Physiological Measurement

PAPER

Interbeat interval-based sleep staging: work in progress toward real-time implementation

Gary Garcia-Molina and Jiewei Jiang
Sleep Number Labs, San Jose, CA, United States of America
E-mail: gary.garciamolina@sleepnumber.com
Keywords: sleep-staging algorithm, real-time, cardiac interbeat interval, deep neural network, sleep disorders

Abstract
Objective. Cardiac activity changes during sleep enable real-time sleep staging. We developed a deep neural network (DNN) to detect sleep stages using interbeat intervals (IBIs) extracted from electrocardiogram signals. Approach. Data from healthy and apnea subjects were used for training and validation; 2 additional datasets (healthy and sleep disorders subjects) were used for testing. R-peak detection was used to determine IBIs before resampling at 2 Hz; the resulting signal was segmented into 150 s windows (30 s shift). DNN output approximated the probabilities of a window belonging to light, deep, REM, or wake stages. Cohen’s Kappa, accuracy, and sensitivity/specificity per stage were determined, and Kappa was optimized using thresholds on probability ratios for each stage versus light sleep. Main results. Mean (SD) Kappa and accuracy for 4 sleep stages were 0.44 (0.09) and 0.65 (0.07), respectively, in healthy subjects. For 3 sleep stages (light+deep, REM, and wake), Kappa and accuracy were 0.52 (0.12) and 0.76 (0.07), respectively. Algorithm performance on data from subjects with REM behavior disorder or periodic limb movement disorder was significantly worse, with Kappa of 0.24 (0.09) and 0.36 (0.12), respectively. Average processing time by an ARM microprocessor for a 300-sample window was 19.2 ms. Significance. IBIs can be obtained from a variety of cardiac signals, including electrocardiogram, photoplethysmography, and ballistocardiography. The DNN algorithm presented is 3 orders of magnitude smaller compared with state-of-the-art algorithms and was developed to perform real-time, IBI-based sleep staging. With high specificity and moderate sensitivity for deep and REM sleep, small footprint, and causal processing, this algorithm may be used across different platforms to perform real-time sleep staging and direct intervention strategies. Novelty & Significance (92/100 words) This article describes the development and testing of a deep neural network-based algorithm to detect sleep stages using interbeat intervals, which can be obtained from a variety of cardiac signals including photoplethysmography, electrocardiogram, and ballistocardiography. Based on the interbeat intervals identified in electrocardiogram signals, the algorithm architecture included a group of convolution layers and a group of long short-term memory layers. With its small footprint, fast processing time, high specificity and good sensitivity for deep and REM sleep, this algorithm may provide a good option for real-time sleep staging to direct interventions.

Introduction

Human sleep follows a cyclical pattern in which periods of non-rapid eye movement (NREM) sleep and rapid eye movement (REM) sleep alternate with an approximate periodicity of 90 min. Typically, a nocturnal sleep session comprises between 4 and 6 sleep cycles (Carskadon and Dement 2011, Patel et al 2021). NREM and REM sleep occupy 75%–80% and 20%–25% of a sleep session, respectively (Carskadon and Dement 2011). To reflect the degree of sleep depth, NREM is subdivided into stages known as N1 (light), N2 (intermediate), and N3 (deep) (Javaheri and Redline 2012, Patel et al 2021). The N3 stage is also known as slow-wave sleep and is the most
| References        | Signal | Features                                                                 | Classifier                                                                                     | Stages             | Temporal window                          | Kappa  | Accuracy |
|-------------------|--------|--------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|--------------------|------------------------------------------|--------|----------|
| Fonseca et al     | PPG    | 86 features from PPG-based IBIs                                         | Multiclass Bayesian linear discriminant (time-varying priors)                                     | W, N1 + N2, N3, REM | Entire sleep required for feature normalization | 0.42   | 0.59     |
| Geng et al 2020   | ECG    | 14 features; Hilbert-Huang transform, wavelet packet decomposition, singular value decomposition | Random forest                                                                                   | W, REM, NREM (N1 + N2 + N3) | 5 min interval around a 30 s epoch          | 0.44   | 0.73     |
| Radha et al 2019  | ECG    | 132 HRV features                                                         | DNN: perceptron layer, 3 LSTM, 2 perceptron layers (~260 000 parameters)                      | W, REM, N1 + N2, N3 | Entire sleep required for feature normalization | 0.61   | 0.77     |
| Sridhar et al 2020| ECG    | 72 000 IHR values resampled at 2 Hz                                      | DNN with convolutional and dilated convolutional layers (est at >1.5 million parameters)     | W, REM, Light (N1 + N2), Deep (N3) | Entire sleep is classified at once         | 0.66   | 0.77     |
| Yücelbaş et al 2018| ECG   | 39 features: singular value decomposition, variational mode decomposition, Hilbert-Huang transform, morphological analysis | Random forest                                                                                   | W, REM, NREM (N1 + N2 + N3) | 30 s window                              | 0.52   | 0.77     |
| Wei et al 2018    | ECG    | 11 features: time domain, spectral features, nonlinear analysis         | Stacked autoencoder 4-layer DNN                                                                | W, REM, NREM       | Not reported                              | 0.56   | 0.77     |
| Wei et al 2019    | ECG    | 25 features from HRV, respiration rate, amplitude variability           | Masking, normalization, 2 full connections, LSTM, dropout, output^{a}                         | W, REM, light, deep | 5 min window centered around each 30 s epoch | 0.55   | 0.78     |
| Zhao and Sun 2021 | PPG    | 21 features: time domain, frequency domain, nonlinear analysis         | Light gradient boosting machine                                                                | W, REM, (N1 + N2), N3 | 30 s window for time domain feature Not specified for frequency domain | 0.69   | 0.77     |

^{a} LSTM is a recurrent neural network that can learn long temporal dependencies.

^{b} We confirmed the number of trainable parameters in Sridhar et al (1.6 million) with the first author. The estimate provided in the table comes from our independent replication of the Sridhar DNN architecture, based on the information provided in the publication (Section: Code and Software availability).
restorative type of sleep (Javaheri and Redline 2012). On average, the N1, N2, and N3 stages are 5%, 50%, and 20% of a sleep session, respectively (Javaheri and Redline 2012, Patel et al 2021). In general, the proportion of N1 and N2 increases with age while the proportion of REM and N3 decreases (Ohayon et al 2004). However, gender-specific analysis found that women had no change in N3 with age compared with men, who showed a 1.7% decrease in N3 per decade (Dorfman et al 2015). Moreover, women showed a smaller rate of N1 increase, a greater rate of increase in N2, and a greater rate of decline in REM as they age (Dorfman et al 2015).

Substantial and persistent changes in the proportions of NREM and REM sleep during the sleep session can potentially lead to issues with cognition, mood, and normal physical function, which are all affected by disrupted sleep (Maquet 2000, Fransen and Buyse 2008, Yordanova et al 2010, Slater and Steier 2012, Zhou et al 2017). Monitoring sleep and identifying the proportions of NREM and REM sleep can help identify alterations in normal sleep patterns and sleep disorders, leading to treatment and improved health.

