Accurate Deep Learning-Based Sleep Staging in a Clinical Population with Suspected Obstructive Sleep Apnea

Henri Korkalainen, Juhani Aakko, Sami Nikkonen, Samu Kainulainen, Akseli Leino, Brett Duce, Isaac O. Afara, Sami Myllymaa, Juha Töyräs, Timo Leppänen

***Abstract*—The identification of sleep stages is essential in the diagnostics of sleep disorders, among which obstructive sleep apnea (OSA) is one of the most prevalent. However, manual scoring of sleep stages is time-consuming, subjective, and costly. To overcome this shortcoming, we aimed to develop an accurate deep learning approach for automatic classification of sleep stages and to study the effect of OSA severity on the classification accuracy. Overnight polysomnographic recordings from a public dataset of healthy individuals (Sleep-EDF, *n*=153) and from a clinical dataset (*n*=891) of patients with suspected OSA were used to develop a combined convolutional and long short-term memory neural network. On the public dataset, the model achieved sleep staging accuracy of 83.7% (κ=0.77) with a single frontal EEG channel and 83.9% (κ=0.78) when supplemented with EOG. For the clinical dataset, the model achieved accuracies of 82.9% (κ=0.77) and 83.8% (κ=0.78) with a single EEG channel and two channels (EEG+EOG), respectively. The sleep staging accuracy decreased with increasing OSA severity. The single-channel accuracy ranged from 84.5% (κ=0.79) for individuals without OSA diagnosis to 76.5% (κ=0.68) for severe OSA patients. In conclusion, deep learning enables automatic sleep staging for suspected OSA patients with high accuracy and expectedly, the accuracy lowered with increasing OSA severity. Furthermore, the accuracies achieved in the public dataset were superior to previously published state-of-the-art methods. Adding an EOG channel did not significantly increase the accuracy. The automatic, single-channel-based sleep staging could enable easy, accurate, and cost-efficient integration of EEG recording into diagnostic ambulatory recordings.**

***Index Terms*—Deep learning, Electroencephalography, Obstructive sleep apnea, Recurrent neural network, Sleep staging**

I. Introduction

IDENTIFICATION of sleep stages is crucial in diagnostics of various sleep disorders. One of the most common sleep disorders is obstructive sleep apnea (OSA) which has been estimated to affect up to 38% of the general population [1]. In the diagnosis of OSA, sleep staging is conducted to assess the sleep characteristics and to accurately determine the total sleep time [2]. Accurate determination of total sleep time is of paramount importance as it significantly affects the parameters used to assess the severity of OSA.

According to the current sleep staging criteria [2], sleep is classified into five different stages: wake, rapid eye movement (REM) sleep and three stages of non-REM sleep (N1–N3). Classification into these stages is performed manually for 30-second epochs of sleep using electroencephalography (EEG), electrooculogram (EOG), and submental electromyogram (EMG) signals measured during polysomnography (PSG). Currently, at least 13 electrodes, with the positions determined by the International 10-20 System, are required for the measurement protocol [2]. Thus, the overall measurement protocol and the sleep staging process is time-consuming, laborsome and requires experienced professionals [3].

Despite the major effort and expenses that go into manual sleep staging, there are still shortcomings. Mainly, the agreement of two different scores is generally unsatisfactory [4]–[9]. The inter-rater reliability (IRR), measured with Cohen's kappa, between two scorers using the current sleep scoring criteria is commonly around 0.78 [4]. However, between international sleep centers, the reliability can be as low

This work was financially supported by the Research Committee of the Kuopio University Hospital Catchment Area for the State Research Funding (projects 5041767, 5041768, 5041770, 5041776, 5041779, 5041780, 5041781, and 5041783), by the Academy of Finland (decision numbers 313697 and 323536), by the Respiratory Foundation of Kuopio Region, by the Research Foundation of the Pulmonary Diseases, by Foundation of the Finnish Anti-Tuberculosis Association, by the Päivikki and Sakari Sohlberg Foundation, by Orion Research Foundation, by Instrumentarium Science Foundation, by the Finnish Cultural Foundation via the Post Docs in Companies program and via the Central Fund, by the Paulo Foundation, by the Tampere Tuberculosis Foundation, and by Business Finland (decision number 5133/31/2018). *(Corresponding author: Henri Korkalainen.)*

H. Korkalainen, S. Nikkonen, S. Kainulainen, A. Leino, S. Myllymaa, and T. Leppänen are with the Department of Applied Physics, University of Eastern Finland, Kuopio, Finland and the Diagnostic Imaging Center, Kuopio University Hospital, Kuopio, Finland. (e-mails: henri.korkalainen@uef.fi, sami.nikkonen@uef.fi, samu.kainulainen@uef.fi, akseli.leino@uef.fi, sami.myllymaa@uef.fi, timo.leppanen@uef.fi).

J. Aakko is with CGI Suomi Oy, Helsinki, Finland (email: juhani.aakko@cgi.com).

J. Töyräs is with the Department of Applied Physics, University of Eastern Finland, Kuopio, Finland, the Diagnostic Imaging Center, Kuopio University Hospital, Kuopio, Finland, and with the School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia. (email: juha.toyras@uef.fi).

B. Duce is with the Department of Respiratory & Sleep Medicine, Sleep Disorders Centre, Princess Alexandra Hospital, Ipswich Rd, Woolloongabba, QLD, Australia, and with the Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane City, QLD, Australia (email: brett.duce@health.qld.gov.au).

I.O. Afara is with the Department of Applied Physics, University of Eastern Finland, Kuopio, Finland (email: isaac.afara@uef.fi).

as 0.58 to 0.63 [5], [6], particularly due to poor scoring of N1 sleep [7], [8]. It has been shown that the agreement of N1 is approximately only 0.46 between sleep laboratories within Europe [4] and as low as 0.19 to 0.31 between international centers [5], [6]. Furthermore, the overall reliability of manual sleep staging may further decrease if an individual is experiencing medical conditions, for example, with OSA patients the reliability is worse than that of healthy individuals [8], [9]. Automatic scoring methods could potentially improve the consistency of sleep staging between different hospitals and health care systems. Furthermore, automatic methods capable of accurate sleep staging with a minimal number of measured signals could simplify the measurement protocol and reduce the related costs.

