Automatic IHR-based sleep stage detection using features of residual neural network

Bhekumuzi M. Mathunjwa [1], Yin-Tsong Lin [2], Chien-Hung Lin [2], Maysam F. Abbod [3], Muammar Sadrawi [4], Jiann-Shing Shieh [1,*]

[1] Department of Mechanical Engineering and Innovation Center for Big Data and Digital Convergence, Yuan Ze University, Taoyuan, Chung-Li 32003 Taiwan;
[2] AI R&D Department, New Era AI Robotic Inc., Taipei City 105, Taiwan; ,
[3] Department of Electronics and Computer Engineering, Brunel University London, UB8 3PH UK;
[4] Bioinformatics Department, School of Life Sciences, Indonesia International Institute for Life Sciences, JI. Pulomas Barat Kav 88, Jakarta 13210

Received: date; Accepted: date; **Published**: date

ABSTRACT

Untreated sleep disorders can harm bodily functions, and a sleep study and monitoring of sleep stages are the first steps in diagnosing these disorders. Using Polysomnography (PSG), signal scoring for sleep stage determination has become a familiar investigation in recent years. Despite its effectiveness, the procedure is time-consuming and costly. This study presents a cost-effective method for sleep classification based on Electrocardiogram (ECG) input signals. We proposed a multi-ethnic study of the Atherosclerosis dataset, including 1700 PSG, to develop a Residual Neural Network (RNN) classifier to stage sleep from Instantaneous Heart Rate (IHR) extracted from the ECG signals. The proposed system follows the following steps: ECG collection, signal preprocessing (including ECG normalization and segmentation, instant heart rate calculation and normalization, resampling, and filtering), and classification using an RNN. A Convolutional Neural Network (CNN) is used to detect sleep stages using preprocessed segments of the IHR time series of 240 samples centered on 30-s epochs as inputs. The proposed algorithm in the five-fold cross-validation achieved an accuracy of 85.32%, a kappa of 77.11%, a Sensitivity of 81.14%, a Specificity of 82.68%, and an F-1 score of 81.87%. The results show that ECG data provide valuable information about sleep stages for a large population.

Keywords: **Electrocardiography, Instantaneous heart rate, Residual neural network, Sleep stages classification**

The abbreviations used in this paper are in alphabetical order as follows:

Nomenclature	Referred to	Nomenclature	Referred to
Acc	Accuracy	MGH	Massachusetts General Hospital
AAMI	Association for the Advancement of Medical Instrumentation	NREM	Non-Rapid Eye Movement
AASM	American Academy of Sleep Medicine	OSAS	Obstructive Sleep Apnea Syndrome
CNN	Convolutional Neural Network	PSG	Polysomnography
CPAP	Constant Positive Airway Pressure	ReLU	Rectilinear Unit
ECG	Electrocardiogram	REM	Rapid Eye Movement
EDR	ECG-Derived Respiration	RNN	Residual Neural Network
EEG	Electroencephalography	R&K	Rechtschaffen and Kales
EMG	Electromyography	ResNet	Residual Neural Network
EOG	Electrooculography	Sens	Sensitivity

FN	False Negative	SLPDB	MIT-BIH Polysomnographic Database
FP	False Positive	Sp	Specificity
F1	F-1 Score	SHHS	Sleep Heart Health Study
HRV	Heart Rate Variability	S1	Sleep Stage 1
IBI	Inter Beat Intervals	S2	Sleep Stage 2
ILSVRC	ImageNet Large Scale Visual Recognition Challenge	S3	Sleep Stage 3
IHR	Instantaneous Heart Rate	TP	True Positive
Kappa	Cohen's Kappa	TN	True Negative
MESA	Multi-Ethnic Study of Atherosclerosis		

1. Introduction

When thinking about how much sleep you need, it is normal to take into account the amount of time you will be in bed. The importance of sleeping enough is undeniable, but it is not the only factor. Additionally, bedtime should also be considered in terms of quality and whether it actually promotes health restoration. A person must experience all four stages of sleep several times throughout the night in order to reach the level of health restoration sleep. Sleep contributes to your feeling of refreshed in body and mind. Therefore, understanding how the cycle of sleep affects a person's health and sleep can help us detect sleeping disorders and other health conditions early.

The human sleep cycle is decomposed into epochs, usually 30 seconds apart, and each epoch has its own sleep stage. The Association for the Advancement of Medical Instrumentation (AAMI) [1] states that there are five stages of sleep: Rapid Eye Movement sleep (REM), non-REM sleep (Sleep Stage-1 (S1), Sleep Stage-2 (S2), and Sleep Stage-3 (S3)), and Wakefulness (W). The standard method for measuring and analyzing sleep involves the recording of multiple physiological changes such as heart rhythms, brain activity, muscle activity, and eye movements by means of Polysomnography (PSG) [2]. PSG provides a variety of sleep-related signals, such as Electrooculography (EOG), Electroencephalography (EEG), and Electromyography (EMG) [3, 4]. In a process known as sleep staging, experts can estimate sleep duration and quality by using these signals based on sleep manuals like the American Academy of Sleep Medicine (AASM) and Rechtschaffen and Kales (R&K) [5].

PSG recording is labor-intensive and time-consuming due to the need for attaching several sensors to the subjects overnight, which requires visiting sleep centers or laboratories. Furthermore, PSGs are expensive to record and process and they are not readily available for analysis [6, 7]. While PSGs offer an accurate assessment of sleep, they are not practical for long-term use due to their high costs and because so many sensors need to be implanted into the body of the participant, which can disturb their sleep.

According to statistics, about 70% of Americans suffer from sleep disorders, a condition that causes chronic sleep deprivation and interferes with normal daytime functioning. Many patients struggle to identify sleep disorders because they occur at night and most go undiagnosed [8]. Sleep disorders can also result in chronic health issues and be very expensive to manage. Each year in the US, about $400 billion is lost due to insufficient sleep [9]. Therefore, early detection and diagnosis of sleep can improve daytime functioning and save resources.

An electrocardiogram (ECG) is a simple test that measures your heart rhythm and activities that affect the heart. ECG signals are among the low-cost methods for stimulating the heartbeat by measuring the voltage over time. Sensors attached to your skin detect the electrical signals produced by your heartbeat. A machine collects ECG signals, which are analyzed by professionals to find the information contained within the heart. Several studies investigated the use of ECGs in sleep apnea detection [10, 11]. The ECG indicates the overall health of the heart. An ECG interpretation includes a structured assessment of the ECG waves and intervals. In a normal ECG, P, Q, R, S, T, and U waves are present, along with various intervals, including P-R, Q-T, R-R, and S-T. The amplitude and duration of these intervals are calculated during heartbeat processing and classification.

Several recent studies have demonstrated that networks based on artificial neural networks are very successful at recognizing and categorizing objects [12, 13, 14, 15, 16]. Researchers have been inspired by these successes to design different architectures that solve problems encountered by others more efficiently and study the roles played

by particular parameters in solving problems. One of the most successful deep learning models is the Residual Neural Network (ResNet) architecture proposed by [17]. ResNet is a type of convolutional neural network (CNN) that use identity mapping (skip connections) to incorporate the output of each layer into the next. It was designed with residual blocks to address the problem encountered when training very deep networks. ResNet architecture took part and won first place in the 2015 ILSVRC classification competition. Furthermore, it demonstrated a significant improvement on the famous image recognition database COCO [18].

ECG signals can be a cost-effective and easy-to-use alternative to the more expensive and complex EEG signals for classifying sleep stages. ECG signals can provide information on heart rate variability, which is associated with autonomic nervous system activity, and can be combined with other physiological signals such as respiratory signals for more accurate sleep stage classification. Although ECG signals show promise for sleep stage classification, more research is needed to validate their use and optimize their effectiveness.