In clinical practice, sleep stages are annotated by a sleep technician for specific recording epochs (every 30 s) on the basis of visual examination of polysomnography (PSG) signals (Muzet et al 2016, Chirkos et al 2021). PSG includes an electroencephalogram (EEG), an electrocardiogram (ECG), an electromyogram, an electro-oculogram, oxygen saturation, body position, and measures of breathing effort and airflow (Van de Water et al 2011, Tautan et al 2020). Numerous automatic sleep-staging algorithms have been developed to automate the sleep-staging process (Chirkos et al 2021). Most of these utilize EEG signals alone or in combination with other signals due to higher accuracy of EEG measurements and the fact that brain activity most clearly reflects sleep stages (Carskadon and Dement 2011, Tautan et al 2020).

Indicators of cardiac activity such as heart rate (HR) and heart rate variability (HRV) vary significantly depending on the stage of sleep (Chouchou and Desseilles 2014). HRV analyzes have been used to evaluate cardiovascular function in health and disease and during sleep and wakefulness. REM sleep is characterized by a likely sympathetic predominance with vagal withdrawal, while the opposite trend manifests in NREM sleep (Tobaldini et al 2013). The presence of sleep disorders such as insomnia (Spiegelhalder et al 2011) or REM behavior disorder (Mineri et al 2016) can modify sympathetic/parasympathetic balance. During sleep, brain and cardiac activity interact on multiple levels: (1) at the sleep-cycle level, between cyclic brain oscillations and co-occurring changes in peripheral autonomic activity; and (2) at a shorter time-scale, where phasic EEG events (K-complexes or arousals) are accompanied by cardiac activity fluctuations (de Zambotti et al 2018). Because of this close interaction, many sleep-staging algorithms based on cardiac activity have been developed (Fonseca et al 2017, Wei et al 2018, Yücelbaş et al 2018, Radha et al 2019, Wei et al 2019, Geng et al 2020, Sridhar et al 2020, Zhao and Sun 2021). In these algorithms, typical preprocessing of cardiac signals consists of: (1) detecting IBI; (2) removing outlier values; and (3) interpolating the IBI sequence to obtain a uniformly sampled IBI sequence that is analyzed in the time and/or frequency domains to extract features.

Most of these algorithms extract features within a prespecified temporal window to build a feature vector, which is processed by a classifier to determine a sleep stage assigned to the corresponding temporal window or to a subsegment thereof (table 1) (Fonseca et al 2017, Wei et al 2018, Yücelbaş et al 2018, Radha et al 2019, Wei et al 2019, Geng et al 2020, Sridhar et al 2020, Zhao and Sun 2021). In contrast, the work of Sridhar et al considered the entire sleep session’s time series of 2 Hz-sampled instantaneous heart rate (IHR; composed of the inverse of each IBI value). The feature-extraction-based algorithms result in a number of features (11–132 features) that are orders of magnitude smaller than the IHR signal used in Sridhar et al (72 000 features) (Sridhar et al 2020). The advantage of the approach proposed by Sridhar et al (Sridhar et al 2020) is in the ability of deep neural networks (DNNs) to automatically ‘discover’ the features needed to detect sleep stages, provided that large amounts of data are used in the learning phase. These classification algorithms fall into two categories: (1) DNN-based; and (2) random forest/decision tree-based. An exception to this is the work by Fonseca et al (Fonseca et al 2015, Fonseca et al 2017), which uses a multiclass Bayesian classifier with variable prior probabilities.

DNN, deep neural network; ECG, electrocardiogram; est, estimated; HRV, heart rate variability; IBI, interbeat interval; IHR, instantaneous heart rate; LSTM, long short-term memory; NREM, non-REM sleep; PPG, photoplethysmography; REM, rapid eye movement; W, wake.

Because of the categorical nature of sleep staging, the accuracy of automatic algorithms is commonly reported using Cohen’s Kappa coefficient of agreement, which, in addition to accuracy, also takes into account the agreement that can occur by chance (Cohen 2016), potentially due to an imbalance in the proportion of sleep stages. Performance of existing algorithms has been quantified using both Kappa and accuracy (Fonseca et al 2017, Wei et al 2018, Yücelbaş et al 2018, Radha et al 2019, Wei et al 2019, Geng et al 2020, Sridhar et al 2020, Zhao and Sun 2021). While Kappa and accuracy positively correlate, a wide range of Kappa values can be associated with the same accuracy. For instance, three studies reported accuracy of 0.77 for four stages (Light NREM = N1 + N2, Deep NREM = N3, REM, and wake) while Kappa values were 0.61, 0.66, and 0.69, respectively (Radha et al 2019, Sridhar et al 2020, Zhao and Sun 2021).

This study describes a sleep-staging algorithm that processes sequential 150 s IBI temporal windows, with a shift of 30 s between adjacent windows. IBIw were used because they can be measured using unobtrusive,
noninvasive sensing technology, such as ballistocardiography (BCG) or indirect contact ECG (Lim et al. 2007, Yi et al. 2019). To validate this algorithm and characterize its accuracy, six datasets (five of them are publicly available) were used that include sleep data from healthy, insomnia, REM behavior disorder (RBD), and periodic limb movement disorder (PLMD) subjects. This algorithm provides an option for real-time detection of sleep stages with a small memory footprint (6408 parameters versus 260 000 parameters (Radha et al. 2019) and more than 1.5 million parameters as confirmed in personal communication from Dr Sridhar (Sridhar et al. 2020)) that can be utilized across platforms.

Methods

Datasets

Dataset D1 was publicly available from the PhysioNet Computing in Cardiology Challenge 2018 (Goldberger et al. 2000, Ghassemi et al. 2018). From this dataset, we used a subset of 994 annotated polysomnography recordings from subjects who were monitored at the Massachusetts General Hospital’s Computational Clinical Neurophysiology Laboratory and the Clinical Data Animation Laboratory for the diagnosis of sleep disorders (Ghassemi et al. 2018). This dataset was built for a challenge aiming at automatically detecting respiratory effort-related arousals, and as such, includes subjects with a wide apnea/hypopnea index (AHI) distribution (average 19.4 and standard deviation 14.4). The dataset was partitioned to ensure uniform distribution of AHI (Ghassemi et al. 2018). PSG signals were captured following AASM standards with 1 ECG channel recorded below the right clavicle near the sternum and 1 over the left lateral chest wall. Seven scorers annotated the data, with 1 scorer per PSG (Ghassemi et al. 2018).

Dataset D2 was a publicly available cyclic alternating pattern sleep database from PhysioNet (Goldberger et al. 2000, Terzano et al. 2002), which included 108 polysomnography recordings. For this research, we used data from healthy sleepers (D2-healthy; n = 16), subjects with insomnia (D2-INS; n = 9), subjects with REM behavior disorder (D2-RBD; n = 22), and subjects with periodic leg movement disorder (D2-PLMD; n = 10). PSG signals included an ECG channel; however, the precise location of the electrodes was not specified. Expert neurologists trained at the Sleep Disorders Center of the Ospedale Maggiore, Parma, Italy provided the sleep staging; only 1 scoring per PSG recording was available.