A number of automatic sleep staging methods have been previously published [10]–[31]. Traditionally, automated methods have relied on pre-defined rules, carefully selected features extracted from the signals, and classification algorithms [22]–[26]. Recently, a few machine-learning-based solutions utilizing deep learning and artificial neural networks have been presented [10]–[12], [14], [16]–[21], [27]–[31]. For these solutions, the classification rules or features of each sleep stage were not explicitly defined. However, previous studies have generally relied on heavy preprocessing by usually either transforming the signals into 2D images representing the spectral information [19], [27]–[30] or by reducing the signals into a limited number of predefined features [10], [30], [31]. Furthermore, deep learning models developed on research datasets of healthy individuals have generally suffered from a loss of accuracy when generalizing into populations with sleep disorders such as OSA [28]. In addition, a few deep learning-based automation attempts have demonstrated promising outcomes on sleep staging with a single EEG channel [10], [11], [13]–[16], [18]–[21]. While some of these have utilized deep learning [10], [11], [14], [16], [18], [19], [21], they have mostly relied on publicly available research datasets with a limited number of healthy individuals. Large clinical and well-balanced datasets have rarely been used, and the effect of sleep disorders on automatic sleep staging has not been thoroughly investigated.

We aimed to develop an accurate deep learning-based automatic method for the classification of sleep stages in patients with suspected OSA. We further aimed to achieve this by utilizing the raw signals without conducting heavy preprocessing. Furthermore, we aimed to study the effect of OSA severity on the performance of automatic sleep staging. We hypothesize that deep learning methods enable accurate sleep staging based on a single EEG channel for patients with suspected OSA and that the sleep staging accuracy decreases with increasing OSA severity.

II. Methods

A. Datasets

1) Sleep-EDF

We first utilized a public dataset, Physionet Sleep-EDF [32], [33], to allow comparison of the proposed deep learning-based approach with previous state-of-the-art methods. We utilized the version 2 of the expanded Sleep-EDF dataset released in March 2018. The dataset comprises 153 PSGs of 37 males and 41 females from a study investigating the effects of age on sleep in a healthy population (Sleep Cassette). We utilized the Fpz-Cz EEG signal for a single-channel input and combined it with a single horizontal EOG signal for two-channel input. Both signals were sampled with a 100 Hz frequency. No preprocessing was implemented on the signals. EMG recording was left out of this study due to its lower sampling frequency.

The sleep stages were originally scored according to the Rechtschaffen and Kales manual [34] into following stages: wake, N1, N2, N3, N4, REM, M (movement), and '?' (not scored). We combined the stages N3 and N4 into a single sleep stage to comply with the AASM guidelines [2]. Furthermore, the stages M and '?' were excluded from the study. The PSG recordings included long periods of wake in the beginning and end of the recording. Similarly to previous studies [11], [18], we only included 30 minutes of the wake before and after the sleep to obtain more realistic results and to enable comparison.

With the Sleep-EDF dataset, we conducted 10-fold cross-validation to assess the performance of the network, meaning that with each fold, 90% of the population was used for training and 10% as an independent test set. Furthermore, 10% of the training set was further used as the validation set during each fold. This was done to avoid overfitting during training, to choose an optimal model, and to keep the test set separate during each fold. 10-fold cross-validation was chosen over a single split to training, validation, and test set due to relatively small dataset and to enable comparison with the previous studies [11], [17]–[21].

2) Clinical dataset

The clinical dataset utilized in this study consists of 933 consecutive diagnostic overnight polysomnographies (PSG) of patients with clinical suspicion of OSA. Out of these, 891 individuals had successful recordings of all the required signals together with complete sleep stage scorings and were thus included in this study. The PSGs were conducted at the Princess Alexandra Hospital, Brisbane, Australia during 2015–2017 and recorded with the Compumedics Grael acquisition system (Compumedics, Abbotsford, Australia). The sleep stages were initially scored manually by multiple experienced scorers who participate regularly in intra- and inter-laboratory scoring concordance activities. Scoring was conducted based on the AASM rules [2] and the prevailing clinical practice of the Princess Alexandra Hospital. Ethical permissions for the data collection and processing were obtained from The Institutional Human Research Ethics Committee of the Princess Alexandra Hospital (HREC/16/QPAH/021).

From the recorded PSGs, EEG (derivation F4-M1) was used for single-channel input and it was complemented with EOG (derivation E1-M2) for two-channel input. EMG was not included to enable comparison with the public dataset. The signals were recorded with 1024 Hz sampling frequency and were downsampled to 64 Hz to reduce the computational load. No additional preprocessing was applied. The frontal EEG

channel was selected due to its simple measurement setup. The dataset was split into three individual sets: a training set, a validation set, and a test set. The training set comprised 717 whole night recordings (80%), and the validation and test sets comprised 87 recordings (10%) each.

Out of the 891 studied individuals, 493 were males and 398 females. The patients were mostly middle-aged and obese. According to the current severity classification of OSA, based on apnea-hypopnea index (AHI) [35], 152 individuals had no OSA ($5 < \text{AHI}$), 278 suffered from mild OSA ($5 \leq \text{AHI} < 15$), 208 from moderate OSA ($15 \leq \text{AHI} < 30$), and 254 had severe OSA ($\text{AHI} \geq 30$). Furthermore, 142 of the individuals were smokers, 197 suffered from diabetes, 368 had hypertension, 96 had cardiac arrhythmia, 22 had cardiac failure, and 41 had suffered a stroke. Table I shows the medians and interquartile ranges for sleep parameters and demographic information.

