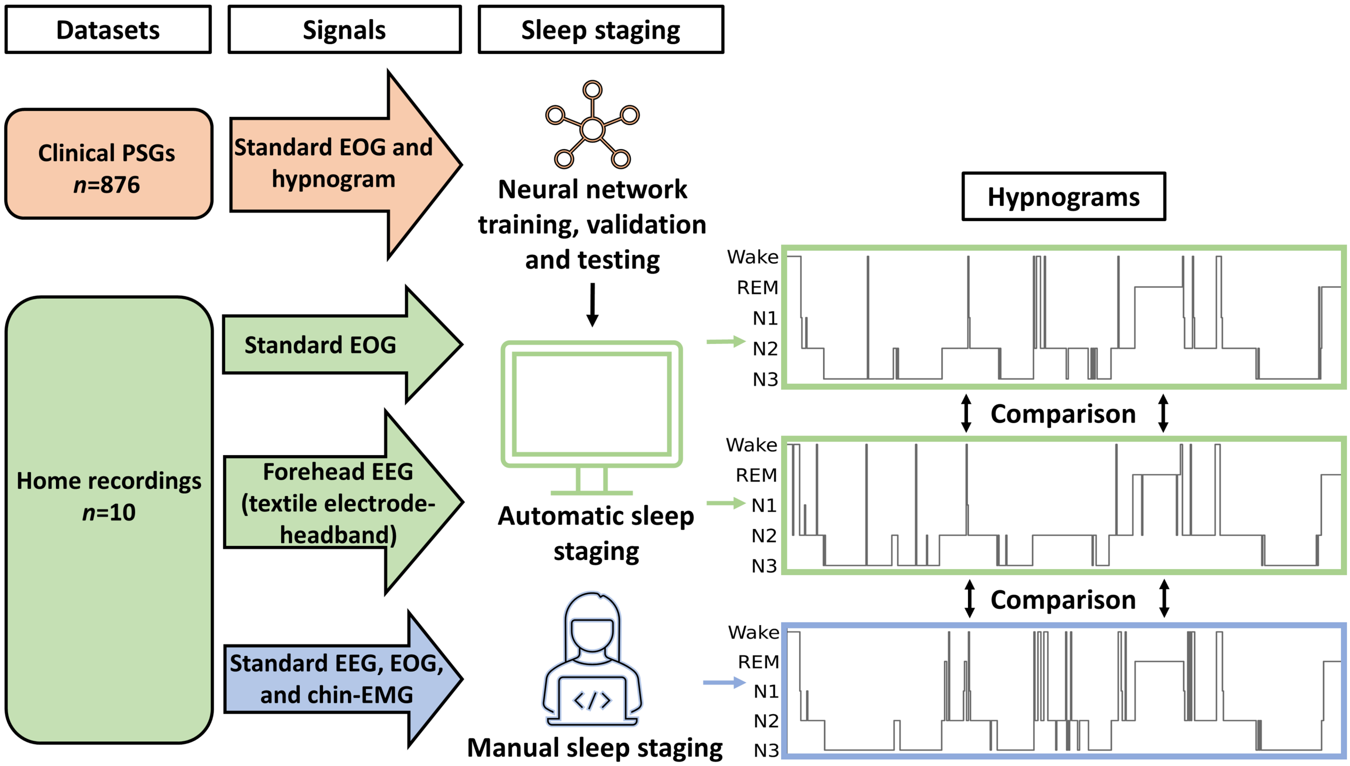

Fig. 2. The structure of the present study illustrated as a flowchart. The overall aim of the study was to test how a deep neural network-based automatic sleep staging method generalizes from clinical polysomnographic (PSG) data to textile electrode-based home recordings without fine-tuning of the network. Abbreviations: EEG = electroencephalography, EOG = electrooculography, EMG = electromyography, REM = rapid eye movement sleep, N1 = Stage N1 sleep, N2 = Stage N2 sleep, N3 = Stage N3 sleep.

PSGs, 876 in total, were included in the training, validation, and testing processes. 57 recordings were excluded due to insufficient amount of total sleep time (<1 h) or poor signal quality. Recordings were conducted using Compumedics (Abbotsford, Australia) Grael 4K PSG system. The Compumedics device records EEG and EOG signals with 1024 Hz sampling frequency. Subject preparation, sensor attachment, and PSG setup were conducted by medical experts following the standards set by the American Academy of Sleep Medicine (AASM) [6].

2) Home Recordings: Home sleep recordings were conducted for ten young healthy adults (aged between 23 and 37 years), of which seven were men and three were women. All subjects slept in their natural sleeping environment and followed their usual sleeping rhythm. The recordings were conducted using a portable Nox A1 (Nox Medical, Reykjavik, Iceland) PSG device. The Nox device has an original sampling rate of 256 kHz for EEG and EOG signals, but the recording device saves the signals using 200 Hz. Standard EEG (F4, C4, O2), EOG (E1 and E2), and chin-EMG signals, as well as the reference potentials on mastoids (M1 and M2), were recorded with medical-grade electrodes (Neuroline 720 and 726, Ambu A/S, Copenhagen, Denmark) (Fig. 3). Self-adhesive and pre-gelled silver/silver chloride electrodes (Neuroline 720) were used on facial areas to record the standard EOG and chin-EMG signals as well as for patient ground. The EEG signals and the reference were recorded using silver/silver-chloride cup electrodes (Neuroline 726) to ensure stable skin-electrode contact on hairy areas. The cup electrodes were attached by investigators following the clinical practices from Kuopio University Hospital. Nuprep skin prep gel and Ten20 conductive paste (Weaver and Company, Aurora, CO, USA) were used for the preparation of the electrode attachment site and to fill the cup electrode. Finally, the cup electrodes were firmly fixed using EC2 (Natus Medical Inc., Pleasanton, CA, USA) adhesive electrode cream and skin tape.

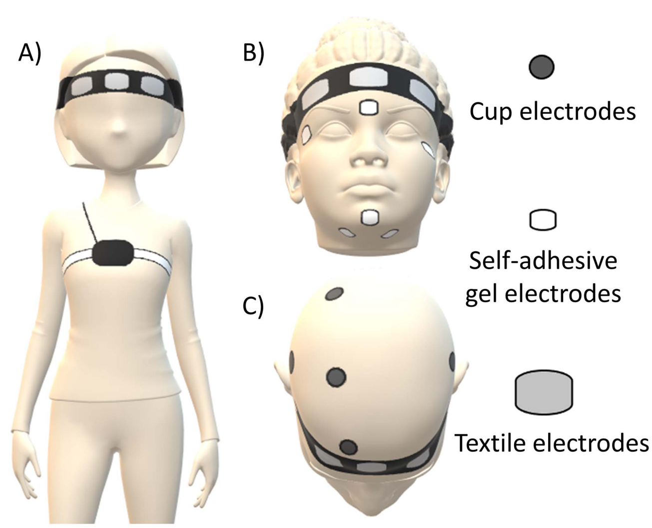

Fig. 3. The device and electrode configurations used in home sleep recordings. (a) The wearable textile electrode headband and a portable recording device illustrated on a subject apart from the medical-grade cup and gel electrodes. (b) Electrode placements on facial area. (c) Electrode placement on scalp.