Dataset D3 consisted of PSG data from self-reported healthy sleepers (n = 45) who participated in a sleep laboratory study to assess the accuracy of sleep metrics obtained from a Sleep Number smart bed compared to PSG-based sleep metrics. PSG recordings were conducted using a standard montage, including an ECG channel with 1 electrode placed below the right clavicle and a second electrode placed on the torso on the left side. PSG recordings were scored by 3 independent, registered sleep technicians using AASM guidelines, and sleep stages were chosen by consensus of at least 2 of the 3 technicians.
Ethics statement
This report used 3 sources of data (D1, D2, and D3). D1 and D2 are publicly available and contain only anonymized dataset information. The study that generated Dataset D3 was submitted to the Institutional Review Board of the University of Chicago (IRB00000331, IRB00000735, IRB00002169) and received ‘exempt status.’

ECG signal processing
R-peaks of the ECG signal from each dataset were detected by applying a high-pass filter followed by a normalization procedure and local dynamic thresholding. These R-peaks were used to calculate the sequence of interbeat intervals, IBI, (figure 1), as previously described (Vollmer 2014). IBIs were processed to replace missing values using linear interpolation (Lippman et al 1994) and to remove excessively long and short IBI values (mean ± 5 SD) before final resampling at 2 Hz (Sridhar et al 2020). Using a 2 Hz resampling frequency has been recommended in the analysis of IBI signals for HRV (Kemper et al 2007, Sridhar et al 2020); the appropriateness of this frequency was confirmed by the performance analysis of the DNN for other sampling frequencies as presented in the supplementary material (table S1). The resampled IBI time series is referred to as the ‘IBI signal.’ To remove signal drifts that can interfere with the DNN calculations, we applied a second-order Butterworth high-pass filter (cutoff frequency: 1 × 10^{-3} Hz) to the IBI signal.

CONV, convolution; D, deep; ECG, electrocardiogram; HP, high pass; IBI, interbeat interval; L, light; LSTM, long short-term memory; R/REM, rapid eye movement; W, wake.

To understand the relationship between the IBI signal and sleep stages, we analyzed the mean power spectrum density of the IBI signal per sleep stage and dataset and calculated the average low/high frequency (LF: HF) spectral ratio across datasets and sleep stages (LF: 0.04–0.15 Hz; HF: 0.15–0.40 Hz) (Shaffer and Ginsberg 2017). LF is believed to reflect sympathetic activity associated with vasomotor waves, while HF may reflect parasympathetic (cardiac vagal) activity (Elsenbruch et al 1999, Shaffer and Ginsberg 2017). The LF:HF ratio has been proposed as an indicator of sympathetic modulation, assuming that the vagal component to LF would be at least partially normalized by the vagal component in the HF (TFESCNASPE 1996).

DNN architecture and evaluation
Input signal
Previous work on HRV estimation and cardiac signal-based sleep staging suggests that the IBI signal should be analyzed in a time period that lasts for a few minutes to extract meaningful information (Shaffer and Ginsberg 2017, Wei et al 2019, Geng et al 2020). In our implementation, the IBI signal [IBI(t)] was processed in sequences of 300 sample-long temporal windows (i.e., 150 s) with a shift of 60 samples (i.e., 30 s) and overlap of 240 samples between consecutive windows. Each 300 sample window consisted of a central 60-sample portion in which the sleep stage is assigned, preceded, and followed by 120 samples (figure 1).

DNN architecture
The DNN architecture included a group of 3 one-dimensional convolution layers (CONV) followed by a group of 3 long short-term memory (LSTM) layers (figure 1). This architecture was inspired by the DNN hyperparameter optimization previously developed for EEG-based sleep staging (Bresch et al 2018) and the real-time use of this DNN in previous studies (Garcia-Molina et al 2019). The output of the CONV group was flattened to generate a 36-component vector that was used as input to the recurrent LSTM layers, each with 16 units. This architecture resulted in 6408 trainable parameters. The output of the LSTM group was a set of unnormalized log probabilities ([z_{D}, z_{L}, z_{R}, z_{W}]) corresponding to 4 stages: deep sleep (D; N3), light sleep (L; N1 + N2), REM (R) sleep, and wake (W). The probability for each stage was approximated using the soft-max function in equation (1) (Goodfellow et al 2016).

\[
q_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)},
\]

where \(i, j \in \{D, L, R, W\}\) and \(z_i\) is the unnormalized log probability associated with stage \(i\).

The assignment of a stage to the window was based on the highest probability rule, with the assigned stage corresponding to the highest probability \(\text{argmax}_i(q_i)\). The compilation of all sleep-stage decisions constitutes the estimated hypnogram.

DNN training and validation
A portion of the D1 dataset was used for training and validation (700 of 994 recordings). A three-fold cross-validation strategy was used to train the DNN. The testing and validation set for each cross validation included 480 and 220 recordings, respectively.
The number of DNN-epochs for training and validation was set to 1000 and the batch size to 128 sequences. The sequences were randomly selected from the training set. Each sequence was 128 consecutive windows long, spanning a duration of 66 min ($= 60 \times 127 + 300$ samples). A DNN-epoch corresponds to a pass of the entire training data through the DNN. Categorical cross-entropy (equation (2)) was set as loss function using the Adam optimization algorithm (Kingma and Ba 2015) with the following parameters: learning rate $= 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and decay $= 0$. The trainable parameters were initialized according to the Glorot normal initialization strategy (Glorot and Bengio 2010)

$$J = \sum_i p_i \log(q_i), \ i \in \{D, L, R, W\}$$

where $p_i$ is the probability of stage $i$, and $q_i$ is the softmax function (equation (1)).

For each DNN-epoch and cross validation, we calculated the loss and average accuracy (across all stages) for both training and validation sets. DNN models were saved every 100 DNN-epochs. The saved model for which the validation loss was at the minimum was used in the testing phase. The DNN was implemented in Python 3.6 and Keras 2.2 (Chollet 2015) and was deployed in an m2.xlarge Amazon Web Services instance.

**DNN testing**

The testing was performed on 6 sets of data, including the remaining data from D1 (294 of 994 recordings remaining after selecting the training and validation datasets), dataset D3, and the four D2 datasets.

The agreement between the ground-truth hypnogram and the estimated hypnogram was calculated using Cohen’s Kappa coefficient of agreement as follows (equation (3)): given a confusion matrix $C$ with elements $C_{ij}$; $i,j \in \{D, L, R, W\}$ that represent the number of sleep-epochs belonging to true stage $j$ detected as stage $i$:

$$Kappa = \frac{4\sum_i C_{ii} - \sum_i R_i L_i}{16 - \sum_i R_i L_i},$$

where $i \in \{D, L, R, W\}$, $R_i$ is the sum of elements along the $i$th row of the confusion matrix, and $L_i$ is the sum of elements along the $i$th column of the confusion matrix.

Kappa was interpreted as indicating no agreement if $Kappa < 0.20$, fair agreement if $0.20 \leq Kappa < 0.40$, moderate agreement if $0.40 \leq Kappa < 0.60$, substantial agreement if $0.60 \leq Kappa < 0.80$, and perfect agreement if $Kappa \geq 0.80$ (Landis and Koch 1977). The accuracy was calculated as the proportion of elements along the diagonal

---

Figure 2. Adjustment of sleep-stage decision based on probability-ratio threshold.
of the confusion matrix. In addition to Kappa and accuracy, sensitivity and specificity for each stage were calculated by considering $2 \times 2$ stage-specific confusion matrices with 2 categories: belonging, and not belonging to the target stage.