TABLE I
DEMOGRAPHIC INFORMATION OF THE CLINICAL DATASET (*N*=891)

	Median	Lower and upper quartiles
Apnea-hypopnea index (events/hour)	15.8	7.0–32.8
Age (years)	55.8	44.7–65.8
Body mass index (kg/m^2)	34.5	29.4–40.4
Arousal index (arousals/hour)	20.8	14.0–31.4
Total recording time (min)	442.5	409.5–474.5
Total sleep time (min)	308.8	253.8–359.8
Wake after sleep onset (min)	102.8	61.3–150
Sleep latency (min)	17.5	9.0-34.5
N1 (%)	11.0	6.8–18.9
N2 (%)	48.3	41.3–56.2
N3 (%)	18.3	9.6–27.1
REM (%)	17.1	11.8–22.1
NREM (%)	82.9	77.8–88.1
Sleep efficiency (%)	70.7	57.9–82.0

N1, N2, N3, and REM mean the percentage of the sleep stage and NREM the percentage of non-REM sleep during total sleep time. Sleep efficiency means the percentage of sleep during total recording time.

3) OSA severity

The effect of OSA severity on the performance of the automatic sleep staging model was assessed by training and evaluating the model separately on each OSA severity group (no OSA, mild, moderate, and severe OSA) of the clinical dataset described above. In this phase, only a single frontal EEG channel (F4-M1) was used, and as with the Sleep-EDF dataset, the performance was evaluated using 10-fold cross-validation. The 10-fold cross-validation was chosen due to reduced size of the dataset compared to the complete clinical dataset, and to get more comprehensive and comparable results over all the severity groups. Table II presents the number of 30-second epochs of each sleep stage in all the utilized datasets.

B. Neural network architecture

The estimation of the sleep stages (wake, N1, N2, N3, and REM) was conducted with a combined convolutional network (CNN) and recurrent neural network (RNN) trained in an end-to-end manner. The CNN aspect of the network was used to learn the characteristic features typical of each sleep stage, while the RNN considered the temporal distribution of the sleep stages overnight. The combined CNN and RNN structure was in essence similar to the architecture presented earlier by Supratak et al. [11]. However, sleep staging was conducted as a sequence-to-sequence classification problem, previously proposed by Phan et al. [29]. The network architecture was identical for the two-channel input and the single-channel input; the only difference was in the input dimension. The network was implemented in Python 3.6 using Keras API 2.2.4 with TensorFlow (version 1.13) backend. The training was conducted on a server with 32-core AMD Ryzen Threadripper 2990WX, 128 GB RAM and NVIDIA GeForce RTX 2080.

The CNN comprised six 1D convolutions each followed by batch normalization and a rectified linear unit (ReLU) activation, two max-pooling layers and a global average pooling layer (Fig. 1). The max-pooling layers were situated after the first two 1D convolutions and after the two following 1D convolutions. The global average pooling layer followed the last two 1D convolutions. The kernel size of the first 1D convolution was 21 and the stride size was 5. The second 1D convolution had a kernel size of 21 and stride size of 1. The number of convolutional filters equaled the sampling frequency (64 Hz for the clinical dataset, 100 Hz for Sleep-EDF) of the used dataset in the first two 1D convolutions. The remaining four 1D convolutions had a kernel size of 5 with a stride size of 1. The number of convolutional filters was two times the sampling frequency for the third and fourth 1D convolution and four times the frequency for the fifth and sixth 1D convolution.

The complete network comprised a time distributed layer of the complete CNN structure, a 0.3 gaussian dropout layer and a bidirectional long short-term memory (LSTM) layer followed by time distributed dense layer with softmax activation (Fig. 1). The number of units in the bidirectional LSTM was 4 times the sampling frequency. The LSTM utilized a tanh activation function and a dropout rate of 0.3. In the recurrent step, a hard sigmoid activation and a dropout rate of 0.5 were used. The last layer of the network comprised a dense layer with softmax activation producing the output sequence of sleep stage probabilities.

TABLE II
THE NUMBER OF 30-SECOND EPOCHS OF EACH SLEEP STAGE IN THE SLEEP-EDF DATASET, CLINICAL DATASET, AND AMONG THE GROUPS WITH DIFFERENT OSA SEVERITY

	Wake	N1	N2	N3	REM	Total
Sleep-EDF	65655 (34%)	21522 (11%)	69132 (35%)	13039 (7%)	25835 (13%)	195183
Clinical dataset	254278 (32%)	74102 (9%)	261317 (33%)	105298 (13%)	95800 (12%)	790795
No OSA	37303 (28%)	7501 (6%)	47782 (35%)	23076 (17%)	19262 (14%)	134924
Mild OSA	70532 (29%)	17947 (7%)	88412 (36%)	37554 (15%)	32485 (13%)	246930
Moderate OSA	59653 (32%)	15938 (9%)	61534 (33%)	25340 (14%)	22820 (12%)	185285
Severe OSA	86790 (39%)	32716 (15%)	63589 (28%)	19328 (9%)	212339 (9%)	223656

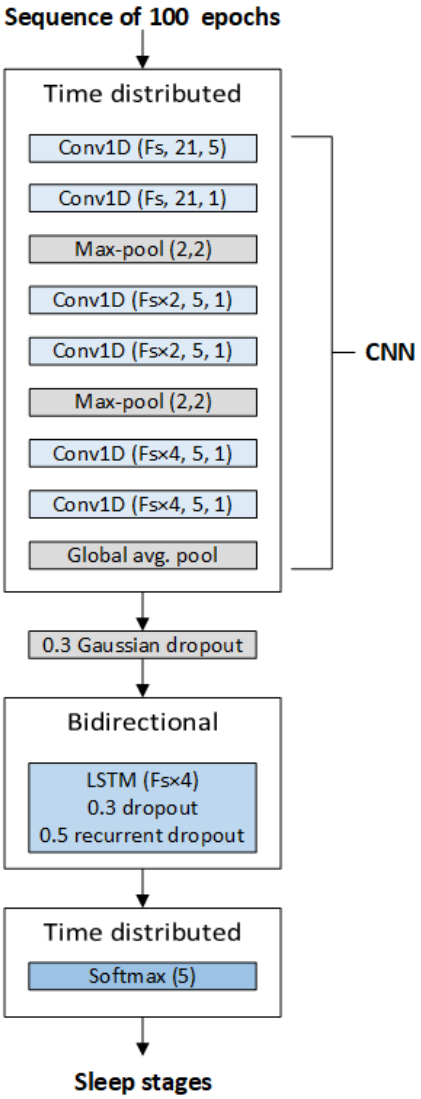

Fig. 1. The architecture of the combined convolutional neural network (CNN) and recurrent neural network (RNN). The parameters of the 1D convolutions (Conv1D) are given as (number of filters, kernel size, stride size) and as (pool size, stride size) for the max-pooling. Fs is the sampling frequency. For the LSTM and dense layer (Softmax) the number of units is given. The rate is given for the dropouts. The dropout layers were only active during training. Sequences of hundred 30-second epochs of the utilized signals were used as an input, and the model produced a sequence of softmax values representing the probabilities of each sleep stage for each epoch.