The FocusBand textile electrodes were connected to unipolar EEG channels of the same Nox A1 recording device as the

TABLE I
DEMOGRAPHIC INFORMATION OF THE UTILIZED DATASETS PRESENTED AS MEDIAN (25–75% QUARTILE)

	PSG Training set (*n*=709)	PSG Validation set (*n*=79)	PSG Test set (*n*=88)	Home recordings, Test set (*n*=10)
Age (years)	55.5 (44.7-65.7)	56.9 (45.8-65.1)	55.8 (43.6-65.9)	27.8 (26.0-31.9)
Body mass index (kg/m^2)	34.7 (29.3-40.4)	34.0 (29.2-38.6)	31.9 (28.8-39.4)	N/A
Total sleep time (min)	309.0 (254.5-360.0)	302.5 (262.5-353.0)	311.0 (245.8-347.4)	390.5 (375.5-411.0)
Sleep efficiency (%)	71.6 (58.2-82.0)	69.4 (61.3-82.4)	67.8 (56.7-80.8)	91.1 (85.3-93.1)
Wake after sleep onset (min)	102.0 (60.6-148.5)	104.5 (62.0-145.8)	109.8 (66.0-158.4)	38.3 (26.0-48.0)
Arousal index (1/h)	20.5 (13.8-31.3)	21.0 (15.0-30.1)	22.6 (14.8-35.8)	16.1 (14.3-21.1)
Stage N1 sleep (%)	11.1 (6.8-18.9)	10.1 (6.9-17.2)	11.8 (6.3-19.6)	2.7 (1.3-3.6)
Stage N2 sleep (%)	48.5 (41.6-56.2)	48.9 (41.5-57.8)	45.7 (38.6-53.3)	49.2 (43.6-51.8)
Stage N3 sleep (%)	18.5 (10.4-27.1)	16.5 (8.6-25.6)	17.9 (9.1-26.5)	27.0 (21.9-32.3)
Stage REM sleep (%)	17.1 (12.0-22.0)	17.2 (13.2-22.2)	19.2 (11.6-23.0)	23.3 (22.1-25.8)
Total recording time (min)	442.0 (410.1-474.5)	440.5 (409.8-473.5)	446.0 (416.4-469.6)	449.4 (429.0-495.9)
Apnea-hypopnea index (1/h)	15.6 (7.0-32.4)	14.0 (6.8-28.9)	19.0 (8.2-43.2)	N/A

The amount of Stage N1, N2, N3, and REM sleep are computed as percentage of total sleep time. Abbreviations: REM = rapid eye movement, PSG = polysomnography.

medical-grade electrodes, which might increase signal inference between the EEG and EOG electrode-recorded signals. The dry textile electrodes of the headband were applied directly to the subject's forehead with no skin preparation. Subjects could adjust the tightness of the headband for maximizing their comfort. Following the findings of our previous in-laboratory study [10], the headband was attached 30 minutes before the start of the recording to allow the skin-electrode interface to stabilize.

A trained sleep technologist from Reykjavik University Sleep Institute (Reykjavik, Iceland) manually annotated the sleep stages based on the standard PSG signals. They used the standard 5-stage classification in compliance with the current AASM rules [6]. The signals from the textile electrode-based headband were not used in the manual sleep staging. Demographic information of the home sleep recordings is presented in Table I. Further details can be found in a previous study considering the technical aspects of the textile electrode headband [11].

B. Neural Network Architecture

A fully convolutional, feed-forward neural network was utilized in this study (Fig. 4). The network architecture was inspired by the U-time sleep staging model [14], which also serves as the basis of the U-sleep model [15]. The fully convolutional structure of U-time was adopted due to its general robustness and good performance over various datasets without further fine-tuning of the network.

The encoder-decoder structure of the neural network architecture consists of blocks utilizing one-dimensional convolutions, batch normalizations, and pooling elements (Fig. 4). Convolutional blocks of kernel size five were implemented with a different number of filters in the encoder structure. The numbers of filters were 32, 48, 64, 96, 128, and 256, respectively for the 6 consecutive blocks of the encoder structure. The convolutional blocks 1-6 included also pooling of sizes of 8, 6, 4, 2, 2, and 1, respectively. Skip connections were implemented between the encoder and decoder to preserve low-level features. The connected upsample blocks 1-5 utilized the same kernel sizes and numbers of filters as the associated convolutional blocks. Upsampling factor was the same as the respective pooling size of the connected convolutional block.

The decoder was followed by a segment classifier utilizing pointwise convolutions, average pooling, and softmax activation to construct the final representation of label predictions i.e., the hypnogram. In the segment classifier, class confidence scores produced in the decoder section for each sample point of the input signal were downsampled with average pooling and trainable convolutional layers to segments of lower temporal resolution. Because of the aforementioned aggregation, the architecture is at this point adjustable to different output resolutions. This enables performing the sleep stage classification in higher temporal frequency, which can be considered an advantage over traditional fixed-size 30-second segments of manual sleep staging.

In addition to U-time-derived structures of the network, a few other state-of-the-art deep learning methods were adapted in the construction of the model's architecture. Atrous Spatial Pyramid Pooling (ASPP) was utilized between encoder and decoder blocks to segment information on variable scales [24]. ASPP is based on using atrous (or dilated) convolutions with different dilation rates (6, 12, and 18 in the present study) that correspond to different receptive fields. The parallel computed fixed-size feature maps are then fused using pointwise convolutions to learn dependencies between the layers of variable scales [24]. Furthermore, each convolutional block in the architecture comprised a squeeze-and-excitation (S&E) adaptive calibration mechanism at the very end of the block [25]. This method was chosen for its ability to increase generalizability and improve the performance of state-of-the-art convolutional neural networks [25].

The final output of the model consisted of softmax values i.e., the probability scores for each of the five sleep stages on a given temporal resolution (30 seconds in this study). The label with the highest softmax value was included in the final hypnogram. Implementation and modifications of the model's architecture were conducted in Python, version 3.8.5, with TensorFlow 2.6.0 and Keras application programming interface.

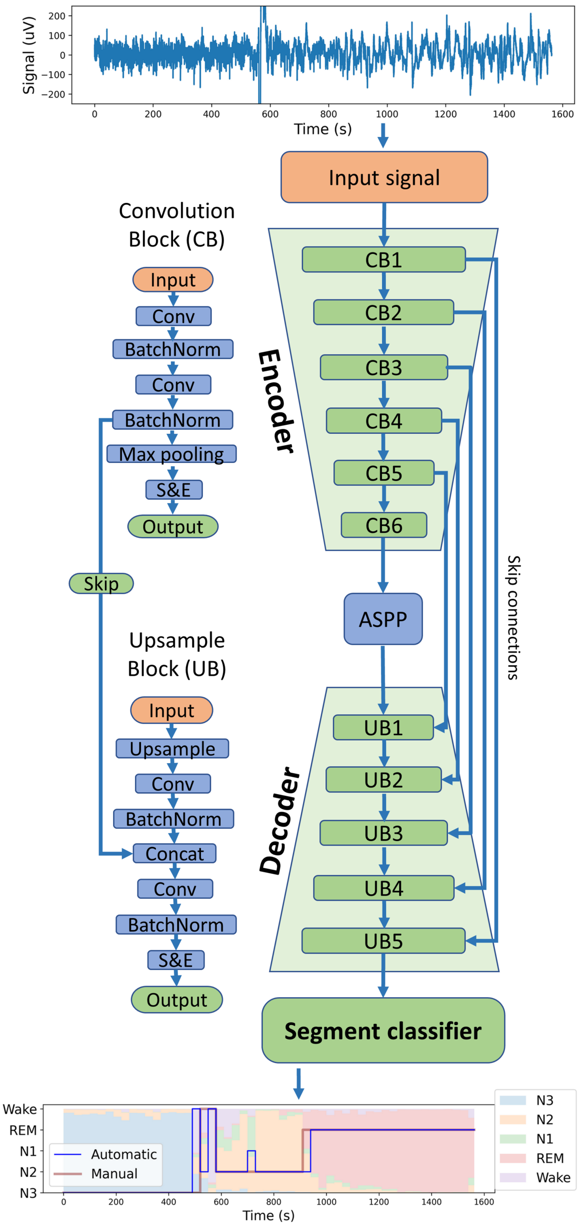

Fig. 4. Neural network architecture for automatic sleep staging method visualized as a schematic diagram. See details in the text. Abbreviations: ASPP = atrous spatial pyramid pooling, Conv = convolutional layer, Concat = concatenate, BatchNorm = batch normalization, S&E = squeeze-and-excitation, REM = rapid eye movement sleep, N1 = Stage N1 sleep, N2 = Stage N2 sleep, N3 = Stage N3 sleep.