**Kappa optimization and correction of light sleep bias**

Given that light sleep is significantly more prevalent than other sleep stages, the DNN can be biased to detect light sleep. To correct the bias, we analyzed the cumulative distribution functions of the probability ratio of light sleep versus other stages ($\frac{q(L)}{q(R)}$, and $\frac{q(L)}{q(W)}$). This analysis allowed us to increase Kappa by performing a confirmation step when $q(L)$ is the highest probability produced by the DNN. The confirmation step depends on the probability ratios (figure 2).

CONV, convolution; D, deep; HP, high pass; IBI, interbeat interval; L, light; R, rapid eye movement; W, wake.

The DNN output was first sorted in order of decreasing probabilities: $q(i_1) > q(i_2) > q(i_3) > q(i_4)$; $i_1, i_2, i_3, i_4 \in \{D, L, R, W\}$. If $i_1 = L$ then the assigned sleep stage was $i_1$; otherwise the ratio $\frac{q(i_1)}{q(i_2)}$ was compared to a probability ratio threshold, $\tau_i > 1$. If $\frac{q(i_1)}{q(i_2)}$ exceeded the threshold, then the assigned stage was light sleep. If $\frac{q(i_1)}{q(i_2)}$ did not exceed the threshold, the assigned sleep stage was $i_2$.

The optimal values of the thresholds $\tau_D$, $\tau_R$, and $\tau_W$ were determined by testing all possible combinations of the triplets $\tau_D$, $\tau_R$, and $\tau_W$ for $\tau_{D,R,W} = 1 : 0.05 : 2.2$ (the notation $1 : 0.05 : 2.2$ indicates all values between 1 and 2.2 with a step of 0.05; i.e., 1, 1.05, 1.10, 1.15...) using the validation dataset to determine the optimal Kappa.

**DNN spectral response**

The DNN architecture (CONV + LSTM) can be considered a feature extraction block in the frequency domain followed by a memory block (Bresch et al. 2018). We analyzed the frequency domain response of the CONV group by using synthetic IBI signals as input to the DNN, as described in equation (4):

$$s(t) = 10^{-3} \sqrt{2} \times SDNN \times \cos(2\pi ft),$$  \hspace{1cm} (4)

where the standard deviation of normal-to-normal intervals (SDNN) (Shaffer and Gingsberg 2017) = $\int s^2(t) dt - \left(\int s(t) dt \right)^2$, which is also the standard deviation of $s(t)$ in milliseconds, and $0 < f < 0.5$ Hz is the frequency of the signal. $s(t)$ was sampled at 2 Hz to match the DNN input signal. The response of the 36 features of the CONV group were visualized against $f$. We also analyzed the response of the CONV layer depending on the spectral LF:HF ratio by calculating the LF:HF of IBI signals in the training set and evaluating the corresponding CONV response.
**Table 2.** Subject demographics and sleep-stage distribution.

|                  | D1-all\(^a\) | D2-healthy | D2-INS | D2-RBD | D2-PLMD | D3  |
|------------------|---------------|------------|--------|--------|---------|-----|
| Subjects, \(n\) | 994           | 16         | 9      | 22     | 10      | 45  |
| Male, \(n\) (%) | 666 (67.0)    | 7 (43.8)   | 4 (44.4)| 19 (86.4)| 7 (70.0)| 20 (44.4) |
| Female, \(n\) (%)| 328 (33.0)    | 9 (56.2)   | 5 (55.6)| 3 (13.6)| 3 (30.0)| 25 (55.6) |
| Age, years, mean (SD) | 55.2 (14.3) | 32.2 (5.5)\(^b\) | 60.9 (10.5) | 70.7 (6.39)\(^b\) | 55.1 (6.71) | 41.2 (10.5)\(^b\) |

**Sleep-stage distribution, mean % (SD)**

|       | Light        | Deep       | REM        | Wake       |
|-------|--------------|------------|------------|------------|
|       | \(56.4\) (13.4) | \(11.2\) (8.26) | \(12.3\) (7.39) | \(20.0\) (13.8) |
| \(^b\) | \(43.3\) (8.06)\(^c\) | \(8.83\) (3.27) | \(38.9\) (10.9) | \(8.94\) (6.00) |
| \(^c\) | \(36.0\) (11.0) | \(9.06\) (5.66) | \(18.4\) (8.89) | \(36.5\) (18.5) |
| \(^c\) | \(37.5\) (10.2) | \(12.3\) (6.29) | \(26.7\) (10.7) | \(23.6\) (11.7) |
| \(^c\) | \(38.8\) (12.2) | \(12.6\) (4.98) | \(28.2\) (11.2) | \(20.9\) (14.1) |
| \(^d\) | \(51.9\) (8.91) | \(13.2\) (6.33) | \(15.1\) (5.54) | \(19.8\) (9.93) |
| \(^d\) | \(P < 1.0 \times 10^{-8}\) versus D1-all. |
| \(^c\) | \(P < 1.0 \times 10^{-6}\) versus D1-all. |
| \(^a\) | D1-all refers to the entire D1 dataset, including training, validation, and testing subsets. |

**Figure 4.** DNN learning curves.

---

**Statistical analysis**

Dataset D1 was the reference for statistically testing differences in age, sleep-stage distribution, and cardiac metrics among datasets. Statistical significance (defined as \(P < 0.05\)) was determined based on generalized linear models where the datasets were predictors and the factor to test was the dependent variable.

Because HRV changes with gender and age (Liu et al. 2014), we tested the effect of these demographic factors on Kappa, using the D3 dataset and a linear mixed model approach. Statistical significance was defined as \(P < 0.05\).
Results

Sleep-stage distribution and spectral IBI signal analysis
The demographic information and sleep stage distribution for each dataset are presented in table 2. Subjects from D3 and D2-healthy were significantly younger than those in D1 ($P < 1.0 \times 10^{-6}$), while subjects in D2-RBD were significantly older than those in D1 ($P < 1.0 \times 10^{-6}$). The relative percent distributions of sleep stages of all D2 subsets were significantly different from those of D1 ($P < 0.001$); the sleep-stage distributions for D3 were not significantly different from D1.

INS, insomnia; PLMD, periodic limb movement disorder; RBD, REM behavior disorder; REM, rapid eye movement; SD, standard deviation.

Analysis of the mean power spectrum density (PSD) of the IBI signal across sleep stages showed that spectral content in the LF band (0.04–0.15 Hz) was highest for the wake stage across all datasets, while the LF spectral content for deep sleep was generally lower than that in wake (figure 3). Trends in LF spectral content depended on the condition investigated (figure 3).

The spectral content of the LF band was highest for wake for most datasets.

IBI, interbeat interval; INS, insomnia; PLMD, periodic limb movement disorder; RBD, REM behavior disorder; REM, rapid eye movement.

The LF:HF ratio was used as a frequency domain metric, similar to HRV, to assess IBI spectral differences. Some LF:HF values for D2 datasets were found to be significantly lower (table S2), including for light sleep ($P = 0.04$), REM ($P = 0.03$), and wake ($P = 0.03$) in the D2-healthy set; and REM and wake for both D2-INS ($P = 0.03$ for both) and D2-RBD ($P = 0.03$ and $P = 0.04$, respectively) sets; there were no significant changes for D2-PLMD or D3 sets compared to the D1 set.