To improve the accuracy of sleep stage classification using ECG features, researchers have attempted to increase the amount of data used for training and explored the use of deep CNNs. However, as the number of layers in a CNN model increases, the model size also increases, resulting in slower training times, higher computational costs, and overfitting. To address these issues, researchers have turned to ResNet architecture, which allows for the addition of more layers while keeping the model size relatively small. ResNet achieves this by using residual connections, which enable the network to skip certain layers and pass the output of one layer directly to another layer further down the network. This has proven to be particularly useful for improving the accuracy of ECG-based sleep stage classification.

Rather than extracting features manually from ECG data, researchers can use CNNs to learn local cardiac features automatically. This approach saves time and eliminates the need for researchers to calculate and interpret these features themselves. By analyzing the raw ECG signal, CNNs can extract key features such as amplitude, frequency, and duration of cardiac activity without human intervention.

In this study, the proposed approach is evaluated according to its ability to automatically stage sleep based on ECG signals acquired from the Multi-Ethnic Study of Atherosclerosis (MESA) dataset into awake (W), REM, and Non-Rapid Eye Movement sleep (NREM) which includes stages 1, 2, and 3. In order to optimize accuracy, and since the data used in stage S1 is insufficient for learning, four stages are classified (W, S1&S2, S3, REM). The three stages in NREM sleep are reduced to two stages by combining S1 and S2 to optimize for accuracy since S1 does not have enough data to distinguish its fractures from the rest of the stages. An individual's sleep state is recorded by recording 30-second epochs according to methods established by the AASM [19]. The classifier uses two-minute Instantaneous Heart Rate (IHR) segments centered at 30-s epochs computed from the ECG signals as one-dimensional input data.

In summary, the paper makes several significant contributions. Firstly, the study investigates the use of the IHR to detect sleep stages through ECG analysis and demonstrates its effectiveness as a dependable and accurate approach. Secondly, the proposed method achieves an impressive 85% accuracy rate in detecting sleep stages, surpassing the established standard of polysomnography. Notably, the ResNet architecture used in the models reduces memory requirements, allowing them to be used in mobile devices. While the data preprocessing followed previously applied methods, the paper's approach stands out by applying an effective filter at the end of the data processing that removes values greater than 5 standard deviations per segment, which can interfere with the training process. Lastly, the IHR-based method is computationally efficient, enabling it to be easily integrated into existing ECG monitoring systems. In light of the obtained results, ResNet can be used with IHR to stage sleep.

The rest of the paper is organized as follows: Section I provides analyses of related works. Sections II and III provide an analysis of the methodology, including data acquisition, annotation scoring, input signal preprocessing, the algorithm, and performance evaluation and results of our proposed work. Lastly, sections IV and V are dedicated to the discussion and conclusions.

1.1. Related work

Su, et al. [20] classified sleep stage into five labels, which are awake, N1, N2, N3, and REM. Fourteen PSG records from the NicoletOne v44 sleep Diagnostic system were trained, where signal entropy and a frequency spectrograph were derived from a 30-second EEG inputs. The validation dataset included 18 PSD records. The Gaussian SVM machine learning method was used for feature learning. A median filter with three 30 s epochs was

utilized to smooth the predicted stages. The resulting testing accuracy was 79%. They concluded that the entropy and spectral power features are stable across different classification methods.

Researchers [21] conducted a study on sleep stage classification by combining recurrence and CNNs to learn sleep apnea events, clinical labels, and leg movements. PSG reports from 15 804 subjects at the SHHS and the Massachusetts General Hospital (MGH) sleep laboratory were included in their study. Spectrogram and raw waveform representation from EEG and EMG data were utilized as input data for the model. 30-second epochs were segmented into 29 sub epochs of 2-seconds with an overlap of 1-second. A multi-taper method was used to estimate the power spectrum density. The overall accuracy for the RCNN was 87.5%.

Garcia-Molina et al. [22] investigated the impact of ECG-generated Inter Beat Intervals (IBIs) on four stages of sleep (light, deep, REM, and wake). PSG data from 1147 subjects were analyzed using two datasets from the PhysioNet resource and data from self-reported healthy sleep participants. Among the subjects are healthy sleepers, people with insomnia, people with REM behavior disorder, and people with periodic leg movement disorder. Garcia-Molina et al used a 150-second window with a 30-second shift. According to the DNN, their accuracy was 76% and the kappa was 52%, indicating the potential of using ECGs to screen for sleep disorders.

Surantha et al. [23] utilized input signals from the MIT-BIH polysomnographic dataset and categorized sleep stage into four sets of stages, 6, 4, 3, and 2 classes. Particle Swarm Optimizations and Extreme Learning Machine were integrated as a machine learning approach to learn features from physiological signals in the MIT-BIH. Eighteen HRV features were extracted from the ECG, which includes AVNN, SDNN, RMSSD, SDSD, NN50, pNN50, HRV Triangular Index, SD1, SD2, SDISD2 Ratio, S, TP, VLF, LF, HF, LFHF Ratio, LFnorm, and HFnorm. Training and testing data was split at 70% and 30 respectively. The testing results for the model were 62.66%, 71.52%, 76.77% and 82.1% respectively for 6, 4, 3, and 2 classes. Their proposed method was compared with support vector machine and ELM methods and it was concluded that the integration of ELM and PSO performed better than SVM and ELM.

The study by Sharan et al. [24] investigated ECG-Derived Respiration (EDR) and Heart Rate Variability (HRV) from a single-lead ECG and applied them to two stages of sleep (sleep and awake). The MIT-BIH PSG database was used to analyze 18 ECG recordings. REM and NREM were classified as one class in the binary classification. In Sharan's et al study, a 5-minute window was used for feature extraction. A total of 74 features were extracted, 32 EDR frequency domain, 32 RR interval frequency domain, and 10 RR interval time domain features. The combination of HRV and EDR features achieved and accuracy of 80%, concluding the potential of using ECG for the screening of sleep disorder.

According to [25], a deep learning method has been developed for sleep scoring based on single-channel EEG data. To learn transition rules between sleep stages, their work used bidirectional long-short-term memory and time-variant features extracted from EEG epochs. A two-part classifier, consisting of sequence residual learning and representation learning, scores the 30s EEG epochs according to AAMI and R&K standards. Supratak et al. [25] utilized sleep data from two public datasets to classify five stages of sleep. PSG recording from 20 subjects from sleep-EDF and 62 subjects came from Montreal Archive of sleep studies. Their model archived an accuracy of 82.0% and Fi-score of 76.0% in the Mass dataset, while in sleep-EDF, the accuracy was 86.2% and F1-score of 81.7%.