C. Neural Network Training

The clinical dataset was divided patient-wise into training, validation, and test sets. First, a sample of 88 PSGs (10% of all) was randomly chosen as a completely separate test set. Then a validation set ($n = 79$) was chosen randomly from the remaining 788 PSGs. Finally, the remaining 709 recordings were used in the training of the model. The demographic information of each of the datasets is shown in Table I.

As our goal was to generalize the model for the headband-recorded forehead signals, we used single-channel EOG (E1-M2) signals in the training process. This decision was based on previous findings of the high similarity of the textile electrode-recorded forehead EEG signals and the simultaneously recorded EOG signals [11]. Moreover, we first tested to generalize the model on clinical EEG data, and then further on the headband-recorded forehead EEG. The preprocessing steps applied to all signals before training and testing the model were the following: re-sampling to 64 Hz, high pass filtering with a 0.3 Hz cutoff, and standardization with an interquartile range. A fifth-order type II Chebyshev high pass filter with a 40 dB minimum attenuation in the stopband was chosen for filtering to allow a smooth passband for the filtered signals. Patient-wise interquartile range standardization i.e., subtracting the median of the signal and dividing by the 25–75% interquartile range, was utilized to exclude the effect of high-amplitude artifacts on the signal standardization process.

Stochastic gradient descent with momentum and decoupled weight decay regularization was used as the optimization method for the neural network's weights [26]. Categorical cross-entropy was utilized as the objective function to be minimized in the training process. A disciplined approach for hyper-parameter selection was adopted following the method described in the literature [27]. This method was used for selecting the optimal learning rate, batch size, momentum, and weight decay for well-balanced training. Setting the initial and maximum learning rates, between which the disciplined approach was used, was based on a learning rate range test [27]. This method resulted in 0.05 and 0.5 as the initial and maximum learning rates, respectively. Furthermore, the network was trained for 100 training epochs, each including iteration over the whole training data. The recordings were shuffled and randomly sampled into smaller batches between the epochs to avoid local minima during weight optimization. For iterating over each epoch, batches of eight recordings were chosen from randomly shuffled PSGs until the whole set was used. From each recording, a two-hour length segment, starting on a randomly chosen 30-second epoch, was fed to the network at a time.

D. Performance Evaluation

The performance of the automatic model was evaluated in both datasets against the manual sleep staging. To allow comparison with other studies, commonly used agreement metrics were computed; accuracy, F1-scores, Cohen's kappa (κ), precision, and recall. Both, the overall metrics as well as the sleep stage-specific metrics for 5-stage classification were computed. For analyzing the classification results in a more detailed manner, we computed confusion matrices between the compared methods. In addition, performance metrics (accuracy, F1-score, and κ) were computed in a subject-by-subject manner for the home recordings to give more insight into the variability of the performance metrics between subjects. Finally, Bland-Altman plots were utilized to compare widely used sleep parameters (total sleep time, sleep latency, and wake after sleep onset (WASO)) values derived from automatically versus manually conducted sleep staging for the home recordings.

III. Results

A. Model Performance in Clinical Dataset

The model's overall accuracy in 5-stage classification was 80% ($\kappa = 0.73$, F1-score = 0.74) based on standard single-channel EOG in the test set of the clinical dataset (Table II).

TABLE II
OVERALL RESULTS AND SLEEP STAGE SPECIFIC METRICS FOR AUTOMATIC SLEEP STAGING ON EOG-BASED MODEL IN THE TEST SET ($N = 88$) OF THE CLINICAL DATA

Signal	Sleep Stage	Precision	Recall	F1-score	Support
Standard EOG (E1-M2)	Wake	0.86	0.90	0.88	26058
	N1	0.58	0.26	0.36	7507
	N2	0.76	0.82	0.79	25390
	N3	0.82	0.78	0.80	9681
	REM	0.83	0.93	0.88	9748
	Macro average	0.77	0.74	0.74	78384
	Cohen's kappa		0.73		
	Accuracy		80%		
Standard EEG (C4-M1)	Wake	0.94	0.82	0.88	26058
	N1	0.46	0.47	0.47	7507
	N2	0.67	0.85	0.75	25390
	N3	0.79	0.81	0.80	9681
	REM	0.93	0.55	0.69	9748
	Macro average	0.76	0.70	0.72	78384
	Cohen's kappa		0.68		
	Accuracy		76%		

The metrics are computed against the manual sleep staging. Macro average is the unweighted average over the five classes. Standard EOG and EEG were recorded using medical-grade electrodes in part of diagnostic PSG recording. Abbreviations: EEG = electroencephalography, EOG = electrooculography, Wake = wakefulness, N1 = Stage N1 sleep, N2 = Stage N2 sleep, N3 = Stage N3 sleep, REM = rapid eye movement sleep.

From all sleep stages, Wake and REM were detected with the best performance (F1-scores = 0.88) using the automatic method. Stages N2 and N3 were also classified with good performance i.e., with F1-scores of 0.79 and 0.80, respectively. The model detected Stage N1 with a moderate precision (0.58), but with a poor recall (0.26), leading to an F1-score of 0.36.

The model's accuracy was 76% ($\kappa = 0.68$, F1-score = 0.72) when tested in a direct transfer manner using standard EEG (C4-M1) (Table II). In particular, the model detected Wake, N2, and N3 sleep with good performance (F1-scores > 0.75), and N1 and REM with moderate performance (F1-scores 0.47 and 0.69, respectively).

B. Model Performance on Home Recordings

1) Forehead EEG – Textile Electrode Headband: Compared to manual sleep staging, the automatic model reached an 82% ($\kappa = 0.75$, F1-score = 0.72) overall accuracy on 5-stage classification based on the textile electrode headband (Table III). Most of the uncertainties in the automatic scoring of the sleep stages were related to Stage N1 misclassification (Fig. 5). The model's performance on the detection of Stage N1 sleep was poor (F1-score = 0.27), however, the model performed well in detecting all other sleep stages: Wake, N2, N3, and REM (F1-scores > 0.80) using the forehead EEG.

Subject-by-subject computed performance metrics showed that the kappa values of the automatic sleep stage prediction were on a high level ($\kappa > 0.78$) for most of the subjects (subjects #1, #2, #3, #4, #5, #8) when using the forehead EEG as an input (Table IV). Accuracy was over 70% ($\kappa > 0.58$) for all subjects with successful recordings (9/10), and 67% ($\kappa = 0.54$) for subject #10 who reported electrode movement due to excessive loosening of the headband during the night of recording. Compared to standard EOG, the utilization of forehead EEG signals as an input led to better results on one subject (subject #2, Table IV). The mean difference (± 1.96 standard deviations) between manually and automatically (from the forehead EEG) derived total sleep time was -2.65 (-23.73–18.43) minutes. Similarly, the mean difference in sleep latency was 8.8 (-10.96–28.56) minutes, and in the amount of wakefulness after sleep onset (WASO), it was -6.15 (-37.23–24.93) minutes.

2) Standard EOG – Clinical Electrodes: When using the standard EOG signal from the home recordings, the model's overall accuracy was 87% ($\kappa = 0.82$, F1-score = 0.78) for 5 sleep stages (Table III). The F1-score of 0.36 for predicting Stage N1 sleep was again the lowest of the stage-specific results. However, the agreement between the automatic and manual classification of other sleep stages was on a high level (F1-scores > 0.86) using the standard EOG as an input.

Subject-by-subject computed metrics showed that the kappa values were 0.79 or higher for 8 subjects, and 0.75 and 0.69 for subjects #6 and #9, respectively (Table IV). The mean differences in the automatically and manually derived sleep parameters were also computed for the standard EOG signals (Fig. 6). The mean difference in total sleep time was -3.5 (-25.91–18.91) minutes, in sleep latency it was 11.65 (-60.17–83.47) minutes, and in the amount of WASO, -8.15 (-74.73–58.43) minutes.