As PSD curve analysis suggested a decline of LF power from the wake stage to the deep-sleep stage, we calculated the difference of [LF:HF in wake] minus [LF:HF in deep sleep] for all datasets. These differences varied significantly across subjects with sleep disorders (D2-INS: $P = 0.04$; D3-RBD: $P = 0.04$; and D2-PLMD: $P = 0.03$).

DNN training, validation, and testing results
The training learning curves for the three cross validations can be observed in figure 4(A). The demographic information and sleep architecture for each cross validation is in table 3. While some differences in the rate of loss reduction can be observed up to DNN-epoch 200, the trends are almost indistinguishable among cross validations for DNN-epoch > 300. Given the similarity in the training learning curves and the dataset properties for all cross validation folds (table 3), we focused on the analysis of the validation accuracy and loss for cross validation 1 (figure 4(B)).

Panel A: Learning curves for the three cross validation strategies. Panel B: top graph shows training and validation accuracy for cross validation 1. Bottom graph shows the loss function. DNN models were saved every 100 epochs. Both the training and validation datasets showed an increasing trend in accuracy.
Training accuracy increased with the DNN-epoch, while the validation accuracy fluctuated more than the training accuracy (figure 4(B)). Training and validation showed an increasing trend in accuracy overall. We compared this DNN architecture against a smaller one with a pure CONV-based architecture that takes the time dimension into account by processing three windows (previous, current, and future, table S3) and found that the accuracy of the pure-CONV-based architecture was lower than that of one using CONV and LSTM.

The loss curves in figure 4(B) suggested an adequate DNN fit based on the narrow gap between the training and validation losses. However, a slight increase in validation loss was noticeable after 500 DNN-epochs. A second-degree polynomial was fit to the validation loss as follows:

\[
\text{Validation loss} = 2.5 \times 10^{-7} \times e^2 - 3.7 \times 10^{-4} \times e + 1.1,
\]

where \(e\) is the DNN-epoch, to determine the number of DNN-epochs for which the loss was at a minimum: \(\text{argmin}_e (\text{validation loss}) = 740\). The DNN model saved at 700 DNN-epochs was used to calculate performance values in the testing datasets.

The DNN performance values, including the average Kappa and accuracy, and sensitivity and specificity for each sleep stage are reported in table 4. Results of the dataset D1-test were used as reference for statistical comparisons. The Kappa and accuracy for the D1-test were not significantly different from those in the training and validation sets, which is consistent with the absence of overfitting or underfitting issues that we observed in figure 4.

All data presented as mean (SD).

INS, insomnia; PLMD, periodic limb movement disorder; RBD, REM behavior disorder; REM, rapid eye movement; Sens, sensitivity; SD, standard deviation; Spec, specificity.

For all datasets, the DNN failed to detect deep sleep; the sensitivity and specificity are 0 and 1, respectively (table 4). Furthermore, the high sensitivity and moderate specificity for light-sleep detection across all datasets, coupled with the relatively low sensitivity for REM and wake, suggest that the DNN was biased toward detecting light sleep, which is the most prevalent sleep stage (as observed in table 2).

The accuracy and/or Kappa were significantly lower for the D2 datasets (table 4). This lower performance in DNN was reflected in detection deficiencies for light sleep and REM sleep in the D2 datasets, as well as for wake. For D3, the DNN performed similarly to the D1-test with respect to Kappa and accuracy.

**Correction of the light-sleep bias using probability-ratio thresholds**

The light-sleep bias was corrected as described in the Methods section. The bias could be observed by analyzing pair-wise cumulative distribution functions (CDFs) between each sleep stage \(S \in (D, R, W)\) and light sleep (figure S1 (available online at stacks.iop.org/PMEA/43/025004/mmedia)).

For each stage other than light sleep, the CDF of \(S\) is portrayed versus the ratio between the probability of light sleep and the probability of sleep stage \(S\): \(q_{L} \over q_{S}\). In addition to the CDF for stage \(S\), the CDF for light sleep versus \(q_{L} \over q_{S}\) is also shown. The faster increase in the CDF of sleep stage \(S = L\) compared with that of light sleep as \(q_{L} \over q_{S}\) suggests that the bias for light sleep could be corrected by setting a threshold on \(q_{L} \over q_{S}\) (figure 2).

To determine the optimal values for thresholds \(\tau_{D}, \tau_{R}, \text{and } \tau_{W}\), a search strategy was adopted in which all possible combinations of the triplets \((\tau_{D}, \tau_{R}, \tau_{W})\) for \(\tau_{D, R, W} = 1, 1.05, 1.1, \ldots, 2.2\), were applied to the validation

### Table 4. Sleep-staging accuracy for all testing datasets.

| D1-test | D2-healthy | D2-INS | D2-RBD | D2-PLMD | D3 |
|---------|------------|--------|--------|---------|----|
| Kappa   | 0.32 (0.15) | 0.24 (0.08)* | 0.30 (0.13) | 0.18 (0.06)* | 0.23 (0.11)* | 0.33 (0.11) |
| Accuracy | 0.64 (0.12) | 0.56 (0.07)* | 0.54 (0.13)* | 0.48 (0.07)* | 0.51 (0.12)* | 0.63 (0.09) |
| Stage   | Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec Sens Spec explo.
The optimal triplet was determined. An illustration of the results is shown in Figure S2, with the mean Kappa against each threshold $\tau$. The optimal triplet was $\tau_D = 1.45, \tau_R = 1.50$, and $\tau_W = 1.25$. Using this triplet, mean Kappa, accuracy, and sensitivity/specificity for each sleep stage were calculated for the testing datasets (Table 5). The mean Kappa increased for all datasets compared with Kappa reported in Table 4 (+0.07 SD; 0.11), with significant increases for the D1-test (+0.06; $P < 10^{-5}$) and D3 (+0.1; $P < 10^{-5}$). Accuracy did not significantly change.

INS, insomnia; PLMD, periodic limb movement disorder; RBD, REM behavior disorder; REM, rapid eye movement; Sens, sensitivity; SD, standard deviation; Spec, specificity.

The mean normalized confusion matrices for all datasets show a diagonal dominance consistent with the results in Table 5 for the D1-test and D3 (Figure 5). Lower Kappa values were associated with D2 datasets. The absence of diagonal dominance is particularly visible in D2-RBD, where REM sleep sensitivity was low, presumably due to the nature of the RBD condition. Bland-Altman analyzes for the detected duration of each sleep stage and dataset are reported in the supplementary material (Figure S3, Table S4).
The confusion matrices show a diagonal dominance, with lower Kappa values associated with the D2 datasets. The fraction of true stages is reported along the columns of the confusion matrix, and the DNN-detected stages are reported on the rows. The normalization process is such that the columns add up to 1; thus, the element $C_{i,j}; i, j \in \{D, L, R, W\}$ indicates the fraction of actual stage '$j$' detected as stage '$i$'.

Acc, accuracy; INS, insomnia; PLMD, periodic limb movement disorder; RBD, REM behavior disorder; REM, rapid eye movement sleep.