The model was trained with sequences of hundred 30-second epochs. An overlap of 75% was used when forming the sequences in the training set to increase its size fourfold. No overlap was used in the validation set or the test set. The model was trained with categorical cross-entropy as the loss function and an Adam optimizer with warm restarts [36] using a learning rate range of 0.001 to 0.00001. This learning rate range was optimized with a learning rate finder [37]. The model was validated with the validation set after each training cycle i.e. after the entire training set was passed through the network. The model was trained for a maximum of 200 training cycles or until the value of the loss function in the validation set no longer decreased during 20 consecutive training cycles The performance of the model was then assessed using in an independent test set.

C. *Interpretation of the results*

The accuracies were calculated in an epoch-by-epoch manner. Moreover, the inter-rater agreement between manual and automatic sleep staging was evaluated using Cohen's kappa coefficient (κ) [38] and the sensitivity and specificity of identifying sleep were calculated.

III. Results

A. *Sleep-EDF*

During the 10-fold cross-validation, the model achieved 89.8% training accuracy, 83.0% validation accuracy, and 83.9% testing accuracy with the two-channel input comprising single EEG and EOG channels. These accuracies corresponded to kappa values of 0.86, 0.77, and 0.78 in the training, validation, and test sets, respectively. Based on the guidelines by Landis and Koch [39], the kappa values indicate almost perfect agreement between manual and automatic sleep staging in the training set, and substantial agreement in the validation and test sets. In the test set, sleep was identified with 96.2% sensitivity and 93.7% specificity. For the individual sleep stages, the accuracy was 93.7% for wake, 87.3% for N2, 78.0% for N3, and 85.4% for REM in the test sets. The lowest concordance was seen with N1 (45.1%, Fig. 2 A).

With the single EEG channel, the obtained accuracies were 89.2%, 82.8%, and 83.7% in training, validation, and test sets, respectively. These correspond to kappa values of 0.85, 0.77, 0.77, respectively, indicating almost perfect or substantial agreement. In the test set, sleep was identified with 96.0% sensitivity and 93.4% specificity. Wake was identified with 93.4%, N1 with 43.4%, N2 with 87.3%, N3 with 78.7%, and REM with 85.4% accuracy (Fig. 2 B). The obtained accuracies and kappa values with single and two-channel input, alongside previous state-of-the-art results, are presented in Table III.

B. *Clinical dataset*

In the clinical dataset with the F4-M1 EEG and E1-M2 EOG channels, the model achieved 85.5% training accuracy and

TABLE III
Performance comparison

	Recordings (*n*)	Cross-validation	Accuracy	κ
Two-channel: Fpz-Cz and EOG				
Present work	**153**	**10-fold**	**83.9%**	**0.78**
Phan et al. [19]	39	20-fold	82.3%	0.75
Andreotti et al. [17]	38	20-fold	76.8%	0.68
Single-channel: Fpz-Cz				
Present work	**153**	**10-fold**	**83.7%**	**0.77**
Mousavi et al. [18]	153	10-fold	80.03%	0.73
Mousavi et al. [18]	39	20-fold	84.26%	0.79
Supratak et al. [11]	39	20-fold	82.0%	0.76
Phan et al. [19]	39	20-fold	81.9%	0.74
Tsinalis et al. [20]	39	20-fold	78.9%	-
Tsinalis et al. [21]	39	20-fold	74.8%	-

Only studies utilizing the sleep cassette dataset of the Sleep-EDF, conducting cross-validation with an independent test set, and having truncated the excess wake periods from the recordings are included.

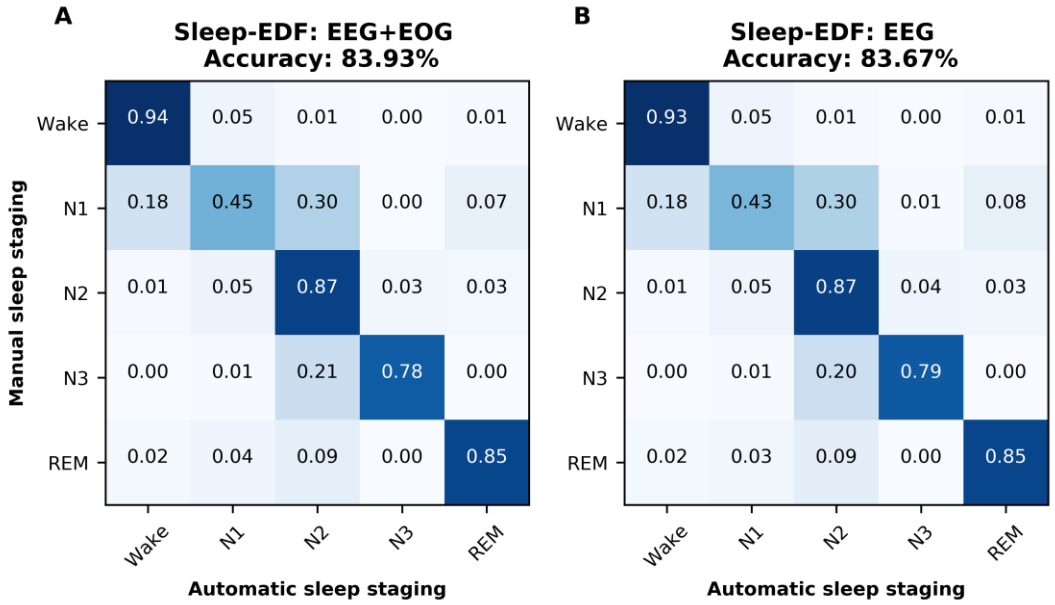

Fig. 2. Normalized confusion matrices of the classification accuracies from Sleep-EDF with (A) two-channel input (Fpz-Cz EEG and horizontal EOG) and (B) single EEG channel (Fpz-Cz) input.