IV. DISCUSSION

The aim of the study was to develop and test a deep learning-based sleep staging method based on a wearable forehead EEG headband. Due to a limited number of home recordings collected for testing of the headband, we used a separate dataset i.e., diagnostic PSG recordings for training the deep learning model. Standard EOGs from the PSG recordings were first used in the training of the network, leading to 80% ($\kappa = 0.73$) accuracy in sleep staging in the test set of the clinical dataset against manual sleep staging.

The trained fully convolutional neural network generalized well for the home recordings; the model achieved potential performance using the headband-recorded forehead EEG (accuracy = 82%, $\kappa = 0.75$) and standard EOG (accuracy = 87%, $\kappa = 0.82$) against manual sleep staging in young healthy adults. Overall, the achieved accuracies surpassed those of non-EEG-based wearables [7], [28].

A. Performance of the Model

The performance of the automatic sleep staging model was on par with the inter-rater agreement of manual sleep stage scoring (κ between 0.71–0.81) [29]. As deep learning methods rely on manual annotations, the performance of the proposed sleep staging model is somewhat limited to the inter-rater agreement

TABLE III
OVERALL RESULTS AND SLEEP STAGE SPECIFIC METRICS FOR AUTOMATIC SLEEP STAGING ON HOME RECORDINGS ($N = 10$)

	Forehead EEG				Standard EOG			
	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
Wake	0.83	0.80	0.81	1434	0.88	0.83	0.86	1434
N1	0.35	0.21	0.27	215	0.34	0.39	0.36	215
N2	0.79	0.86	0.82	3814	0.88	0.87	0.87	3814
N3	0.89	0.77	0.83	2020	0.94	0.85	0.89	2020
REM	0.85	0.86	0.86	1854	0.85	0.97	0.91	1854
Macro average	0.74	0.70	0.72	9337	0.78	0.78	0.78	9337
Cohen's kappa	0.75				0.82			
Accuracy	82%				87%			

The metrics are computed against the manual sleep staging. Macro average is the unweighted average over all sleep stages. Forehead EEG (Fp1-Fp2) was recorded using the textile electrode headband and standard EOG (E1-M2) using medical-grade wet electrodes. Abbreviations: EEG = electroencephalography, EOG = electrooculography, N1 = Stage N1 sleep, N2 = Stage N2 sleep, N3 = Stage N3 sleep, REM = rapid eye movement sleep.

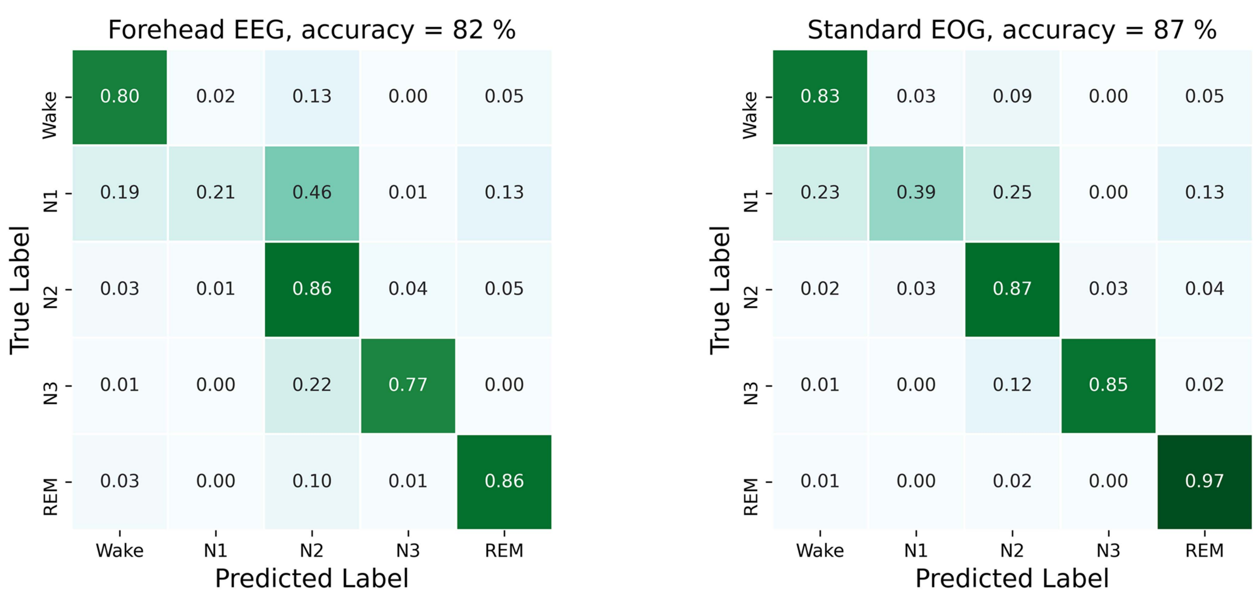

Fig. 5. Confusion matrixes between the manual sleep staging (true label) and automatic sleep staging (predicted label) on home recordings. Forehead EEG (Fp1-Fp2) was recorded with the textile electrode headband and standard EOG (E1-M2) using medical-grade wet electrodes. Abbreviations: EEG = electroencephalography, EOG = electrooculography, Wake = wakefulness, N1 = Stage N1 sleep, N2 = Stage N2 sleep, N3 = Stage N3 sleep, REM = rapid eye movement sleep.

of manual scorers [30]. However, our model is trained by sleep annotations from several individual scorers. Following this, a comparison against an individual scorer might not be the best practice, and using the consensus of multiple scorers as the true labels of the sleep stages would be more optimal for performance evaluation [31]. On datasets utilized in this study, each recording was manually analyzed only once, excluding the possibility of the aforementioned inspection. Nevertheless, the current results illustrate a similar performance of automatic sleep staging compared to manual sleep staging.

The automatic sleep staging model presented in this study performed on a similar level as other state-of-the-art automatic methods utilizing only single-channel EOG signal (accuracies between 76%–91%) [32], [33], [34]. It also reached similar F1-scores as what has been reported with another U-time-derived network i.e., the U-sleep utilizing one EOG channel combined with one forehead EEG channel (F1-scores = 0.76–0.77) [15]. Moreover, our model surpassed another state-of-the-art network (*SeqSleepNet*) in direct transfer without any finetuning, tested between EEG and EOG channel mismatches (accuracies 81%