### Three-stage case

In all datasets, a high degree of confusion was observed, associated with deep sleep being detected as light sleep. For D2-RBD and D2-PLMD, 68% and 74% of actual deep-sleep epochs were detected as light sleep, respectively. This observation motivated an additional analysis that considers 3 stages instead of 4: NREM sleep (which combines light and deep sleep), REM sleep, and wake. The average Kappa, accuracy, and sensitivity specificity for the 3-stage case are reported in Table 6; REM and wake sensitivity and specificity were identical to the 4-stage case, as expected (Table 5).

Kappa values for the 3-stage case were higher for all datasets (Table 6). The highest increase in Kappa was +0.08 for D3, followed by +0.07 for the D1-test, and +0.06 for D2-RBD and D2-PLMD. A smaller improvement was found for D2-healthy (+0.04) and D2-INS (+0.02). Kappa and accuracy and NREM

### Table 6. Three-stage sleep analysis.

| Mean (SD)   | D1-test | D2-healthy | D2-INS | D2-RBD | D2-PLMD | D3      |
|-------------|---------|------------|--------|--------|---------|---------|
| Kappa       | 0.46(0.17) | 0.35 (0.10)$^b$ | 0.41 (0.14) | 0.30 (0.13)$^c$ | 0.42 (0.11) | 0.52 (0.12)$^c$ |
| Accuracy    | 0.73 (0.11) | 0.65 (0.09)$^c$ | 0.64 (0.12)$^c$ | 0.59 (0.10)$^c$ | 0.68 (0.08) | 0.76 (0.07)$^c$ |
| Stage       | Sens    | Spec      | Sens    | Spec    | Sens    | Spec    |
| NREM        | 0.77 (0.15) | 0.73 (0.17) | 0.48 (0.09) | 0.72 (0.07)$^a$ | 0.79 (0.18) | 0.54 (0.14)$^c$ |
|             | 0.73 (0.17) | 0.48 (0.09) | 0.72 (0.07)$^a$ | 0.79 (0.18) | 0.54 (0.14)$^c$ | 0.79 (0.18) | 0.54 (0.14)$^c$ |
| REM         | 0.58 (0.30) | 0.41 (0.09) | 0.89 (0.09) | 0.50 (0.12) | 0.85 (0.14)$^c$ | 0.23 (0.14)$^c$ |
|             | 0.90 (0.08) | 0.41 (0.09) | 0.89 (0.09) | 0.50 (0.12) | 0.85 (0.14)$^c$ | 0.23 (0.14)$^c$ |
| Wake        | 0.65 (0.21) | 0.88 (0.12) | 0.63 (0.26) | 0.57 (0.01)$^b$ | 0.86 (0.15) | 0.54 (0.26)$^b$ |
|             | 0.65 (0.21) | 0.88 (0.12) | 0.63 (0.26) | 0.57 (0.01)$^b$ | 0.86 (0.15) | 0.54 (0.26)$^b$ |

$^a$ $P < 0.05.$  
$^b$ $P < 0.01.$  
$^c$ $P < 0.001.$ INS, insomnia; PLMD, periodic limb movement disorder; RBD, REM behavior disorder; REM, rapid eye movement; Sens, sensitivity; SD, standard deviation; Spec, specificity.

Figure 6. Convolution layer output depending on frequency.
sensitivity were significantly higher for D3 compared with the D1-test ($P < 0.05$); however, similar to the 4-stage case, marked performance deficiencies were associated with D2 datasets (table 6).

The difference in performance metrics may have been due to differences in spectral IBI content (figure 3, table S2). Therefore, we analyzed the DNN spectral response using the synthetic signals generated by equation (4).

**DNN spectral response**

Synthetic signals were generated for 4 values of SDNN ∈ {100, 150, 200, 250 milliseconds} and 50 frequency values in the interval from 0.01 to 0.5 Hz, with a step of 0.01 Hz. We plotted the values of the 36 features generated as output of the CONV group versus the frequency of the synthetic signal and for the 4 SDNN values (figure 6). Based on the profile of the spectral response, 4 clusters could be distinguished, each with 9 components. The CONV outputs in cluster 1 in general had higher values compared with the rest of the clusters in all but the lower frequencies ($<0.05$ Hz). The shape of the CONV output curves for cluster 2 and cluster 4 members resembled that of a band-pass filter in the frequency range from 0.1 to 0.4 Hz, although the values for cluster 4 were lower. The shape of cluster 3 curves resembled that of a low-pass filter with an approximate cutoff frequency of 0.1 Hz.

The amplitude of the CONV output in each cluster increased as the amplitude of the synthetic signal (i.e., SDNN) increased. As the amplitude increased, the CONV output values for cluster 2 approached that of cluster 1 in the vicinity of 0.25 Hz. The peak at 0.25 Hz that is most prominently visible for clusters 2 and 4 may reflect typical respiration frequency (15 breaths min$^{-1}$); this frequency is related to the phasic effects of tidal respiration on the cardiovascular system (Russo et al 2017). The analysis of the power spectrum density of signals in dataset D3 suggests the validity of this hypothesis (figure S4).

Each line represents a CONV output, with 36 outputs in total. Similar outputs are clustered, with clusters corresponding to CONV outputs as follows: Cluster 1 = outputs 1–9; Cluster 2 = outputs 10–18; Cluster 3 = outputs 19–27; and Cluster 4 = outputs 28–36. Clusters 2 and 4 show peaks at 0.25 Hz, a frequency that corresponds to a respiration rate of 15 breaths per minute.

CONV, convolution layer; SDNN, standard deviation of normal-to-normal intervals.

The analysis of the CONV responses (figure 7) depending on the spectral LF:HF ratio shows 3 qualitative clusters: (1) inverse U-shaped response depending on LF:HF; (2) quasi-constant response across a wide range of LF:HF values; and (3) high response for low LF:HF and decreasing as LF:HF increases.

This figure shows the responses of the 36 CONV outputs depending on LF:HF. The responses form three clusters corresponding to a U-shaped response, a quasi-constant response, or a response that is inversely related to LF:HF.

CONV, convolution; HF, high frequency; LF, low frequency.

**Estimation of the DNN processing speed**

We tested the processing speed of the network by embedding the DNN in an NXP® imX nano CPU-ARM processor. The test consisted of processing a thousand 300 sample-long windows. The average time to process a 300 sample window was 19.2 milliseconds.

**Effect of gender and age on Kappa**

Using the D3 dataset, we developed a linear mixed model (equation (5)) to predict Kappa using gender and age as predictors.
Kappa $\sim 0.52875 - 0.006 \times \text{male gender} - 0.0016 \times \text{age}$.

The statistical significance for the coefficients of gender and age was 0.235 and 0.244, respectively. Thus, neither gender nor age significantly affected Kappa.

**Discussion**

Interbeat intervals are among the basic types of information that can be extracted from cardiac monitoring technologies such as electrocardiography, photoplethysmography (including devices attached to the wrist and the finger), ballistocardiography, and video-based cardiography. Using the IBI signal for automatic sleep staging enables the development of algorithms that are agnostic of the cardiac monitoring technology. Moreover, the IBI’s temporal unit facilitates algorithm portability across technologies, because contrary to other PSG signals, scaling or filtering of electrical amplitude signals are not necessary. Different types of features (including time domain, frequency domain, or nonlinear features) can be extracted from IBI signals (table 1) and used as input for machine-learning classifiers.