2. **Materials and Methods**

2.1. Data acquisition

Data from MESA is used [26, 27]. These data were made available by the MESA Coordination Center. An online portal, www.sleepdata.org, is used to grant users permission and access to the dataset. The data includes PSGs and actigraphy raw data for black, white, Chinese-American, and Hispanic men and women aged between 45 and 84 years. PSGs were recorded following the AASM standard. The sampling frequency is 200 Hz for all signals. PSGs are annotated in 30-second non-overlapping epochs as one of the five-sleep stages: W, S1, S2, S3 and REM. In total, 2,237 subjects participated in an overnight sleep study with unattended PSGs [27], seven-day wrist-worn actigraphy, and sleep questionnaire. Data from 1700 sleeping nights were processed and used to develop the model (Table 1). Training, validation, and testing data are randomly divided into three subsets of 70:20:10. Training and validation sets are used to develop the model, while the test set is used to test it without being exposed to the model. The MESA sleep data is

advantageous for research due to several factors. First, the dataset includes a large and diverse sample size of thousands of participants, providing greater statistical power and the ability to detect small effects. Additionally, the MESA study is longitudinal, with data collected at multiple time points, allowing for an examination of changes in sleep patterns and health outcomes over time. The dataset also includes rich data on a variety of sleep parameters and disorders, promoting a comprehensive analysis of sleep patterns and their relationship to health outcomes. As a multi-ethnic study, the MESA data allows for the examination of potential ethnic differences in sleep patterns and health outcomes. We used data from the MIT-BIH Polysomnographic Database (SLPDB) to test the generalization of our model. The SLPDB is a collection of over 80 hours of physiological recordings obtained during sleep from individuals with Obstructive Sleep Apnea Syndrome (OSAS) who were monitored in a sleep laboratory. The database includes ECG, EEG, and respiration signals, which are annotated beat-by-beat and with respect to sleep stages and apnea. The purpose of the database is to evaluate the efficacy of Constant Positive Airway Pressure (CPAP) therapy in OSAS. It is a valuable resource for researchers and clinicians interested in studying sleep physiology, and developing new algorithms for sleep staging, apnea detection, and CPAP optimization.

Table 1
Number of different sleep segments per sleep

Sleep stage	5 stage segments	4 stage segments
W	717159	717159
S1	185474	(S1&S2) 971376
S2	785902	
S3	144703	(S3) 144703
REM	250830	250830

2.2. Annotation scoring

The data used in this study is part of PSG recording of sleep in the MESA database. Annotation for scoring sleep and other events in the MESA database is based on the AASM. A sleep expert scored every night signal. The data was divided into 30-second epoch and each epoch assigned to one of five sleep stage scores. The AASM scores sleep into five stages, W, S1, S2, S3, and REM. For epochs with more than one stage, the stage that reflects the greatest portion of the epoch is assigned to that stage. In the case of two sleep stages that are evenly distributed in one epoch and one of the stages is the same as the preceding epoch, the same stage as the preceding epoch is assigned to that epoch.

2.3. Input features preprocessing

The study applies 2 min segments of heart rate extracted from the ECG signals as input signals. To extract the features, first the signal is normalized, and the R-peak detection algorithm is applied. The IBI is calculated by finding the difference between consecutive R positions. The IHR is then calculated by taking the reciprocal of the interbeat interval. Each individual's heart rate is normalized separately by subtracting the mean and dividing by the standard deviation. Segments with less than 25 R-peaks per 30 sec epochs are removed to make sure that only data with uninterrupted beats is used. After removing segments with less R peak than normal, the time series is interpolated to 30 samples per 30 sec epochs to have an exact match between the segments and the annotations. The time-series data is resampled using linear interpolation with a sampling rate of 2 Hz. To smooth the time series, anomalous values exceeding five standard deviations that are the result of missed or false peaks are removed. Four 30-s epochs are merged together to form one segment for the input signals. Fig. 1 shows the signal processing flowchart for the proposed work.

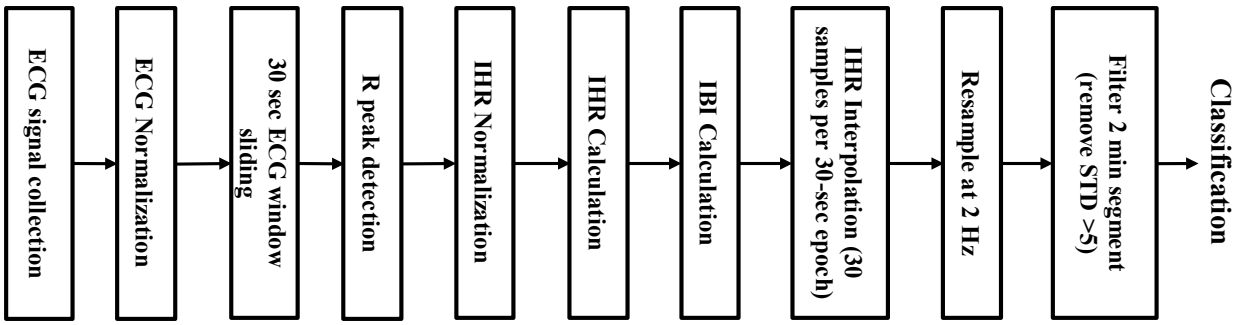

Fig. 1. Signal collection and feature extraction flowchart.

2.4. Algorithm

In this paper, we propose a sleep stage classification model that can be applied on mobile devices. To address the memory size issue encountered in our previous research [24], we utilized the ResNet architecture. Moreover, ResNet is recommended for addressing challenges during training of deeper neural networks.

The input to the model is an IHR with a shape of (240, 1) centred on 30-s epochs. The network architecture has nine layers, utilizing shortcut connections similar to ResNet [28]. Three residual blocks are constructed with three convolutional layers per block, followed by average pooling and a softmax layer. Every block consists of three convolutional layers followed by batch normalization and Rectilinear Unit (ReLU) between them. The training procedure is the same as that described in [29]. There are 64, 128, and 128 convolutional filters in each of the 3 blocks. The first layer is a batch normalization layer, which improves the performance of the model by normalizing the input data to have a zero mean and unit variance. The next layer is a convolutional layer with 64 filters, each with a size of 240 × 1. The output of this layer is passed through another batch normalization layer, followed by an activation function, in this case, ReLU. The next two layers are a pair of convolutional, batch normalization, and activation layers, each with 64 filters. The outputs of these two layers are added together via a skip connection. The next pair of convolutional layers are similar to the previous pair, but this time with 128 filters each. Again, the outputs are added together via a skip connection. Finally, there are two more convolutional layers with 128 filters each, followed by batch normalization and activation layers. The output of the final convolutional layer is the output of the model, which classifies between four stages of sleep (W, S1&S2, S3, and REM). Fig. 2 shows the structure of the model used in this research.

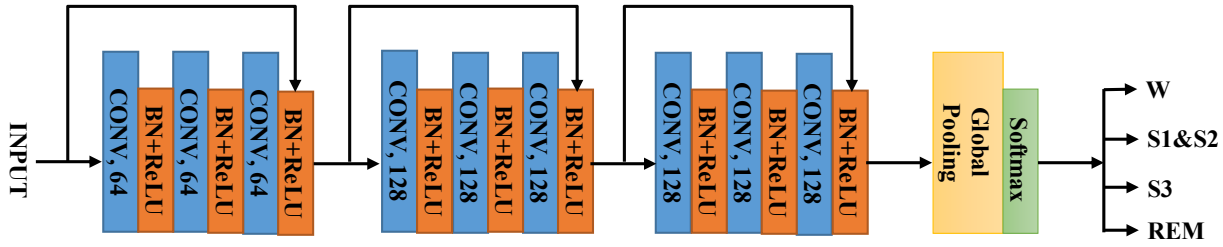

Fig. 2. Network structure**.**

CNN training can be a challenging and lengthy process due to its high computational requirements, particularly when working with larger and more complex models and datasets. As a result, high computing power, such as GPUs, is often required to achieve both accuracy and efficiency in the training process. In this case, a Nvidia GPU with 32 GB was utilized, hosted in a supercomputer provided by Yuan Ze University, to train a model with 150 epochs. The training process took 21 hours, and the resulting model size was 17.2 MB.