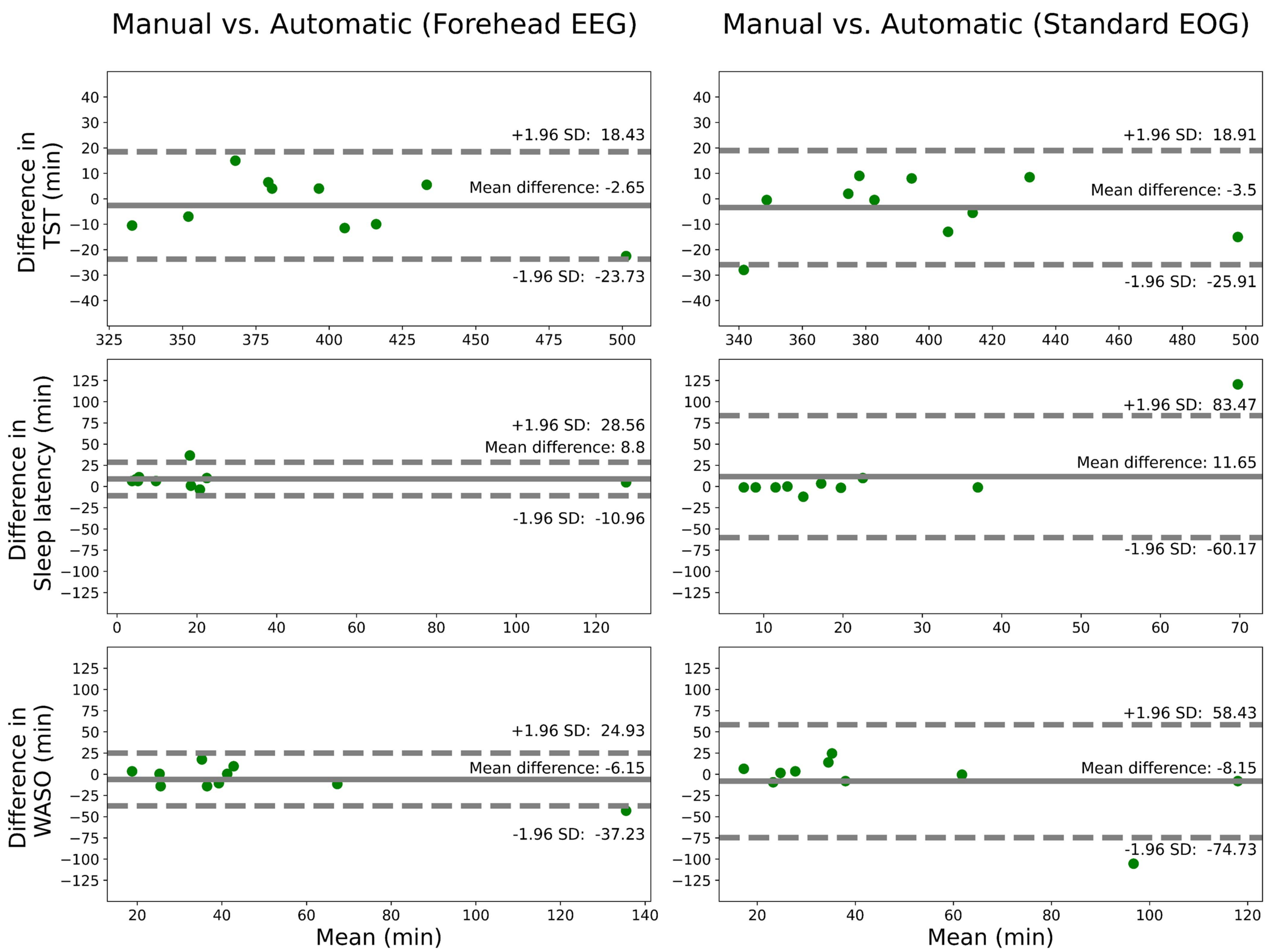

Fig. 6. Mean-difference plots of the selected sleep parameters. Parameters are computed using hypnograms derived from automatic sleep staging, which was based on forehead EEG (Fp1-Fp2) recorded using the textile electrode headband or standard EOG (E1-M2) signal recorded using medical-grade wet electrodes, and those derived from manual sleep staging. The differences are computed subject-wise ($n = 10$) by subtracting the automatically derived parameter from the manually derived one. Abbreviations: TST = total sleep time, WASO = wake after sleep onset, EEG = electroencephalography, EOG = electrooculography, SD = standard deviation.

and 52% for EEG->EEG and EEG->EOG channel mismatches, respectively) [23].

B. Generalizability of the Model

The current results show the generalizability of the developed automatic sleep staging method from EOG to EEG channel and to a different recording environment as well as with the headband-recorded signals. Contrary to the general trend reported in the literature [17], [18], [22], [23], [31], [35], [36] the performance of the neural network seems to increase when evaluated across a separate test set. This increase is, however, most likely due to only having ten healthy subjects in the home recordings; a performance drop more similar to that observed in the literature would be most probably seen with a larger and more diverse population. The subjects in the home recording dataset had, on average, notably less N1 sleep, less WASO, and higher sleep efficiency (Table I). Therefore, as these factors usually affect positively the performance of automatic sleep staging solutions [37], this must be acknowledged when interpreting the results against those achieved in the clinical populations with suspected sleep disorders. When generalizability was tested in the clinical population from EOG to EEG channel, the performance drop (κ decreases from 0.73 to 0.68) was more consistent with the trend in the literature. In addition, [35] and [36] use z-score normalization for preprocessing (subtract mean and divide by standard deviation). In our experiments, we noticed similar degradation in performance when using z-score normalization. However, when we used medians and interquartile ranges to normalize the signals, the performance on cross-domain predictions remained high. We believe this is related to the effect of high-amplitude artifacts, which can drastically alter the mean and standard deviation, but have only a minor effect on the median and interquartile range.

Although the model shows considerable performance in generalizability from EOG to EEG signals, it must be noted that this might be related to recording settings. The underlying physiological phenomena behind EOG and EEG are different, and if these would be recorded in an ideally isolated setting or separated later with signal processing, the model might be unable

TABLE IV
SUBJECT-BY-SUBJECT PERFORMANCE METRICS OF THE AUTOMATIC SLEEP STAGING ON HOME RECORDINGS

Signal	Subject	F1-score	Cohen's kappa	Accuracy
Forehead EEG	#1	0.76	0.83	89%
	#2	0.76	0.83	88%
	#3	0.69	0.81	86%
	#4	0.71	0.78	85%
	#5	0.82	0.83	88%
	#6	0.64	0.58	71%
	#7	0.62	0.68	78%
	#8	0.72	0.80	86%
	#9	0.70	0.68	77%
	#10	0.54	0.54	67%
Standard EOG	#1	0.81	0.89	93%
	#2	0.78	0.80	86%
	#3	0.77	0.87	90%
	#4	0.77	0.86	90%
	#5	0.80	0.86	90%
	#6	0.74	0.75	82%
	#7	0.79	0.79	85%
	#8	0.77	0.83	88%
	#9	0.67	0.69	78%
	#10	0.75	0.80	86%

The metrics are computed between the automatically predicted sleep stages and manual sleep staging. Forehead EEG was recorded with the textile electrode headband and standard EOG with medical-grade wet electrodes. Abbreviations: EEG = electroencephalography, EOG = electrooculography.

to generalize between the two signals. However, in practical settings where surface electrodes are used, neural as well as corneo-retinal activities are both picked up in the recordings, enabling cross-channel predictions.

C. Textile Electrode Headband

The performance of our wearable headband-based automatic sleep staging model is not directly comparable to previous studies on wearable devices due to differences in the test set demographics, recording environments, and lack of comparable performance metrics for the automatic scoring method. However, by acknowledging these differences, comparable solutions can be found in the literature. One of the earliest self-applicable EEG recording devices, called QUISI, already utilized a neural network with only two hidden layers for automatic sleep staging, and with relatively promising results [38]. More recently, Casciola et al. validated a Cognionics 2-channel EEG headband (Cognionics, San Diego, CA, USA) combined with a deep learning-based sleep staging approach, which resulted in 74% sleep staging accuracy on the headband-based in-lab recordings [39]. Another wearable headband called The Dreem Headband (Dreem, Paris, France) was reported to reach 84% $\pm$ 6% ($\kappa = 0.75 \pm 0.10$, $n = 25$) accuracy against the consensus of five manual scorers with its integrated automatic sleep staging algorithm [40]. Advanced Brain Monitoring (Carlsbad, California, United States) also has a multichannel frontopolar EEG device called Sleep Profiler. It has been tested in home environment for sleep-disordered patients, and has reached a kappa value of 0.67 between automatic and manual sleep staging [41]. A similar headband solution utilizing silicone-based dry sensors was introduced by Lin et al. [42], reporting an overall accuracy of 77% ($\kappa = 0.69$, $n = 10$) in automatic sleep staging against the manual one.