We developed a DNN that enables detection of sleep stages based on IBI signals. Our DNN holds promise for real world implementation due to its small size and real-time measurement capability. We tested and validated our DNN across a large set of studies, incorporating both healthy sleepers and those with sleep disorders. Furthermore, we verified that gender and age do not affect the accuracy of the DNN.

We found that the spectral content of the IBI signals was highest for wake across all datasets and lowest for deep sleep. This may be due to changes in autonomic activity that occur from wakefulness to deep sleep (Chouchou and Desseilles 2014). The LF:HF ratio was used to assess IBI spectral differences across datasets—the D2 dataset had lower ratios, while D3 was not different from D1. The decline in LF power from wake to deep sleep was significantly different for subjects with sleep disorders (insomnia, RBD, or PLMD). Reduced HRV has been reported for insomnia (Spiegelhalder et al 2011), RBD (Ministeri et al 2016), and PLMD (Barone et al 2017), which may contribute, in a condition-specific manner, to the observed differences in LF power decline.

As an alternative to feature extraction, in this study we leveraged automatic feature discovery enabled by use of a DNN composed of a CONV group, which likely extracted spectral features (as suggested by the DNN spectral-response and LF:HF response analyzes), followed by a recurrent LSTM group that accounted for sequential time dependencies. The chosen DNN architecture leveraged the hyperparameter optimization proposed in Bresch et al (2018). The hyperparameters proposed by Bresch et al (2018) in the context of EEG signals were found to be appropriate for the IBI signals in this research because the time-varying spectral content of both types of signals (EEG and IBI) appears to specifically reflect sleep stages and their transitions.

The 1D CONV operated along a single dimension of the input as a ‘feature extractor’ and had lower complexity, making it suitable for real-time applications (Kiranyaz et al 2021). An LSTM layer is a type of recurrent neural network composed of memory layers that have the ability to identify and utilize longer-term dependencies in time series (Hochreiter and Schmidhuber 1997), which are important to accurately identify sleep stages during transition epochs. The outputs of the LSTM approximated the probability of each sleep stage by using the softmax output. DNN spectral response analysis showed that the CONV outputs can be categorized into 4 clusters, each with 9 components, with distinct spectral properties and where amplitude increased with increasing HRV (as quantified by SDNN values). Interestingly, a peak in the CONV spectral response at 0.25 Hz appears to be related to a peak in the respiratory effort signal at 15 breaths min$^{-1}$ that is prominent in the mean power spectrum density of the respiratory effort sigma. We also found that the CONV output depended on the LF:HF ratio—this suggests that the CONV part of the DNN extracts spectral features in frequency bands that are relevant for IBI analysis and are further processed by the LSTM group to produce a decision.

We have tested a much smaller pure CONV-based alternative DNN architecture (Table S3) that processes three (previous, current and future) windows simultaneously but leads to lower performance. Another option for optimization is provided in Shen et al (2021); we implemented this network and adapted the dimensions to our data. This produced $>500,000$ trainable parameters, roughly two orders of magnitude larger than the DNN we developed and exceeding the capacity of the microprocessor for real-time applications. Another future optimization consists in testing an architecture that results in only 4 CONV outputs (i.e., the number of clusters identified in the CONV spectral response) followed by LSTM layers.

Real-time interventions that compensate sleep deficiencies (e.g. deep-sleep prolongation in elderly individuals using a temperature intervention (Raymann et al 2008)) or boost beneficial aspects of sleep (e.g., slow-wave-sleep enhancement through sensory or electromagnetic stimulation (Bellesi et al 2014)) require real-time sleep staging, preferably based on minimally obtrusive sensing. We believe that the algorithm we propose in this research is minimally obtrusive because it relies on IBI signals that can be extracted through contactless sensing to estimate the sleep stage of 30 s sleep epochs, providing sleep staging of a specific epoch within 1 min. Additionally, our DNN model is small in scale compared with other DNN models—our model utilizes 6408
parameters compared with > 100 K parameters for other models relying on DNNs such as those reported by Radha et al. (Radha et al 2019) and Sridhar et al (Sridhar et al 2020)—and has the processing speed to provide real-time results. The latency of our algorithm (1 min) can be considered real-time if we refer to the average bout (uninterrupted sleep-stage period) durations for deep, light, and REM stages were estimated at 11.95, 6.78, and 9.40 min, respectively, in the D3 dataset (healthy subjects). These durations are consistent with a previous study (Usher et al 2012), suggesting that a 1 min latency provides a reasonable opportunity to intervene, as has been demonstrated in other studies (Papalambros et al 2017). Furthermore, a similar DNN architecture to the one proposed in this research was successfully embedded in a consumer device (Garcia-Molina et al 2019, Diep et al 2020). This small size makes our DNN model ideal for embedded platforms.

The performance of the staging algorithm presented in this study as measured by Kappa and accuracy is lower compared with the state-of-the-art offline approaches (table 1). However, in real-time sleep interventions such as deep sleep or REM sleep enhancement where the intervention (e.g., sensory stimulation) occurs during deep and REM sleep stages, it is important to guarantee high specificity with a reasonable sensitivity to prevent untimely stimulation that may otherwise disturb sleep (Papalambros et al 2017, Garcia-Molina et al 2018). For instance, a previously described device enhanced deep sleep through auditory stimulation that was delivered during approximately 50% of the duration of deep sleep (Garcia-Molina et al 2018). The algorithm we propose has average specificity of detecting both deep and REM sleep (≥ 0.90) in the D3 healthy group, similar to previous studies (Debellemansiere et al 2018), while average sensitivity was 0.46 and 0.67 for deep and REM sleep, respectively. The applicability of our sleep staging algorithm for the sleep disorders we considered (RBD and PLM) is much more limited; future work should focus on designing dedicated DNNs for an intervention tailored for a specific sleep disorder.

Understanding the factors that determine the accuracy of a sleep-staging algorithm and its generalization properties is essential to deploying an algorithm. These considerations motivated our approach to validate our DNN model performance using several datasets, including data from groups with sleep disorders. The DNN performance without postprocessing of the LSTM output on the testing datasets revealed a bias for the detection of light sleep (table 4), which results from the minimization of the loss function (equation (2)) and high prevalence of light sleep. The analysis of the empirical distribution of the probability ratios $q_{\ell | q_S}$, $S \in \{D, R, W\}$ enabled us to correct the bias by setting a threshold $\tau_S > 1$ on $q_{\ell | q_S}$ before assigning stage $L$ when $\text{argmax} (q_i) = L$ and $q_i$ is the second highest probability. This strategy resulted in an increase in Kappa across all datasets and a significant increase in Kappa for the D1-test and D3, while accuracy increased only slightly. Although the minimization of the loss function based on the cross-entropy is accompanied by an increase in accuracy (figure 6), Kappa may not necessarily display a similar increasing trend. This is similar to previous studies (table 1) where the same accuracy value is associated with a range of Kappa values, and suggests that the strategy consisting in setting thresholds on probability ratios increases Kappa, preserves accuracy, and crucially enables the detection of deep sleep in all datasets.