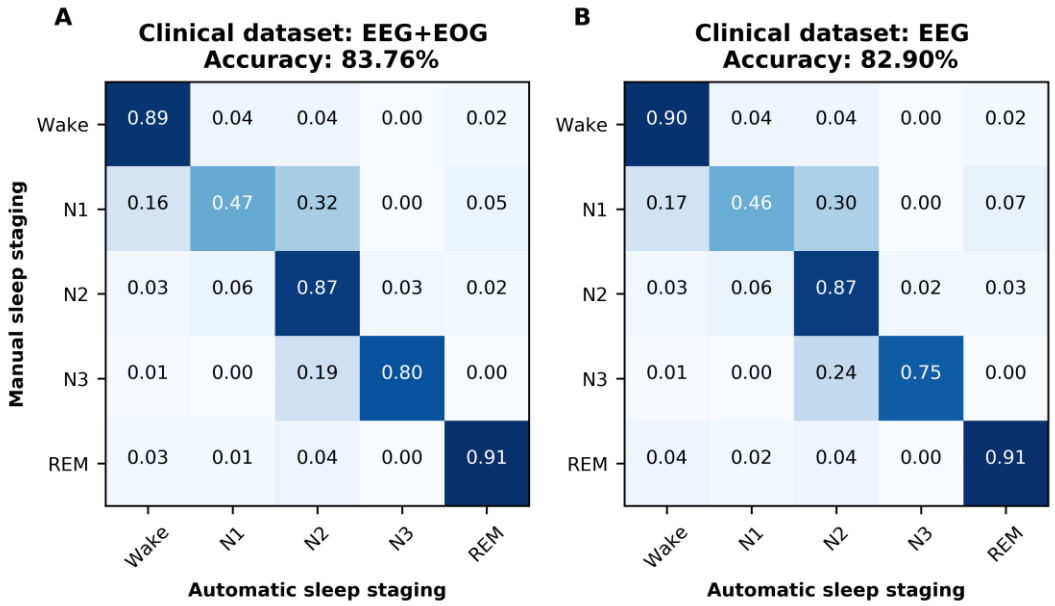

Fig. 3. Normalized confusion matrices of the classification accuracies from the clinical dataset with (A) two-channel input (F4-M1 EEG and E1-M2 EOG) and (B) single EEG channel (F4-M1) input.

83.8% validation accuracy. In the independent test set, the accuracy was 83.8%. These accuracies corresponded to Cohen's kappa values of 0.80, 0.78, and 0.78, respectively, indicating substantial agreement. Furthermore, the sensitivity of identifying sleep was 95.9% with 89.4% specificity in the test set. For individual sleep stages, the accuracy was 89.4% for wake, 87.2% for N2, 79.8% for N3 and 91.4% for REM in the test set. The lowest concordance between manual and automatic sleep staging was obtained in N1 with an accuracy of 46.9% (Fig. 3 A).

With the single frontal EEG channel, the accuracies were 86.3%, 83.4%, and 82.9% in the training, validation and test sets, respectively. These accuracies corresponded to kappa values of 0.82, 0.78, and 0.77. In the test set, the sensitivity for identifying sleep was 95.6% with 89.8% specificity. The N1 sleep stage was the most challenging to identify (classification accuracy of 46.0%). In contrast, wake was identified with 89.8% accuracy, N2 with 86.5%, N3 with 75.4%, and REM with 90.8% accuracy (Fig. 3 B).

C. OSA severity

When comparing the OSA severity groups, the accuracies and kappa values were lowest for patients with severe OSA (Table IV). The accuracy increased with decreasing OSA severity and were the highest for individuals without OSA. Similar behavior was perceived in the individual sleep stages, with the exception of N1 sleep which was most accurately classified for severe OSA patients (Fig. 4).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2951346, IEEE Journal of Biomedical and Health Informatics