2.5. Performance evaluation

The classification performance of the model is measured using the accuracy, Cohen's Kappa, Sensitivity (Sens), Specificity (Sp) and the F-1 score of the 5-classes with the annotations as a reference. Training and validation are

carried out using 70% and 20% of the total data respectively, and 10% is set apart for testing. According to the CNN model, five probabilities are generated to predict the 30-s segment (epoch). The values correspond to the five classes, which are W, S1, S2, S3, and REM. The value with the highest probability is taken as the predicted stage. The confusion matrix table is used to calculate the performances of the model. For calculating performances in a multiclass classification, it is assumed that there are three stages of sleep S1, S2, and S3. A confusion matrix table to visualize the performance example for the three stages model is shown in Table 2. Table 3 summarizes the equations of features used to evaluate the proposed model.

Table 2.
Example of the confusion matrix

	Predicted			
Actual	Sleep stage	S1	S2	S3
	S1	PS11	PS21	PS31
	S2	PS12	PS22	PS32
	S3	PS13	PS23	PS33

Table 3
Equation of features used in the proposed method.

Row	Features	Equation	Row	Features	Equation
1	True	$TP(S1) = P_{S22} + P_{S32} + P_{S23} + P_{S33}$	11	*Sens*	$Sens = \frac{TP}{TP+FN}$ x 100%
2	Positive	$TP(S2) = P_{S11} + P_{S31} + P_{S13} + P_{S33}$			
3	rate	$TP(S3) = P_{S11} + P_{S21} + P_{S12} + P_{S22}$	12	*Sp*	$Sp = \frac{TN}{TN + FP}$ x 100%
4	False	$FP(S1) = P_{S12} + P_{S13}$	13	*Kappa*	$Kappa = \frac{P_0 - P_e}{1 - P_e}$ x 100%
5	Positive	$FP(S2) = P_{S21} + P_{S23}$			
6	rate	$FP(S3) = P_{S31} + P_{S32}$	14	*Po*	$Po = \frac{TP + TN}{TP + FP + TN + FN}$
7	False	$FN(S1) = P_{S21} + P_{S31}$			
8	Negative	$FN(S2) = P_{S12} + P_{S32}$	15	*Pe*	$Pe = (TN + FP) \times (TN + FN) + (FN + TP) \times (FP + TP)$
9	rate	$FN(S3) = P_{S13} + P_{S23}$			
10	Accuracy	$Acc = \frac{TP + TN}{PT + FN + TN + FP}$ x 100%	16	*F1*	$F1 = 2 \times \frac{Sens \times Sp}{Sens + Sp}$ x 100%

where TN, FP, and FN represent true negative, false-positive, and false-negative values, respectively, for each sleep stage. TP is calculated the same way as in binary classification. Multiclass classification, however, calculates the true positive values for each class under investigation. The TP values for the classes in Table 2 are P_{S11}, P_{S22} and P_{S33}. Cohen's kappa measures the degree of agreement between two raters categorizing items into mutually exclusive groups. The po and pe variables used in Kappa equation are the observed proportionate agreement and the probability of random agreement, respectively, as provided in equations 14 and 15. A kappa value of 1 indicates complete agreement between the raters. The kappa value is zero if the raters do not agree. There is a possibility that the statistic will be negative. This can happen by chance if the ratings of the two raters are unrelated. The model is validated using fivefold cross-validation.

3. Results

The features of the time series input were extracted by training a three-block ResNet (Fig. 2). Using two-minute IHR time series, the network classifies five stages of sleep for every 30-second interval, including W, S1, S2, S3 and REM. Due to poor predictions in S1, S1 and S2 are merged into a single stage. The model is trained, validated, and tested using data from the MESA. Initially, 500 subjects' data are selected from the database for training. The number

of subjects' data are varied from 500 to 1000 in order to see how it affected sleep classification. A common difference between 100 subjects is used as an arithmetic progression sequence. Furthermore, additional subjects are added in the training and testing including 1500 and 1700 subjects to advance the study. Different sets of labels are tested including four and five labels in the pursuit of finding the effect of adding more subjects and the results are shown in Tables 4 and 5. This work is evaluated based on the overall accuracy, Kappa value, Sens, Sp, and F-1 score. Table 4 shows the performance results for finding the effect of adding more subjects to training stages of sleep.

Table 4
The results for sleep stage to find the effect of adding data in the classification of 4 and 5 sleep stages

	5-stages		4-stages	
Subjects	Acc	Kappa	Acc	Kappa
500	73.35%	61.96%	72.59%	57.91%
600	72.90%	61.56%	78.96%	67.15%
700	71.01%	58.07%	78.27%	65.98%
800	72.29%	60.13%	77.20%	64.04%
900	70.96%	58.77%	77.94%	66.38%
1000	74.18%	63.36%	78.12%	66.14%
1500	74.69%	64.00%	80.11%	68.86%
1700	77.34%	67.58%	84.17%	75.29%

Looking at the results for the five stages of sleep, a consistently low performance was observed in the performance evaluation of Sens, Sp, and F-1 score. This is because there is relatively low number of segments in S1 compared to the rest of the stages, which lead us to merge the S1 with S2 as per practice with other researchers [30, 31, 32]. The performance evaluation using Sens, Sp, and F-1 for finding the effect of adding more subjects when training the four and five stages of sleep is shown in Table 5.

Table 5
The results for sleep stage using different sets of subjects to find the effect of adding data in the classification.

	5-stages			4-stages		
Subjects	Sens	Sp	F-1	Sens	Sp	F-1
500	63.25±27.31%	64.47±16.83%	63.12±23.16%	67.63±15.25%	66.92±16.47%	66.2±12.33%
600	64.10±24.93%	63.53±19.77%	63.50±22.58%	74.16±9.06%	74.41±12.70%	74.12±10.24%
700	58.78±25.45%	63.11±19.86%	60.30±22.57%	71.94±14.01%	73.98±9.39%	72.67±11.07%
800	62.60±26.26%	62.93±19.14%	62.26±23.35%	68.95±16.05%	73.45±8.16%	70.87±12.26%
900	61.43±26.83%	61.01±17.83%	60.45±23.37%	71.67±13.08%	75.76±7.42%	73.36±9.52%
1000	64.67±24.06%	66.35±20.10%	65.36±22.22%	73.23±12.24%	72.51±11.34%	72.74±11.30%
1500	64.97±24.70%	66.93±19.01%	65.7±22.20%	74.36±11.87%	76.68±7.77%	75.43±9.84%
1700	67.17±26.62%	70.30±16.31%	68.03±22.69%	79.52±9.98%	81.3±6.94%	80.36±8.43%

An example of the sleep stage IHR, hypnograms that were manually scored by experts, hypnograms that were automatically scored by our model, as well as predicted IHR segments for each of the four sleep stages for one subject using the MESA dataset is shown in Fig. 3.

An equivalence of 170 subjects' data (208409 segment) is set apart for testing the 4 and 5 classes of sleep. The study used 1700 subjects' data, of which 70%, 20%, and 10% were used for training, validation, and test, respectively. To evaluate the performance of the model, the results from testing data were used. To compare predicted classification results with actual classification results, a confusion matrix table was created.