In addition to headbands, textile electrodes have been integrated into sleep eye-masks for automatic sleep staging [43], [44]. The eye-mask solution with eight healthy participants has been reported to result in 87% accuracy against manual 4-stage sleep staging [44]. A different design is utilized in in-ear EEG devices that have been tested in automatic sleep staging, reaching an accuracy of 74% [45] and kappa values of 0.61 and 0.73 on healthy subjects [45], [46]. Furthermore, Popovic et al. [47] have validated the use of forehead EEG (Fp1-Fp2) channel in automatic sleep staging on healthy subjects reaching an overall accuracy of 81% ($\kappa = 0.75$) in a test set comprising nap recordings and nocturnal PSG recordings from a controlled laboratory environment. In conclusion, our method performs on a higher or similar level as the previously proposed portable and user-friendly systems for automatic sleep staging. Furthermore, the headband is suitable for self-application and for measuring over several nights, overcoming the limitations of clinically used standard EEG electrodes.

The headband used in this study can be attached to any portable PSG recording device that utilizes standard 1.5 mm (DIN 42 802) touch-proof sockets, making it a universal solution for many users. A similar idea has been earlier adopted with some self-applicable electrode sets proposed for ambulatory sleep recordings [48], [49]. An accuracy of 76% ($\kappa = 0.66$) in the manual inter-rater agreement has been reported for a novel flexible electrode set produced using advanced screen-printing techniques [48]. Although this electrode set expresses a more comprehensive facial montage compared to the headband utilized in the present study, for example for the detection of sleep bruxism events in addition to sleep staging [50], it comprises self-adhesive, hydrogel-coated silver/silver chloride electrodes that cannot be reused. In another study, the applicability of knitted silver-coated textile electrodes (DryodeTM, IDUN Technologies, Zürich, Switzerland) on nocturnal in-home recordings were tested with five healthy subjects [51]. They reported inter-scorer reliability of $\kappa = 0.66$ (accuracy = 78%) in manual scoring between the textile-based and standard electrodes.

Our sleep staging model's accuracy on home recordings was around 5% lower using the headband-recorded forehead EEG compared to using the standard EOG (82% vs. 87%) as an input signal. Similarly, sleep stage-specific F1-scores were systematically higher (0.05 units or more) for the standard EOG predictions. The decreased performance, when comparing forehead EEG and standard EOG, can result from a lower signal quality of the textile electrode-recorded signals or generalizability of the model from standard EOG to forehead EEG. Therefore, it might be possible to further increase the performance for the forehead EEG headband, using signals recorded from the exact positions of the textile electrodes (Fp1, Fpz, Fp2) in the training of the neural network. In our clinical dataset, these signals were not available for testing this approach. However, both of the used signals enable reliable automatic sleep staging, and when considering the benefits of the headband, the decrease in performance between the forehead EEG and standard EOG is acceptable.

Sleep stage-specific results for the automatic method showed that Stage N1 sleep produced most misclassifications relative to the number of epochs. This was expected due to the lower amount of Stage N1 sleep compared to other sleep stages in all datasets. Furthermore, demographic information of the datasets shows a lower amount of Stage N1 sleep for home recordings of healthy subjects compared to training data (medians: 2.7% vs. 11.1% of TST). Moreover, only fair inter-rater reliability ($\kappa \approx 0.24$) is reached also among expert human scorers in the detection of Stage N1 sleep [29]. As we had separate scorers from different sleep laboratories i.e., clinical PSGs were manually analyzed at the Sleep Disorders Center, Princess Alexandra Hospital (Brisbane, Australia) by several scorers, and home recordings were analyzed at the Reykjavik University Sleep Institute (Reykjavik, Iceland) by one scorer, there are most probably differences in practices of scoring Stage N1 sleep between these two institutions/scorers as well, which is a more general issue [52]. These issues are most probably reflected in the agreement of Stage N1 classification between the automatic and manual methods.

Performance metrics computed in a subject-by-subject manner allowed interpreting the variance in the scorings of individual subjects. The results showed that the automatic sleep staging method combined with the headband-recorded forehead EEG performed well on most of the subjects. However, notably lower performance was found in two of the subjects (#6 and #10). For subject #6, the automatic method predicted an excess amount of Stage N2 sleep. Most likely, subject #6 was hard to analyze because of the subject's movement and recurring sleep-wake transitions. Subject #10 reported movement of the headband after loosening the Velcro strap too much during the recording night. This led to textile electrode movement, which induced artifacts in the recorded forehead EEG signals. Therefore, the latter half of the recorded night was at some points not technically analyzable based on the forehead EEG but all manually scorable epochs of the recording were included in the analysis to have a fair comparison against manual sleep staging. Similarly, as for subject #6, the automatic method classified other sleep stages incorrectly into Stage N2 sleep. These are probably caused by model uncertainty, while Stage N2 is the majority class in the training data; when the deep learning model is uncertain about the predictions, it tends to output higher softmax scores for the majority class, since this policy minimizes the categorical cross-entropy loss used during training. In these cases, the automatic sleep staging method would be more interpretable with uncertainty quantification, which has already been implemented for example in [13]. Due to movement-related misclassifications, the automatic method might benefit from additional actigraphy-like input or continuous skin-electrode contact impedance monitoring. Moreover, artifact detection could be beneficial in some extreme cases but is not yet widely implemented alongside automatic sleep staging [17].

Sleep is often quantified with hypnogram-derived parameters, such as the TST, sleep latency, and WASO. Thus, the automatic sleep staging method needs to produce also reliable estimation of these measures. Therefore, we compared the aforementioned parameters derived from automatic versus manually scored hypnograms. Although biases on these parameters were minimal, slight differences between automatic and manual methods can be seen. On average, the automatic method predicts shorter sleep latency and more wake after sleep onset. This might indicate that automatic sleep staging does not have as strong a smoothing effect as manual sleep staging tends to have for hypnograms i.e., overlooking sudden sleep-wake transitions [53]. Only one subject (#2) represented automatically predicted sleep parameters significantly differing from the manually derived ones. These were related to sleep stage predictions from standard EOG signals, not from textile electrode-recorded forehead EEG signals, and can be seen as individual outliers in the mean-difference plots of sleep latency and WASO (Fig. 6). It is worth noting, that even a misclassification of a single epoch at the beginning of the recording can have a significant effect on these two parameters e.g., if there is a long period of wakefulness at the start of the recording. On subject #2, these differences originated from this type of minor misclassification, and the hypnogram was otherwise predicted with good overall performance ($\kappa = 0.80$, Table IV). Apart from those results, TST was quantified with excellent accuracy using the automatic method. Therefore, this method could be used to easily increase the accuracy of HSAT, in which total recording time is currently used as the estimate of TST [54].

D. Limitations

A major limitation of the present study is that the home recordings comprised only a limited set of healthy subjects. Automatic sleep staging methods tend to perform better when analyzing healthy individuals compared to analyzing sleep-disordered patients [37]. One reason for this is sleep fragmentation, which makes sleep staging more difficult with patients suffering from sleep-disordered breathing (SDB) and consequential arousals [52]. However, we believe the neural network model has the potential to generalize also to patients suffering from SDB as the model has been trained on diagnostic recordings of suspected OSA patients. The model already had good performance (accuracy = 80%, $\kappa = 0.73$, F1-score = 0.74) in the test set of the clinical data (median AHI = 19.0 events/hour) using only the single-channel EOG as an input. In addition, the performance in the test set was also good when EEG signals were used to test instead of EOG (accuracy = 76%, $\kappa = 0.68$, F1-score = 0.72). This shows that the model can at least generalize from EOG to EEG channel in the clinical population. Moreover, sleep efficiency is usually better for the same subject in home-based PSG compared to in-lab one [55], making it generally easier for the model to analyze home-based recordings. Moreover, a practical limitation of the method in medical use could be disinfection and washing of the headband. It has not been tested how long-term repetitive disinfection or washing would affect the textile electrodes and therefore the quality of the recording. Finally, another problem could arise from nocturnal sweating, which is a common symptom among SDB patients [56]. Although textile electrodes are considered to require some moisture and sweat in the formation and functioning of the skin-electrode interface [10], excessive sweating might induce unwanted potential fluctuations disturbing the recorded signals. Therefore, a follow-up study of the headband with SDB patients needs to be conducted in the future to continue the validation of this method.