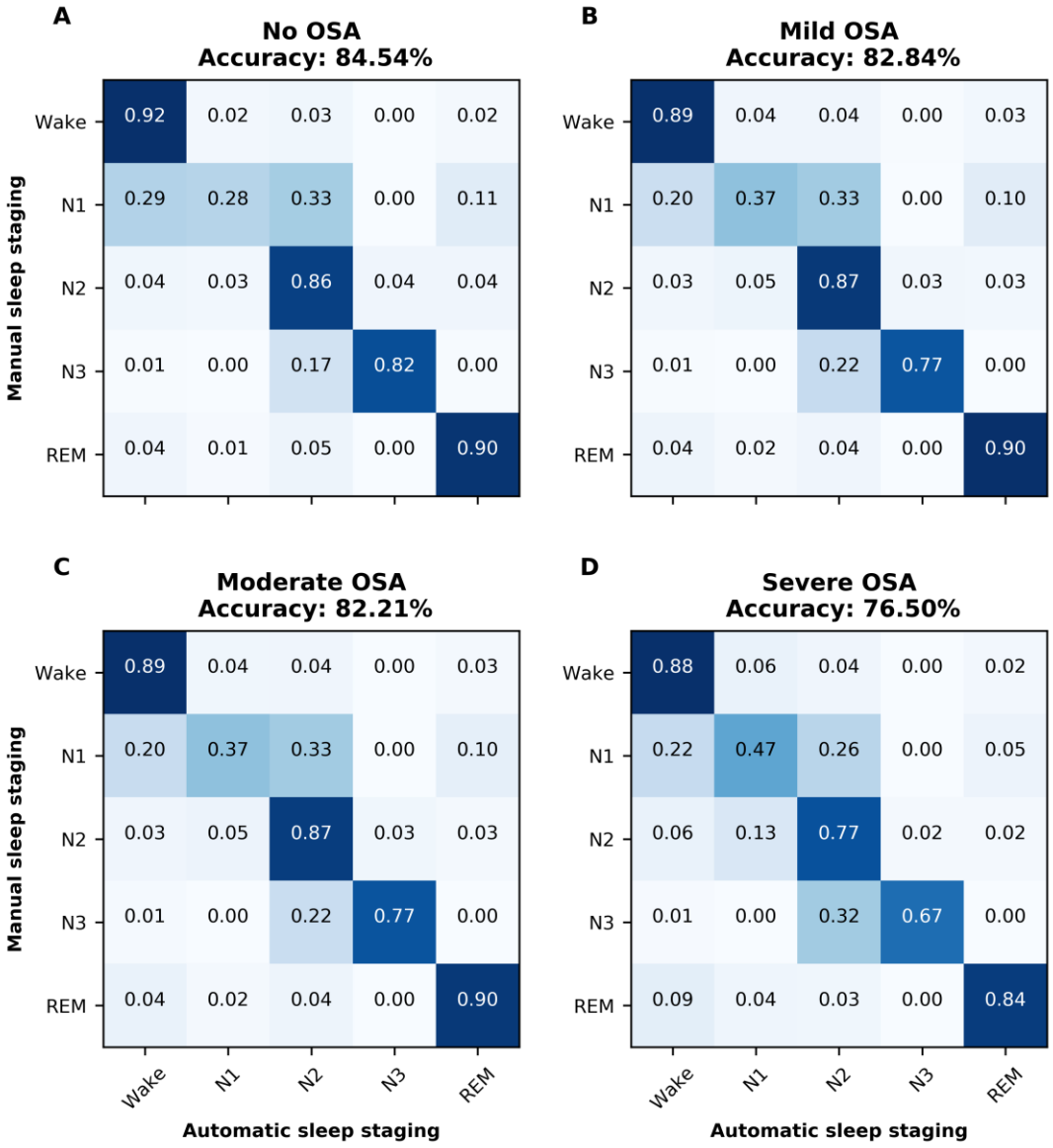

Fig. 4. Normalized confusion matrices of the classification accuracies with a single EEG channel (F4-M1) in individuals (A) with no OSA, (B) with mild OSA, (C) with moderate OSA, and (D) with severe OSA.

TABLE IV
PERFORMANCE IN OSA SEVERITY GROUPS

	n	Accuracy			κ		
		Training	Validation	Test	Training	Validation	Test
No OSA	152	89.4%	84.4%	84.5%	0.86	0.79	0.79
Mild OSA	278	87.7%	82.4%	82.8%	0.83	0.77	0.77
Moderate OSA	208	87.2%	83.0%	82.2%	0.83	0.77	0.76
Severe OSA	254	82.9%	76.7%	76.5%	0.77	0.68	0.68

IV. DISCUSSION

In this study, we developed a deep learning-based method for automatic classification of sleep stages from raw EEG and EOG signals using both a large clinical dataset (n=891) comprising patients with suspected OSA and a publicly available dataset of healthy individuals (n=153). Sleep staging was implemented using both two-channel input and single-channel input. Furthermore, we also studied the effect of OSA severity on the performance of automatic sleep staging. Overall, the automatic sleep staging method achieved high accuracies: 83.9% (κ=0.78) and 83.6% (κ=0.77) with single and two-channel inputs, respectively, in the public dataset, and almost correspondingly 82.9% (κ=0.78) and 83.8% (κ=0.77) in the clinical dataset. The accuracy of the sleep staging decreased with increasing OSA severity with the accuracy being the highest for individuals without OSA and lowest with individuals having severe OSA.

Furthermore, deep learning could enable accurate sleep staging with a single easily measurable frontal EEG channel with practically the same accuracy as with the additional EOG channel. Overall, the reliability of these automatic sleep staging approaches was comparable with the reliability of manual sleep scoring as previously reported in numerous studies [4]–[9].

The developed deep learning model compared favorably to previous studies based on the publicly available Sleep-EDF dataset [32], [33]. Our method slightly surpassed the performance of previously published methods (Table III). Previously, Mousavi et al. have utilized the updated Sleep-EDF dataset with 153 recordings and included only 30 minutes of wake before and after sleep achieving an accuracy of 80.03% (κ=0.73) with a single EEG channel [18]. In comparison, we achieved a single-channel accuracy of 83.7% (κ=0.77) with the same dataset and identically truncated signals. Other studies based on state-of-the-art methods have been conducted with the smaller Sleep-EDF dataset with only 39 recordings [11], [17], [19]–[21] and thus direct comparison is difficult. However, it is noteworthy that Mousavi et al. compared the performance of their sleep staging method in both the smaller and updated datasets and achieved significantly higher accuracy (84.26% vs 80.03%) im the smaller dataset [18]. This indicates that accurate sleep staging may be easier in the smaller dataset when compared to the larger, updated dataset used in the present study. Furthermore, direct comparison with previous studies is difficult due to non-standardized use of the database. The recordings in the database contain excessive wake periods before and after sleep. Inclusion of the excess wake periods to the automatic sleep staging can lead to overly optimistic results. Therefore, we only compared our results to studies truncating the excess amount of wake either by using only 30 minutes of wake before and after sleep [11], [18] or by only using the sleep itself [19], [20], [21]. Furthermore, the results cannot be compared to studies not using an independent test set to assess the performance, as these results could be distorted by overfitting.

The PSGs collected from suspected OSA patients have been problematic for previous automatic sleep staging approaches and even the reliability of manual scoring is known to be lower than with healthy individuals [8], [9], [28]. This is most likely due to a fragmented sleep structure and an increase in N1 sleep stage, which are typical for OSA patients [9]. In the present study, the sleep staging accuracies decreased with increasing OSA severity, with an accuracy of 84.5% for individuals without OSA and 76.5% for patients with severe OSA. Wake and N1 sleep comprised a larger portion of the recording whereas N2, N3, and REM comprised a smaller portion of the recording for patients with severe OSA when compared to the other patient groups (Table II). Especially N1 comprised a significantly larger portion (15%) of the recordings in the severe OSA group compared to the other groups (6–9%). This supports the idea that fragmented sleep structure caused by OSA impairs the accuracy and reliability of sleep staging. However, it is noteworthy that the accuracy of staging N1 was 47% for patients with severe OSA (Fig. 4 D) while it was only 28% for individuals without OSA (Fig. 4 A). This increase in accuracy is likely due to a larger amount of N1 sleep epochs and transitions between wake and N1 available during the training of the deep learning model. Furthermore, it is possible that manually identifying the N1 sleep of an individual patient becomes more reliable when more N1 sleep and especially more transitions between wake and N1 are available. This could improve the automatic scoring of N1 in addition to the accuracy increasing simply due to the larger training material. However, the N1 staging accuracy remained the lowest amongst all sleep stages and the accuracy of the other stages decreased for severe OSA. Thus, the increase in N1 accuracy was insufficient to compensate for the reduction in total accuracy with increasing OSA severity.