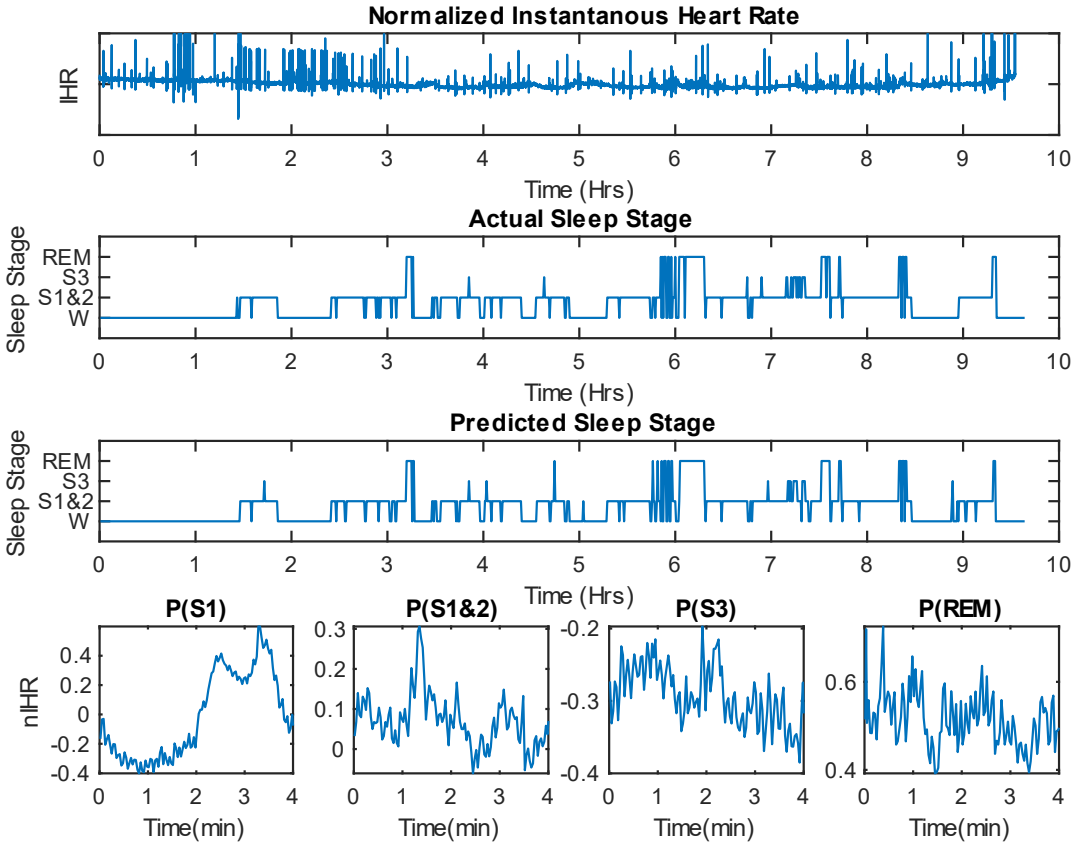

Fig. 3. Sleep stage chart for one of the subjects in the MESA dataset over the course of one night. (a) The Unfiltered IHR, used as input for a classifier after being segmented. (b) A hypnogram of the sleep stage as manually scored by experts according to the four stages of sleep. (c) A hypnogram of the sleep stage as automatically scored by the model. (d) An example of correctly classified 2-minute IHR segments from each sleep stage.

In Tables 6 and 7, the confusion matrices of the proposed model applied to the MESA datasets are shown. To calculate the confusion matrix, the scores for each of the test cases are added together. The rows represent the number of samples scored by experts in their respective sleep stages, while the columns reflect the number of samples predicted by the proposed model. Table 8 tests the generalization of the proposed model by testing it on a different dataset (SLPDB).

Table 6
Normalized classification results of each of the classes in the 5-stage classifier.

Stage	Normalized					Full count				
	W	S1	S2	S3	REM	W	S1	S2	S3	REM
W	86.30%	23.61%	5.85%	3.75%	7.36%	61892	4380	4599	543	1846
S1	2.91%	22.58%	3.80%	0.30%	1.93%	2090	4189	2989	43	484
S2	7.32%	46.60%	83.73%	32.00%	9.27%	5249	8643	65801	4631	2324
S3	0.62%	0.95%	4.04%	62.49%	0.69%	446	177	3174	9043	174
REM	2.84%	6.25%	2.58%	1.46%	80.75%	2039	1159	2028	211	20255

Table 7
Normalized classification results of each of the classes in the 4-staage classifier (MESA data).

	Normalized				Full count			
Stage	Awake	S1	S2	REM	Awake	S1	S2	REM
Awake	87.95%	6.41%	2.25%	6.08%	63073	6228	326	1524
S1	9.71%	88.58%	26.31%	8.26%	6967	86049	3808	2073
S2	0.31%	3.01%	70.62%	0.35%	225	2922	10219	89
REM	2.02%	2.00%	0.82%	85.30%	1451	1940	118	21397

Table 8
Normalized classification results of each of the classes in the 4-staage classifier (sleep).

	Normalized				Full count			
Stage	Awake	S1	S2	REM	Awake	S1	S2	REM
Awake	63.42%	22.76%	16.16%	29.99%	1938	1288	107	209
S1	26.87%	67.11%	63.75%	38.16%	821	3798	422	266
S2	1.73%	2.40%	15.41%	0.86%	53	136	102	6
REM	7.98%	7.72%	4.68%	30.99%	244	437	31	216

Table 6 presents results of testing the 5 stages of sleep using the confusion matrix table. The results show the normalized count and the full count respectively. Table 7 shows the results for the 4-stages of sleep after merging S1 and S2 for the 1700 subjects. The evaluation of testing the model used for training the 5-stages of sleep in the 1700 subjects gives an overall accuracy of 77.34%, and Kappa of 67.58% as shown in Table 4. Evaluating using Sens give a mean and standard deviation of 67.17% ± 26.62%. Sp gives a mean and standard deviation of 70.30% ± 16.31%, while F-1 score results are 68.03% ± 22.69%. In order to improve the accuracy of the model, S1 and S2 were merged to reduce the number of sleep stages. Evaluating the accuracy of the model for 4 stages using the overall accuracy provides 84.17% and kappa is 75.29% respectively and is provided in Table 4. The Sens gives a mean and standard deviation of 79.52% ± 9.98%. Sp provides a mean and standard deviation of 81.3% ± 6.94 %. Evaluation using the F-1 score gives a mean and standard deviation of 80.36% ± 8.43%.

In Table 9, the evaluation results obtained for the four stages of sleep cross-validation using 1,700 subjects are presented. The evaluation results include the mean and standard deviations for Sens, Sp, and F1. Table 10 presents the mean and standard deviations for accuracy Kappa, mean Sens and Sp for all the classes, along with the F-1 score. Additionally, Table 10 presents the results for testing the model using data from the SLPDB. The means and standard deviation for accuracy are 85.3% ± 1.12%, kappa is 78.9% ± 1.77%, Sens is 81.14% ± 1.57%, Sp is 82.68% ± 1.37%, and F-1 score is 81.87% ± 1.48%, respectively. The accuracy for data from the SLPDB is 60.10%, kappa is 44.70%, Sens is 45.98%, Sp is 44.23%, and F-1 score is 45.09%.

Table 9
Five-fold cross-validation results of the 4-stage classifier using the Sens, Sp and F1-score.