In the present study, the textile electrode headband, FocusBand, was attached to a portable PSG device. Therefore, there is no evidence of how the FocusBand works with its integrated recording device in sleep recordings. The current results only show potential in generalizability to Nox A1 devices, and e.g., possibly differing analog prefiltering settings, as well as other hardware specifications, can affect the generalizability to different devices. The analog pre-filtering settings of the devices used in this study were not disclosed by manufacturers. The integrated recording device makes the headband wireless and could therefore be a more comfortable option for sleep recordings as well. With this option, the headband could also be used independently from the chest-mounted PSG if other PSG signals are not needed. However, the option to use the headband with almost any portable EEG/PSG device is a valuable feature. That is because the headband could easily be used to substitute standard EEG electrodes in various research and diagnostic measurements, where a reliable and objective assessment of sleep structure outside the sleep laboratory is needed. In addition, recording over multiple consecutive nights without visiting the sleep laboratory between the nights could be possible due to the reusability of the headband.

V. Conclusion

The developed deep learning-based sleep staging model generalized from clinical studies to textile-based forehead EEG. The model enables reliable, automatic assessment of sleep structure and shows potential to quantify sleep structure in home-based sleep recordings with a reusable and self-applied textile electrode headband on similar reliability as what is achieved with in-lab attended PSG and manual sleep staging.

Acknowledgment

The authors would like to thank Brett Duce at the Princess Alexandra Hospital, Brisbane, Australia for data collection and research collaboration. His important and highly professional work at the Sleep Disorders Center serves as the basis of the present study. FocusBand Technologies has supported this work by providing measuring equipment free of charge and through a consultation agreement with The University of Queensland. Authors claim no personal conflict of interest related to FocusBand Technologies.

References

[1] J. Corral-Peñafiel, J. L. Pepin, and F. Barbe, "Ambulatory monitoring in the diagnosis and management of obstructive sleep apnoea syndrome," *Eur. Respir. Rev.*, vol. 22, no. 129, pp. 312–324, 2013, doi: 10.1183/09059180.00004213.

[2] H. Korkalainen et al., "Self-applied home sleep recordings: The future of sleep medicine," *Sleep Med. Clin.*, vol. 16, no. 4, pp. 545–556, 2021, doi: 10.1016/j.jsmc.2021.07.003.

[3] J. Fischer et al., "Standard procedures for adults in accredited sleep medicine centres in Europe," *J. Sleep Res.*, vol. 21, no. 4, pp. 357–368, 2012, doi: 10.1111/j.1365-2869.2011.00987.x.

[4] N. A. Collop, "Portable monitoring for the diagnosis of obstructive sleep apnea," *Curr. Opin. Pulmonary Med.*, vol. 14, no. 6, pp. 525–529, 2008, doi: 10.1097/MCP.0b013e328312ed4a.

[5] N. A. Collop et al., "Clinical guidelines for the use of unattended portable monitors in the diagnosis of obstructive sleep apnea in adult patients," *J. Clin. Sleep Med.*, vol. 3, no. 7, pp. 737–747, 2007.

[6] R. B. Berry et al., "AASM manual for the scoring of sleep and associated events," *Amer. Acad. Sleep Med. Darien IL, USA*, 2018, doi: 10.1016/j.carbon.2012.07.027.

[7] S. A. Imtiaz, "A systematic review of sensing technologies for wearable sleep staging," *Sensors*, vol. 21, no. 5, 2021, Art. no. 1562, doi: 10.3390/s21051562.

[8] M. Bruyneel and V. Ninane, "Unattended home-based polysomnography for sleep disordered breathing: Current concepts and perspectives," *Sleep Med. Rev.*, vol. 18, no. 4, pp. 341–347, 2014, doi: 10.1016/j.smrv.2013.12.002.

[9] M. D. Ghegan, P. C. Angelos, A. C. Stonebraker, and M. B. Gillespie, "Laboratory versus portable sleep studies: A meta-analysis," *Laryngoscope*, vol. 116, no. 6, pp. 859–864, 2006, doi: 10.1097/01.mlg.0000214866.32050.2e.

[10] M. Rusanen et al., "An in-laboratory comparison of focusband EEG device and textile electrodes against a medical-grade system and wet gel electrodes," *IEEE Access*, vol. 9, pp. 132580–132591, 2021, doi: 10.1109/access.2021.3113049.

[11] M. Rusanen et al., "Technical performance of textile-based dry forehead electrodes compared with medical-grade overnight home sleep recordings," *IEEE Access*, vol. 9, pp. 157902–157915, 2021, doi: 10.1109/access.2021.3128057.

[12] L. Fiorillo et al., "Automated sleep scoring: A review of the latest approaches," *Sleep Med. Rev.*, vol. 48, 2019, Art. no. 101204, doi: 10.1016/j.smrv.2019.07.007.

[13] H. Phan et al., "Sleep transformer: Automatic sleep staging with interpretability and uncertainty quantification," 2021. [Online]. Available: arxiv.org/abs/2105.11043

[14] M. Perslev, M. H. Jensen, S. Darkner, P. J. Jennum, and C. Igel, "U-Time: A fully convolutional network for time series segmentation applied to sleep staging," 2019. [Online]. Available: arxiv.org/abs/1910.11162

[15] M. Perslev et al., "U-Sleep: Resilient high-frequency sleep staging," *NPJ Digit. Med.*, vol. 4, no. 1, pp. 1–12, 2021, doi: 10.1038/s41746-021-00440-5.

[16] H. Korkalainen et al., "Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 7, pp. 2073–2081, Jul. 2020, doi: 10.1109/jbhi.2019.2951346.

[17] H. Phan and K. Mikkelsen, "Automatic sleep staging of EEG signals: Recent development, challenges, and future directions," *Physiol. Meas.*, vol. 43, no. 4, 2022, Art. no. 04TR01, doi: 10.1088/1361-6579/ac6049.

[18] H. Phan et al., "Towards more accurate automatic sleep staging via deep transfer learning," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 6, pp. 1787–1798, Jun. 2021, doi: 10.1109/TBME.2020.3020381.

[19] M. Radha et al., "A deep transfer learning approach for wearable sleep stage classification with photoplethysmography," *NPJ Digit. Med.*, vol. 4, no. 135, 2021, Art. no. 135, doi: 10.1038/s41746-021-00510-8.

[20] R. Saeedi and A. H. Gebremedhin, "A signal-level transfer learning framework for autonomous reconfiguration of wearable systems," *IEEE Trans. Mobile Comput.*, vol. 19, no. 3, pp. 513–527, Mar. 2020, doi: 10.1109/TMC.2018.2878673.

[21] Q. Li et al., "Transfer learning from ECG to PPG for improved sleep staging from wrist-worn wearables," *Physiol. Meas.*, vol. 42, no. 4, 2021, Art. no. 044004, doi: 10.1088/1361-6579/abf1b0.

[22] A. Guillot and V. Thorey, "RobustSleepNet : Transfer learning for automated sleep staging at scale," 2021. [Online]. Available: arxiv.org/abs/2101.02452

[23] H. Phan et al., "Deep transfer learning for single-channel automatic sleep staging with channel mismatch," in *Proc. 27th Eur. Signal Process. Conf.*, 2019, pp. 1–5, doi: 10.23919/EUSIPCO.2019.8902977.

[24] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/TPAMI.2017.2699184.

[25] J. Hu et al., "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020, doi: 10.1109/TPAMI.2019.2913372.