Implementation of automatic sleep staging system in a clinical setting could provide significant benefits over the prevailing practice. Currently, the manual sleep scoring lacks sufficient inter-rater reliability, as perceived from numerous studies [4]–[9]. It could be argued that since our deep learning-based sleep staging method was trained with manual scorings, it cannot be better than human scorers. However, the developed automatic method may produce a consensus over multiple scorers and thus minimize the variability. The developed automatic sleep staging method did not learn only from a single scorer as the clinical PSGs were scored by multiple sleep technicians potentially differing in their scoring preferences and traditions. Thus, the optimal solution is not to mimic a single scorer but rather classify the stages as similarly as possible to the majority of the scorers. Furthermore, after training, the automatic method always scores the sleep stages similarly regardless of the situation. This can be a major advantage over a manual scorer, as the automatic scoring does not depend on factors such as human error, vigilance level, or the current scoring environment.

In addition to high variability, manual sleep staging is highly time-consuming and requires trained specialists for a rather repetitive task. The sleep staging of a single patient could be performed in less than a second with the proposed automatic sleep staging method, whereas the manual scoring can take up to hours even for experienced scorers. Although the automatic sleep staging method is reliable for suspected OSA patients, the reliability of sleep stage classification of individuals with other sleep disorders remains to be studied.

Accurate sleep staging with a single EEG channel may present opportunities for further development and application of various ambulatory EEG and PSG acquisition systems [40], [41]. Currently, conducting PSG is expensive and requires trained specialists. Thus, cheaper ambulatory recordings have been developed and shown to be accurate for the diagnosis of OSA [3]. Ambulatory recordings are even the preferred diagnostic method in some health care systems [42], [43]. However, the major disadvantage of ambulatory recordings is often the lack of EEG recording, preventing identification of sleep stages and resulting in crude approximations of the total sleep time from other signals. Thus, ambulatory EEG recording based on a single frontal channel could enhance the accuracy of the ambulatory recordings whilst ensuring simplicity and cost-efficiency. However, further studies are warranted to assess and

verify the performance of the developed sleep staging method when applied together with an ambulatory recording device.

The most significant limitation of the developed deep learning-based sleep staging method is the scoring of N1 sleep stage. With both the two-channel and single-channel approaches, the agreement with the manual scoring of stage N1 was the lowest of all sleep stages with a variation of 43–47% between the public and clinical datasets. However, N1 is the most difficult sleep stage to identify even for experienced manual scorers [7], [8]. The agreement in N1 we achieved with the automatic sleep staging method is, however, comparable to the inter-rater agreement between manual scorers, which is between 0.19 and 0.46 [4]–[6]. Thus, the limited accuracy of scoring N1 sleep stage may not be due to the developed sleep staging method, but rather in the scoring definitions of N1 resulting in disagreement between experienced manual scorers.

V. Conclusion

The proposed deep learning-based automatic method enables reliable, fast, and accurate sleep staging for suspected OSA patients. The accuracy of the sleep staging decreases with increasing OSA severity but with the utilized large clinical dataset, the sleep staging can be conducted for patients suffering from OSA with almost comparable accuracy to individuals without OSA. Practically, automatic sleep staging can be performed as accurately using either a combination of single EEG and EOG signals or using a single frontal EEG channel. The single-channel approach could enable a cost-efficient, simple, and accurate sleep staging in OSA diagnostics.

References

[1] C. V Senaratna et al., “Prevalence of Obstructive Sleep Apnea in the general population: A systematic review,” Sleep Med. Rev., vol. 34, pp. 70–81, 2017.

[2] R. B. Berry et al., The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications., Version 2.5 Darien, IL: American Academy of Sleep Medicine, 2018.

[3] N. Collop, “Portable monitoring for the diagnosis of obstructive sleep apnea,” Curr. Opin. Pulm. Med., vol. 14, no. 6, pp. 525–529, 2008.

[4] H. Danker-Hopfe et al., “Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard,” J. Sleep Res., vol. 18, no. 1, pp. 74–84, 2009.

[5] U. J. Magalang et al., “Agreement in the Scoring of Respiratory Events and Sleep Among International Sleep Centers,” Sleep, vol. 36, no. 4, pp. 591–596, 2013.

[6] X. Zhang et al., “Process and outcome for international reliability in sleep scoring,” Sleep Breath., vol. 19, no. 1, pp. 191–195, 2015.

[7] R. S. Rosenberg and S. Van Hout, “The American Academy of Sleep Medicine inter-scorer reliability program: Respiratory events,” J. Clin. Sleep Med., vol. 10, no. 4, pp. 447–454, 2014.

[8] 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. 9, no. 1, p. 89, 2013.

[9] 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.

[10] N. Michielli, U. R. Acharya, and F. Molinari, “Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals,” Comput. Biol. Med., vol. 106, no. January, pp. 71–81, 2019.

[11] A. Supratak, H. Dong, C. Wu, and Y. Guo, “DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 25, no. 11, pp. 1998–2008, 2017.

[12] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, “A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 26, no. 4, pp. 758–769, 2018.

[13] A. R. Hassan and M. I. H. Bhuiyan, “Automated identification of sleep states from EEG signals by means of ensemble empirical mode decomposition and random under sampling boosting,” Comput. Methods Programs Biomed., vol. 140, pp. 201–210, 2017.