	CV-1			CV-2			CV-3			CV-4			CV-5		
	Sens	Sp	F-1	Sens	Sp	F-1	Sens	Sp	F-1	Sens	Sp	F-1	Sens	Sp	F-1
Mean (%)	79.52	81.3	80.36	79.73	81.38	80.5	83.11	84.38	83.72	82.28	83.67	82.95	81.05	82.68	81.82
SD (%)	9.98	6.94	8.43	10.29	6.45	8.44	8.45	5.73	7.12	8.76	5.99	7.4	9.53	6.18	7.89

Table 10
Classification results of the 4-stage classifier using the 5-fold cross-validation

5-fold CV	Acc	Kappa	Sens	Sp	F-1

Mean	85.32%	77.11%	81.14%	82.68%	81.87%
SD	1.12%	1.77%	1.57%	1.37%	1.48%
SLPDB	60.10%	44.70%	45.98%	44.23%	45.09%

4. Discussion

The study focuses on using ResNet architecture to classify sleep stages into four categories based on ECG signals. The IHR is calculated from the ECG signals of 1700 participants to demonstrate that ECG contains information about sleep stages. Using IHR is justified by the fact that ECG can be extracted using various wearable sensors, some of which are readily accessible to consumers. ECG signals are a cost-effective and easy-to-use alternative to more expensive EEG signals for sleep stage classification. They provide information on heart rate variability, which is associated with autonomic nervous system activity and can be combined with other physiological signals for better classification accuracy. However, more research is needed to validate their use and improve their effectiveness.

To address the challenges associated with improving classification accuracy, researchers have tried to use more training data and deep CNNs. Adding more layers to a CNN model can lead to longer training times, higher computational costs, and overfitting. Therefore, ResNet architecture is used, which allows for the addition of more layers while keeping the model size relatively small. ResNet uses residual connections to skip certain layers and pass the output of one layer directly to another layer further down the network. This approach has been shown to be effective in improving the accuracy of ECG-based sleep stage classification. The goal is to implement the model for sleep stage classification on smartphones, which is feasible with the solutions proposed above.

Rather than extracting features manually from ECG data, we used CNNs to learn local cardiac features automatically. This approach saves time and eliminates the need for researchers to calculate and interpret these features themselves. By analyzing the raw ECG signal, CNNs can extract key features such as amplitude, frequency, and duration of cardiac activity without human intervention.

Datasets for training are taken from the MESA database, which included heart rate features based on ECG analysis. The MESA database is one of the largest databases used in sleep studies and is annotated according to 30-second intervals. The sleep stage study has not made the much-anticipated breakthrough that other researchers have, and that is largely due to the lack of data. Even though participants usually provide at least 10 hours of sleep data, studies have shown that even more participants are needed to strengthen the diverse features required for a study of sleep. As a test to determine whether sufficient data is available to diagnose sleep, the number of participants is varied from 500 to 1000. Six datasets are created using a sequence with a difference of 100 participants. In addition, two sets of 1500 and 1700 participants are included. This study was conducted for the five stages of sleep and the results are presented in Tables 4 and 5. Across the 5-sleep stage classification, Table 4 indicates fluctuating accuracy from 500 to 900 participants and a consistent rise from 1000 to 1700 participants. Another notable change is the decrease in mean for the Sens, Sp, and F_1-score, as well as the increase in standard deviation in all five stages. This is largely due to the model's low performance at S1, which is often misclassified as classed W, S2, S3, and REM. The majority of misclassifications in S1 are in S2, which is the majority class in all NREM sleep stages (S1, S2, and S3). In the MESA dataset, class W had the best performance for the 5-stage classification. However, it is not the majority class. Across the different datasets, the majority of misclassifications are in class S2, which is the majority class.

To improve model performance and determine whether the low performance in stage S1 interferes with the investigation in Tables 4 and 5, the model is trained to classify sleep in four stages. Stages 1 and 2 were merged to achieve this. Table 4 with the 4 sleep stages shows an increase in accuracy and kappa from 600 to 1700 participants, although there is still some variation in accuracy in the early stages of the classification. The mean Sens, Sp, and F-1 for the four stages are also improved, and the standard deviation is low in comparison with results shown in Table 5. The results in Tables 4 and 5 indicate that more participants are needed in the sleep stage study.

Sleep stage classification is a difficult task, with challenges including inter- and intra-subject variability, noise in signals, and unclear definitions of sleep stages [33, 34]. These challenges can cause classification errors due to EEG signal variability, ambiguity in stage definitions, artifact contamination, and inadequate feature selection [35, 36, 37].

To reduce classification errors, researchers have proposed several techniques. Feature extraction methods, including time-frequency and wavelet analyses, can extract more relevant and distinctive features. Ensemble classifiers, which combine multiple classifiers, can improve accuracy by reducing the impact of misclassifications and signal variability. Incorporating expert knowledge, such as sleep expert annotations, can improve accuracy. Data augmentation, including adding artificial noise or shifting signals, can increase the diversity and size of training data. Transfer learning techniques, which utilize pre-trained models, can improve the accuracy of sleep stage classification models with limited training data [38, 39, 40, 41, 42]. Overall, sleep stage classification is challenging, but researchers can improve accuracy by utilizing these techniques, which can lead to better understanding of sleep patterns and disorders.

The proposed method is compared with other works conducted with different databases, numbers of participants, and types of signals, sleep stages and methods. A summary of the comparison is shown in Table 11. Comparisons include accuracy, TPR, and FPR obtained from their methods. Studies have attempted to stage sleep using ECG features in the past. With the growing interest in the field and the availability of large sample size databases, this method is starting to gain popularity. The use of ECG features to stage sleep also has the advantage that some of the tools are easily accessible, so data can be extracted more easily. The results of Table 11 also highlight the advantage of using ECG features, since more researchers are able to stage sleep on more than 100 participants.

Table 11
Comparison with other research.

Author	Database	Signal	Participants	Stages	Method	Acc (%)	TPR	FPR
[43]	SHHS	ECG and respiratory	8682 (7208)	5, 3	CNN+LSTM	72.54, 88.03	0.76, 0.87	0.13, 0.07
[44]	SHHS, MESA	ECG (IHR)	10,000	4	CNN	SHHS,78 MESA,80	0.76 -	0.08 -
[45]	Siesta	ECG (HRV)	292	4	LSTM	74.65	-	-
[46]	RF-Sleep	RFS	100	4	CNN-RNN	79.80	0.67	0.12
[47]	Sleep-EDF Sleep-EDF-20 Sleep-EDF-78 SHHS	EEG	8 20, 78, 329	5 5	FKSVM AttnSleep	90.20 85.60, 82.90, 86.60	- 0.79, 0.76, 0.75	- 0.04, 0.05, 0.04
[23]	MITBPD	ECG (HRV)	18	6, 4, 3, 2	ELM/PSO	62.66, 71.52, 76.77, 82.10	-	-
[48]	SHHS	EEEG+ ECG+ EMG	8682		XGBoost	85.30	0.85	-
[49]	Apnea-ECG	ECG	35	2	CWT+SVM	91.40	0.89	-
[50]	Apnea-ECG	ECG	32	2	ResNet+ Multiscale	86.00	0.84	-
[51]	SHHS+MESA	ECG	100	4	Multi-Scale Residual Adaptive Network	84.90, 82.70	0.86, 0.83	0.05, 0.05
Proposed	MESA	ECG (IHR)	1700	5 4	ResNet	77.34, 85.32	- 0.89	- 0.05

SVM: Support Vector Machine, RotSVM: Rotational support Vector Machine, MITBPD: MIT-BIH polysomnographic database, LSTM: long short-term memory, RNN: recurrence neural network, AttnSleep: attention-based deep learning

architecture, ELM: Extreme Learning Machine, PSO: Particle Swarm Optimization RFS: Radio frequency spectrogram. XGBoost: Extreme Gradient Boosting

According to Table 11, ECG-based sleep staging in 5 stages is not popular as more researchers prefer staging sleep in 4 stages. It is primarily due to the low performance caused by the high data imbalance, as shown in Table 6. Based on the results from Table 6, S1 has a very low accuracy rate and does not have any distinct features of its own. It may be possible to conduct a study on whether the lack of data in S1 is a contributing factor to its poor performance using the SHHS database. In the SHHS database there is more data that may improve the size of all stages and provide a basis for a study involving balancing the sizes of all stages.