[26] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019. [Online]. Available: arxiv.org/abs/1711.05101

[27] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1-learning rate, batch size, momentum, and weight decay," 2016. [Online]. Available: arxiv.org/abs/1803.09820

[28] H. Scott, L. Lack, and N. Lovato, "A systematic review of the accuracy of sleep wearable devices for estimating sleep onset," *Sleep Med. Rev.*, vol. 49, 2020, Art. no. 101227, doi: 10.1016/j.smrv.2019.101227.

[29] Y. J. Lee et al., "Inter-rater reliability of sleep stage scoring: A meta-analysis," *J. Clin. Sleep Med.*, vol. 18, no. 1, pp. 193–202, 2021, doi: 10.5664/jcsm.9538.

[30] T. Penzel, X. Zhang, and I. Fietze, "Inter-scorer reliability between sleep centers can teach us what to improve in the scoring rules," *J. Clin. Sleep Med.*, vol. 09, no. 1, pp. 89–91, 2013, doi: 10.5664/jcsm.2352.

[31] A. Guillot, F. Sauvet, E. H. During, and V. Thorey, "Dreem open datasets: Multi-scored sleep datasets to compare human and automated sleep staging," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 9, pp. 1955–1965, Sep. 2020, doi: 10.1109/TNSRE.2020.3011181.

[32] J. Fan et al., "EOGNET : A novel deep learning model for sleep stage classification based on single-channel EOG signal," *Front. Neurosci.*, vol. 15, no. 7, 2021, Art. no. 573194, doi: 10.3389/fnins.2021.573194.

[33] A. N. Olesen, J. A. E. Christensen, H. B. D. Sorensen, and P. J. Jennum, "A noise-assisted data analysis method for automatic EOG-based sleep stage classification using ensemble learning," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2016, pp. 3769–3772, doi: 10.1109/EMBC.2016.7591548.

[34] M. M. Rahman, M. I. H. Bhuiyan, and A. R. Hassan, "Sleep stage classification using single-channel EOG," *Comput. Biol. Med.*, vol. 102, no. 11, pp. 211–220, 2018, doi: 10.1016/j.compbiomed.2018.08.022.

[35] D. Alvarez-Estevez and R. M. Rijsman, "Inter-database validation of a deep learning approach for automatic sleep scoring," *PLoS One*, vol. 16, no. 8, 2021, Art. no. e0256111, doi: 10.1371/journal.pone.0256111.

[36] A. N. Olesen et al., "Automatic sleep stage classification with deep residual networks in a mixed-cohort setting," *Sleep*, vol. 44, no. 1, 2021, Art. no. zsaa161, doi: 10.1093/sleep/zsaa161.

[37] R. Boostani, F. Karimzadeh, and M. Nami, "A comparative review on sleep stage classification methods in patients and healthy individuals," *Comput. Methods Programs Biomed.*, vol. 140, no. 3, pp. 77–91, 2017, doi: 10.1016/j.cmpb.2016.12.004.

[38] I. Ehlert et al., "A comparison between EEG-recording and scoring by QUISI version 1.0 and standard PSG with visual scoring-a one-channel ambulatory EEG recording device using neural network techniques for automatic sleep stage classification," *Somnologie*, vol. 2, no. 3, pp. 104–116, 1998, doi: 10.1007/s11818-998-0015-y.

[39] A. A. Casciola et al., "A deep learning strategy for automatic sleep staging based on two-channel EEG headband data," *Sensors*, vol. 21, no. 10, 2021, Art. no. 3316, doi: 10.3390/s21103316.

[40] P. J. Arnal et al., "The dreem headband compared to polysomnography for electroencephalographic signal acquisition and sleep staging," *Sleep*, vol. 43, no. 11, 2020, Art. no. zsaa097, doi: 10.1093/sleep/zsaa097.

[41] D. J. Levendowski et al., "The accuracy, night-to-night variability, and stability of frontopolar sleep electroencephalography biomarkers," *J. Clin. Sleep Med.*, vol. 13, no. 6, pp. 791–803, 2017, doi: 10.5664/jcsm.6618.

[42] C.-T. Lin et al., "Forehead EEG in support of future feasible personal healthcare solutions: Sleep management, headache prevention, and depression treatment," *IEEE Access*, vol. 5, pp. 10612–10621, 2017, doi: 10.1109/ACCESS.2017.2675884.

[43] S.-F. Liang et al., "Development of an EOG-based automatic sleep-monitoring eye mask," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 11, pp. 2977–2985, Nov. 2015, doi: 10.1109/TIM.2015.2433652.

[44] T.-H. Hsieh et al., "Home-use and real-time sleep-staging system based on eye masks and mobile devices with a deep learning model," *J. Med. Biol. Eng.*, vol. 41, no. 5, pp. 659–668, 2021, doi: 10.1007/s40846-021-00649-5.

[45] T. Nakamura, Y. D. Alqurashi, M. J. Morrell, and D. P. Mandic, "Hearables: Automatic overnight sleep monitoring with standardized in-ear EEG sensor," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 1, pp. 203–212, Jan. 2020, doi: 10.1109/TBME.2019.2911423.

[46] K. B. Mikkelsen et al., "Accurate whole-night sleep monitoring with dry-contact ear-EEG," *Sci. Rep.*, vol. 9, no. 1, 2019, Art. no. 16824, doi: 10.1038/s41598-019-53115-3.

[47] D. Popovic, M. Khoo, and P. Westbrook, "Two electrodes on the forehead: Validation in healthy adults," *J. Sleep Res.*, vol. 23, no. 2, pp. 211–221, 2015, doi: 10.1111/jsr.12105.Automatic.

[48] S. Myllymaa, A. Muraja-Murro, and K. Myllymaa, "Assessment of the suitability of using a forehead EEG electrode set and chin EMG electrodes for sleep staging in polysomnography," *J. Sleep Res.*, vol. 25, no. 6, pp. 636–645, 2016, doi: 10.1111/jsr.12425.

[49] S. Kainulainen et al., "Comparison of EEG signal characteristics between polysomnography and self applied somnography setup in a pediatric cohort," *IEEE Access*, vol. 9, pp. 110916–110926, 2021, doi: 10.1109/ACCESS.2021.3099987.

[50] T. Miettinen et al., "Success rate and technical quality of home polysomnography with self-applicable electrode set in subjects with possible sleep bruxism," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 4, pp. 1124–1132, Jul. 2018, doi: 10.1109/JBHI.2017.2741522.

[51] S. Leach et al., "A protocol for comparing dry and wet EEG electrodes during sleep," *Front. Neurosci.*, vol. 14, no. 6, 2020, Art. no. 586, doi: 10.3389/fnins.2020.00586.

[52] R. G. Norman, I. Pal, C. Stewart, J. A. Walsleben, and D. M. Rapoport, "Interobserver agreement among sleep scorers from different centers in a large dataset," *Sleep*, vol. 23, no. 7, pp. 901–908, 2000, doi: 10.1093/sleep/23.7.1e.

[53] S. L. Himanen and J. Hasan, "Limitations of rechtschaffen and kales," *Sleep Med. Rev.*, vol. 4, no. 2, pp. 149–167, 2000, doi: 10.1053/smrv.1999.0086.

[54] M. T. Bianchi and B. Goparaju, "Potential underestimation of sleep apnea severity by at-home kits: Rescoring in-laboratory polysomnography without sleep staging," *J. Clin. Sleep Med.*, vol. 13, no. 4, pp. 551–555, 2017, doi: 10.5664/jcsm.6540.

[55] M. Bruyneel et al., "Sleep efficiency during sleep studies: Results of a prospective study comparing home-based and in-hospital polysomnography," *J. Sleep Res.*, vol. 20, no. 1, pp. 201–206, 2011, doi: 10.1111/j.1365-2869.2010.00859.x.

[56] E. S. Arnardottir et al., "Nocturnal sweating a common symptom of obstructive sleep apnoea: The Icelandic sleep apnoea cohort," *BMJ Open*, vol. 3, no. 5, 2013, Art. no. e002795, doi: 10.1136/bmjopen-2013-002795.