[14] E. Bresch, U. Großekathöfer, and G. Garcia-Molina, “Recurrent Deep Neural Networks for Real-Time Sleep Stage Classification From Single Channel EEG,” Front. Comput. Neurosci., vol. 12, no. October, pp. 1–12, 2018.

[15] A. R. Hassan and M. I. H. Bhuiyan, “An automated method for sleep staging from EEG signals using normal inverse Gaussian parameters and adaptive boosting,” Neurocomputing, vol. 219, pp. 76–87, 2017.

[16] A. I. Humayun, A. S. Sushmit, T. Hasan, and M. I. H. Bhuiyan, “End-to-end Sleep Staging with Raw Single Channel EEG using Deep Residual ConvNets,” pp. 1–5, 2019.

[17] F. Andreotti, H. Phan, N. Cooray, C. Lo, M. T. M. Hu, and M. De Vos, “Multichannel Sleep Stage Classification and Transfer Learning using Convolutional Neural Networks,” Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, vol. 2018-July, pp. 171–174, 2018.

[18] S. Mousavi, F. Afghah, and U. R. Acharya, “SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach,” PLoS One, vol. 14, no. 5, p. e0216456, 2019.

[19] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, and M. De Vos, “Joint Classification and Prediction CNN Framework for Automatic Sleep Stage Classification,” IEEE Trans. Biomed. Eng., vol. 66, no. 5, pp. 1285–1296, 2019.

[20] O. Tsinalis, P. M. Matthews, and Y. Guo, “Automatic Sleep Stage Scoring Using Time-Frequency Analysis and Stacked Sparse Autoencoders,” Ann. Biomed. Eng., vol. 44, no. 5, pp. 1587–1597, 2016.

[21] O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou, “Automatic Sleep Stage Scoring with Single-Channel EEG Using Convolutional Neural Networks,” 2016.

[22] B. Koley and D. Dey, “An ensemble system for automatic sleep stage classification using single channel EEG signal,” Comput. Biol. Med., vol. 42, no. 12, pp. 1186–1195, 2012.

[23] U. R. Acharya, E. C.-P. Chua, K. C. Chua, L. C. Min, and T. Tamura, “Analysis and automatic identification of sleep stages using higher order spectra,” Int. J. Neural Syst., vol. 20, no. 06, pp. 509–521, 2010.

[24] P. Anderer et al., “An E-Health solution for automatic sleep classification according to Rechtschaffen and Kales: Validation study of the somnolyzer 24 x 7 utilizing the siesta database,” Neuropsychobiology, vol. 51, no. 3, pp. 115–133, 2005.

[25] S. F. Liang, C. E. Kuo, Y. H. Hu, Y. H. Pan, and Y. H. Wang, “Automatic stage scoring of single-channel sleep EEG by using multiscale entropy and autoregressive models,” IEEE Trans. Instrum. Meas., vol. 61, no. 6, pp. 1649–1657, 2012.

[26] C. Stepnowsky, D. Levendowski, D. Popovic, I. Ayappa, and D. M. Rapoport, “Scoring accuracy of automated sleep staging from a bipolar electroocular recording compared to manual scoring by multiple raters,” Sleep Med., vol. 14, no. 11, pp. 1199–1207, 2013.

[27] S. Biswal, H. Sun, B. Goparaju, M. Brandon Westover, J. Sun, and M. T. Bianchi, “Expert-level sleep scoring with deep neural networks,” J. Am. Med. Informatics Assoc., vol. 25, no. 12, pp. 1643–1650, 2018.

[28] A. Patanaik, J. L. Ong, J. J. Gooley, S. Ancoli-Israel, and M. W. L. Chee, “An end-to-end framework for real-time automatic sleep stage classification,” Sleep, vol. 41, no. 5, pp. 1–11, 2018.

[29] H. Phan, F. Andreotti, N. Cooray, O. Chén, and M. de Vos, “SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 27, no. 3, pp. 400–410, Jan. 2019.

[30] A. Malafeev et al., “Automatic Human Sleep Stage Scoring Using Deep Neural Networks.,” Front. Neurosci., vol. 12, no. November, p. 781, 2018.

[31] H. Sun et al., “Large-Scale Automated Sleep Staging,” Sleep, vol. 40, no. 10, 2017.

[32] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberyé, “Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG,” IEEE Trans. Biomed. Eng., vol. 47, no. 9, pp. 1185–1194, 2000.

[33] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," Circulation, vol. 101, no. 23, p. e220, 2000.
[34] A. Rechtschaffen and A. Kales, A manual of standardized terminology, techniques and scoring system of sleep stages in human subjects. University of California, Brain Information Service/Brain Research Institute, Los Angeles, 1968.
[35] AASM, "Sleep–Related Breathing Disorders in Adults: Recommendations for Syndrome Definition and Measurement Techniques in Clinical Research," Sleep, 1999.
[36] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," pp. 1–16, 2016.
[37] L. N. Smith, "Cyclical learning rates for training neural networks," Proc. - 2017 IEEE Winter Conf. Appl. Comput. Vision, WACV 2017, no. April, pp. 464–472, 2017.
[38] J. Cohen, "A coefficient of agreement for nominal scales," Educ. Psychol. Meas., vol. 20, no. 1, pp. 37–46, 1960.
[39] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," Biometrics, vol. 33, no. 1, pp. 159–174, 1977.
[40] S. Myllymaa et al., "Assessment of the suitability of using a forehead EEG electrode set and chin EMG electrodes for sleep staging in polysomnography.," J. Sleep Res., vol. 25, no. 6, pp. 636–645, 2016.
[41] 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. Heal. Informatics, vol. 22, no. 4, pp. 1124–1132, 2018.
[42] W. W. Flemons, N. J. Douglas, S. T. Kuna, D. O. Rodenstein, and J. Wheatley, "Access to Diagnosis and Treatment of Patients with Suspected Sleep Apnea," Am. J. Respir. Crit. Care Med., vol. 169, no. 6, pp. 668–672, 2004.
[43] E. S. Arnardottir et al., "Variability in recording and scoring of respiratory events during sleep in Europe: a need for uniform standards," J. Sleep Res., vol. 25, no. 2, pp. 144–157, 2016.