A comparison of the results of the proposed work with previous multi-classification works based on features extracted from ECG signals is provided. In [43], researchers used a combination of an ECG and abdominal breathing to stage sleep using CNN and LSTM networks. In 5-stage classification, the proposed method outperformed the previous method by 4.80%. Due to the different number of sleep stages pursued in both studies, research on the other sleep stages is not comparable. In comparison to [44], ResNet performed better in the SHHS and MESA databases by 7.32% and 5.32%, respectively. According to [45], the proposed method performed better than their LSTM-based method by 10.67%. Researcher in [23] also used ELM and PSO in combination to study sleep stages based on ECG HRV features. In spite of using different sets of stages, our method performed better than all of them. When it comes to the 3, 4, and 6 stages of sleep, our method beats theirs in each of the four and five stages. For the 4-stages, the accuracy rate was 22.66%, 13.8%, and 8.55%, and for the 5-stages, it was 14.68%, 5.82% and 0.57% better than their method. Despite some of them being less than the proposed method, the proposed method in four stages performed better than all the sets of stages examined in [23].

In addition, the proposed study is compared with studies that investigated sleep stage studies using features from other types of signals. Researchers [47, 51] used EEG signals to investigate sleep in 5-stages. In [51], the researchers used a rotational support vector machine model trained on data from eight participants to classify sleep stage. In spite of training their model on twenty participants' data, they outperformed the proposed model by 5.78%. When compared to research from [47], the proposed work performed better than their proposed work based on sleep-EDF database data on 20 and 78 participants. Their study on the SHHS dataset outperformed the proposed work by 1.28%. Research conducted by [46] on the radio frequency spectrogram performed better than the proposed study by 5.52%.

The performance of the proposed classifier in terms of TPR and FPR is also presented in Table 11. Comparing the TPR of our study with the rest of the studies in Table 11 suggests that our work is superior to most, with the exception of [49], which has the same TPR detection. Two studies performed better in terms of FPR compared to ours. Three of the studies are level with our study in terms of FPR, and the rest were slightly poorer. It is worth mentioning that most of the TPR and FPR values were not provided in the studies. We had to calculate them ourselves from the data that was provided. Most of the data provided is not the final assessment by the studies, and some is from experimentations they conducted.

ResNet has been combined with other networks to classify sleep in other studies [50, 51]. Compared to these two studies, our classifier performed slightly better in terms of accuracy. One noticeable advantage of our study is the amount of data used, compared to the majority of the studies in Table 11. Although our proposed study produced promising results in sleep stage classification, it is worth mentioning that the model's generalization is not good. The SLPDB dataset was used to test the proposed model, producing an accuracy of 60.1%. This means that the model is unable to perform well on new, unseen data, despite being trained on a large dataset. Overfitting, where the model learns the noise or idiosyncrasies of the training data too well, is a common cause of poor generalization. In this case, utilized the SLPDB dataset, which comprises a distinct patient population, to evaluate the efficacy of our model on data from a diverse patient group. We selected the SLPDB dataset as it consists of sleep data for patients with OSAS, and we wanted to determine how our model would perform on unfamiliar data. Our model's results on the SLPDB dataset were satisfactory and reasonable (although an accuracy of 60.1% only), because it had not been exposed to the features of sleep data from individuals with sleep apnea. To enhance the model's generalization capacity to classify sleep stages in individuals of all types, including those with sleep apnea, we will integrate this dataset into our training data in future studies. Furthermore,

we plan to experiment with different models to identify an appropriate classifier that satisfies our needs. We also intend to modify the ResNet model's architecture to optimize its suitability for the data.

Filtering is a crucial step in removing outside influences from the data that would interfere with the analysis. Researchers have used the threshold from the five standard deviations to remove outliers in similar features. In their filtering process, [44] removed values greater than five standard deviations per night. Although the filter is effective, the threshold is high in cases such as the one shown in Fig. 3 (a), which may render it ineffective. In order to compensate for that, we calculated the threshold segment by segment and removed anomalous values >5 standard deviations in every segment.

5. Conclusion

This paper proposes a 1-D automatic sleep stage detection method based on the IHR. The MESA dataset was used to acquire ECG signals. Preprocessing of ECG data includes normalization and segmentation, R-peak detection to calculate the inter-beat interval, IHR calculation and normalization, resampling, and filtering. In order to ensure that the annotations provided match the segments perfectly, interpolation is used. Anomalies greater than five standard deviations were filtered from IHR segments. Sleep staging is based on the 2-minute segments of the IHR inputs to the three blocks ResNet. An experiment is conducted that varied the number of subjects in the training of sleep staging in order to explore the effect of adding more subjects. Based on the five-fold cross-validation, the 4-stage classifier has an accuracy of 85.32%, a kappa of 77.11%, a Sens of 81.14%, a Sp of 82.68%, and an F-1 score of 81.87%. The proposed method performed well on the MESA dataset. In the future, other databases will be incorporated in order to add more subjects since Tables 4 and 5 indicate that adding more subjects improves the accuracy of the classifier.

Acknowledgements

This study was financially supported by the Ministry of Science and Technology, Taiwan (Grant number: MOST 107-2221-E-155-009-MY2).

Author contributions

B.M.M., C.-H.L., and Y.-T.L. developed the algorithms and analyzed the data. B.M.M. wrote the paper. J.-S.S. and M.F.A. evaluated and supervised this study.

Conflicts of interest

The authors declare that they have no conflicts of interest.

References

1. Berry RB, Budhiraja R, Gottlieb DJ, Gozal D, Iber C, Kapur VK, et al. Rules for scoring respiratory events in sleep: update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events. Deliberations of the Sleep Apnea Definitions Task Force of the American Academy of Sleep Medicine. J Clin Sleep Med. 2012;8(5):597-619.
2. Armon C, Roy A, Nowack W. Polysomnography: Overview and clinical application. E-Medicine March. 2007.
3. Shepard JW. Atlas of sleep medicine. Futura Publications Ltd.: London, UK. 1991.
4. Shrivastava D, Jung S, Saadat M, Sirohi R, Crewson K. How to interpret the results of a sleep study. Journal of community hospital internal medicine perspectives. 2014;4(5):24983.
5. Iber C, Ancoli-Israel S, Chesson A, Quan S. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications: American Academy of Sleep Medicine. Westchester, IL. 2007.
6. Chokroverty S, Bhatt M, Goldhammer T. Polysomnographic recording technique. Atlas of sleep medicine. Elsevier Inc. https://doi. org/10.1016/B978-0-7506-7398-3.50005-X; 2005.
7. Prasad B, Carley DW, Herdegen JJ. Continuous positive airway pressure device-based automated detection of obstructive sleep apnea compared to standard laboratory polysomnography. Sleep and Breathing. 2010;14:101-7.

8. Kapur V, Strohl KP, Redline S, Iber C, O'connor G, Nieto J. Underdiagnosis of sleep apnea syndrome in US communities. Sleep and Breathing. 2002;6(02):049-54.
9. Moore M, Gelfeld B, Okunogbe A, Paul C. Identifying future disease hot spots: infectious disease vulnerability index. Rand health quarterly. 2017;6(3).
10. Faust O, Acharya UR, Ng E, Fujita H. A review of ECG-based diagnosis support systems for obstructive sleep apnea. Journal of Mechanics in Medicine and Biology. 2016;16(01):1640004.
11. Hassan AR, editor Automatic screening of obstructive sleep apnea from single-lead electrocardiogram. 2015 international conference on electrical engineering and information communication technology (ICEEICT); 2015: IEEE.
12. Chachadi K, Nirmala S, editors. Gender recognition from speech signal using 1-D CNN. Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2021; 2022: Springer.
13. Deepak S, Ameer P. Automated categorization of brain tumor from mri using cnn features and svm. Journal of Ambient Intelligence and Humanized Computing. 2021;12:8357-69.
14. Hasan MM, Salehin I, Moon NN, Kamruzzaman T, Hasan M, editors. A computer vision system for the categorization of citrus fruits using convolutional neural network. 2021 International Symposium on Electronics and Smart Devices (ISESD); 2021: IEEE.
15. Melekoodappattu JG, Dhas AS, Kandathil BK, Adarsh K. Breast cancer detection in mammogram: Combining modified CNN and texture feature based approach. Journal of Ambient Intelligence and Humanized Computing. 2022:1-10.
16. Švorc D, Tichý T, Růžička M. An infrared video detection and categorization system based on machine learning. Neural Network World. 2021;31(4):261.
17. He K, Zhang X, Ren S, Sun J, editors. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.
18. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al., editors. Microsoft coco: Common objects in context. Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13; 2014: Springer.
19. McGrogan N, Braithwaite E, Tarassenko L, editors. Biosleep: a comprehensive sleep analysis system. 2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 2001: IEEE.
20. Su C-H, Ko L-W, Juang J-C, Hsu C-Y. Hybrid sleep stage classification for clinical practices across different polysomnography systems using frontal EEG. Processes. 2021;9(12):2265.
21. Biswal S, Sun H, Goparaju B, Westover MB, Sun J, Bianchi MT. Expert-level sleep scoring with deep neural networks. Journal of the American Medical Informatics Association. 2018;25(12):1643-50.
22. Garcia-Molina G, Jiang J. Interbeat interval-based sleep staging: work in progress toward real-time implementation. Physiological Measurement. 2022;43(2):025004.
23. Surantha N, Lesmana TF, Isa SM. Sleep stage classification using extreme learning machine and particle swarm optimization for healthcare big data. Journal of Big Data. 2021;8(1):1-17.
24. Sharan RV, editor ECG-derived respiration for sleep-wake stage classification. 2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET); 2021: IEEE.
25. Supratak A, Dong H, Wu C, Guo Y. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2017;25(11):1998-2008.
26. Chen X, Wang R, Zee P, Lutsey PL, Javaheri S, Alcántara C, et al. Racial/ethnic differences in sleep disturbances: the Multi-Ethnic Study of Atherosclerosis (MESA). Sleep. 2015;38(6):877-88.
27. Zhang G-Q, Cui L, Mueller R, Tao S, Kim M, Rueschman M, et al. The National Sleep Research Resource: towards a sleep data commons. Journal of the American Medical Informatics Association. 2018;25(10):1351-8.

28. He K, Zhang X, Ren S, Sun J, editors. Identity mappings in deep residual networks. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14; 2016: Springer.
29. Wang Z, Yan W, Oates T, editors. Time series classification from scratch with deep neural networks: A strong baseline. 2017 International joint conference on neural networks (IJCNN); 2017: IEEE.
30. Liang Z, Chapa-Martell MA. A multi-level classification approach for sleep stage prediction with processed data derived from consumer wearable activity trackers. Frontiers in Digital Health. 2021;3:665946.
31. Phan H, Do Q, Do T-L, Vu D-L, editors. Metric learning for automatic sleep stage classification. 2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC); 2013: IEEE.
32. Radha M, Fonseca P, Moreau A, Ross M, Cerny A, Anderer P, et al. A deep transfer learning approach for wearable sleep stage classification with photoplethysmography. NPJ digital medicine. 2021;4(1):135.
33. Satapathy S, Loganathan D, Kondaveeti HK, Rath R. Performance analysis of machine learning algorithms on automated sleep staging feature sets. CAAI Transactions on Intelligence Technology. 2021;6(2):155-74.
34. Tsinalis O, Matthews PM, Guo Y, Zafeiriou S. Automatic sleep stage scoring with single-channel EEG using convolutional neural networks. arXiv preprint arXiv:161001683. 2016.
35. Eltrass AS, Ghanem NH. Investigation of automatic spindle detection in sleep EEG signals contaminated with noise and artifact sources. Journal of Ambient Intelligence and Humanized Computing. 2022:1-22.
36. Satapathy SK, Loganathan D. Automated classification of sleep stages using single-channel EEG: A machine learning-based method. International Journal of Information Retrieval Research (IJIRR). 2022;12(2):1-19.
37. Shen N, Luo T, Chen C, Zhang Y, Zhu H, Zhou Y, et al. Towards an automatic narcolepsy detection on ambiguous sleep staging and sleep transition dynamics joint model. Journal of Neural Engineering. 2022;19(5):056009.
38. Abdulla S, Diykh M, Siuly S, Ali M. An Intelligent Model Involving Multi-Channels Spectrum Patterns Based Features for Automatic Sleep Stage Classification. International Journal of Medical Informatics. 2023:105001.
39. He Z, Du L, Wang P, Xia P, Liu Z, Song Y, et al. Single-channel EEG sleep staging based on data augmentation and cross-subject discrepancy alleviation. Computers in Biology and Medicine. 2022;149:106044.
40. Yang C-Y, Chen P-C, Huang W-C. Cross-Domain Transfer of EEG to EEG or ECG Learning for CNN Classification Models. Sensors. 2023;23(5):2458.
41. You Y, Zhong X, Liu G, Yang Z. Automatic sleep stage classification: A light and efficient deep neural network model based on time, frequency and fractional Fourier transform domain features. Artificial Intelligence in Medicine. 2022;127:102279.
42. Zhang H, Wang X, Li H, Mehendale S, Guan Y. Auto-annotating sleep stages based on polysomnographic data. Patterns. 2022;3(1):100371.
43. Sun H, Ganglberger W, Panneerselvam E, Leone MJ, Quadri SA, Goparaju B, et al. Sleep staging from electrocardiography and respiration with deep learning. Sleep. 2020;43(7):zsz306.
44. Niranjan S, Shoeb A, Stephens P, Alaa K, Shimol DB, Burkart J, et al. Deep learning for automated sleep staging using instantaneous heart rate. NPJ Digital Medicine. 2020;3(1).
45. Radha M, Fonseca P, Ross M, Cerny A, Anderer P, Aarts RM. LSTM knowledge transfer for HRV-based sleep staging. arXiv preprint arXiv:180906221. 2018.
46. Zhao M, Yue S, Katabi D, Jaakkola TS, Bianchi MT, editors. Learning sleep stages from radio signals: A conditional adversarial architecture. International Conference on Machine Learning; 2017: PMLR.
47. Eldele E, Chen Z, Liu C, Wu M, Kwoh C-K, Li X, et al. An attention-based deep learning approach for sleep stage classification with single-channel EEG. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2021;29:809-18.
48. Choi J, Kwon S, Park S, Han S. Validation of the influence of biosignals on performance of machine learning algorithms for sleep stage classification. Digital Health. 2023;9:20552076231163783.
49. Fang H, Lu C, Hong F, Jiang W, Wang T. Sleep Apnea Detection Based on Multi-Scale Residual Network. Life. 2022;12(1):119.

50. Zhang Z, Tang M. A Domain-Based, Adaptive, Multi-Scale, Inter-Subject Sleep Stage Classification Network. Applied Sciences. 2023;13(6):3474.
51. Alickovic E, Subasi A. Ensemble SVM method for automatic sleep stage classification. IEEE Transactions on Instrumentation and Measurement. 2018;67(6):1258-65.