Using a weighted cross-entropy loss function provides an alternative to correct bias due to class imbalance (Cui et al 2019). This can be tested in a straightforward manner using the Python software utilized in this research (see Supplement for more details). The results on the D3 dataset were Kappa ± SD = 0.38 ± 0.11 and accuracy ± SD = 0.64 ± 0.09, which were higher compared with those in table 4 but lower compared with those in table 5, suggesting that our probability-ratio thresholding approach results in the highest Kappa and accuracy for these data. Class imbalance is a pervasive issue in real world applications using machine learning (Kumar et al 2021). Random oversampling/undersampling and class-imbalance-corrected loss functions (such as the weighted cross-entropy) are commonly used to tackle imbalance. Further research on optimizing thresholds on probability-ratios may reveal useful to address class imbalance.

The performance across datasets was significantly different with a consistently higher Kappa and accuracy for D3 and lower performance for the D2 datasets (except the insomnia cases). The results with D3 are encouraging for the generalizability of the DNN because D3 was recorded at a different sleep laboratory, under different circumstances and in a different population compared with D1. The reduced performance in the D2 datasets may be due to significant differences in sleep architecture and the spectral content of the IBI signal (table 2 and table S2). D2 performance appears to challenge the generalization of the DNN but also highlights the importance of considering heterogeneity in the training dataset.

The obvious question that emerges from this consideration is whether different probability ratios can be optimized on the D2 datasets? We conducted a proof-of-concept experiment to explore this possibility through exhaustive searches of $\tau_D$, $\tau_R$, and $\tau_W$ using the D2-healthy and D2-RBD datasets. For the 4-stage case, we found that the average Kappa of the D2-healthy set can increase to 0.38 (+0.07, with respect to the result in table 5) if $\tau_D = 1.05$, $\tau_R = 1.45$, and $\tau_W = 1.65$. Similarly, the Kappa value of the D2-RBD set can be increased to 0.28 (+0.04, with respect to the result in table 5) if $\tau_D = 1.05$, $\tau_R = 1.65$, and $\tau_W = 1.95$. This possibility of adapting the DNN to process new IBI data with different sleep-stage distribution and spectral properties by modifying the
thresholds on the probability ratios offers an option for computationally efficient transfer learning, which would otherwise consist in retraining portions of the DNN with the new IBI data.

From the perspective of an intervention such as slow wave sleep enhancement, the distinction between light sleep and deep sleep may not be strictly required. Both light sleep and deep sleep are categories of NREM sleep and, as such, share the same physiological generators. Therefore, we considered a 3-stage scenario (NREM = light sleep + deep sleep, REM, and wake); as expected, performance increased for all datasets and, for the D3 set, Kappa reaches 0.52, significantly better compared with all other datasets ($P < 0.05$).

Given the high prevalence of undiagnosed sleep disorders in the general population, it is important that automatic sleep-staging algorithms also include data from sleep disorders during the training phase. The proportion of data from each disorder may be guided by disorder prevalence and demographic factors. The data available from sleep disorders in this study was not sufficient to include in a proportion consistent with the prevalence in the general population (Ohayon and Roth 2001, 2002; Haba-Rubio et al 2018); these considerations should be accounted for in future research.

One limitation of this study is that the DNN optimization relied on minimization of the cross-entropy loss that favors accuracy without considering Kappa. Recent developments in deep learning have recommended the optimization of Kappa (de la Torre et al 2018), which is presumably more appropriate for sleep staging. Additionally, the approach presented here is not an end-to-end DNN solution: we used a high-pass filter of the IBI signal before DNN processing and used thresholds on the probability ratios to correct the bias for light sleep. Neural networks are considered universal approximation functions, and it can therefore be argued that end-to-end neural architectures are possible and may even be more suitable because they can automatically discover parameters such as filter coefficients and thresholds. However, this is critically dependent on the choice of loss function and network architecture. Another limitation of this work is the absence of multiple expert-scoring per PSG across all datasets. Finally, the IBI signal-based sleep-staging algorithm proposed here does not account for the presence of sleep/cardiac conditions and use of medications that may modify HR and/or HRV. Indeed, the algorithm performance for the disorder datasets is lower compared with that of the D3 healthy dataset.

Our algorithm may provide sleep staging in real-time compared with existing algorithms. The algorithms described in 3 reports require the data from the entire sleep duration to either postprocess (including detrending, normalizing, and special handling of the data from epochs at the beginning or end of the sleep recording) the features (Fonseca et al 2017, Radha et al 2019, Sridhar et al 2020) or to classify the entire sleep session at once (Sridhar et al 2020). It is important to mention that the postprocessing step may be accomplished in real-time if detrending and normalization parameters are extracted from an individualized training phase. The rest of the approaches in table 1, while producing a sleep-stage estimation for every 30 s window (termed ‘sleep epoch’), analyze a longer window (usually 5 min) centered around the targeted sleep-epoch. Our DNN-based algorithm has a small memory footprint with a number of parameters that was at least 3 orders of magnitude smaller than comparable approaches, which utilized either approximately 260 000 (Radha et al 2019) or more than 1 million parameters, as estimated in independent replication of the DNN architecture in Sridhar et al study (Sridhar et al 2020). The causal processing and small size make our proposed algorithm ideal for real-time implementation.

**Conclusion**

The DNN model presented here utilized IBI signals instead of feature extraction to generate real-time sleep-stage identification. Utilizing signals extracted from contactless sensing, our model provides sleep staging of 30 s epochs within 1 min, allowing for real-time assessment and intervention. Additionally, the small size of our model (approximately 6400 parameters, several orders of magnitude smaller compared to similar approaches) allows it to be utilized across different platforms. Kappa optimization of the algorithm used thresholds based on probability ratios for specific sleep stages, which enabled rapid adaptation of the algorithm to sleep-stage detection using different datasets. However, performance of the model was variable across datasets, particularly for datasets with markedly different demographic parameters or that included sleep disorders. This suggests that future algorithms should incorporate additional variable datasets to ensure broader applicability of the analysis, potentially leveraging techniques such as transfer learning (Panigrahi et al 2021). The DNN architecture proposed here was based on the optimization work on EEG signals described in Bresch et al 2018. Future work should focus on architecture options that directly use IBI signals and consider deeper networks and/or longer convolution outputs. Further investigation should also focus on the mathematical foundation of the Kappa optimization approach based on thresholds on the probability ratios and the conditions under which this outperforms other alternatives such as weighted cross-entropy loss.
Acknowledgments

This study received funding from Sleep Number Corporation. G G-M and J J are employees of Sleep Number Corporation, the manufacturer of the Sleep Number smart bed. The authors declare no other competing interests.

Author contributions are as follows: G G-M, algorithm design, data analysis, and manuscript writing; J J, data analysis, reproduction of the Sridhar et al DNN architecture, and manuscript writing.

All authors reviewed the manuscript and agreed to be accountable for the content.

The authors would like to thank Daniel Polimac, Farzad Siyahjani, Cristina Jocson, Anna Pastushko, and Shawn Barr from Sleep Number labs for critically reviewing this work and providing insightful comments to improve the content presented in the manuscript.

Dataset D1 was made publicly available through PhysioNet by the Computational Clinical Neurophysiology Laboratory and the Clinical Data Animation Laboratory at Massachusetts General Hospital, Boston, MA, USA.

Dataset D2 was made publicly available through PhysioNet by the Sleep Disorders Center of the Ospedale Maggiore of Parma, Italy.

Medical writing and editorial support were provided by Rachel C. Brown, PhD, and Sandra J. Page, PhD, of Oxford PharmaGenesis Inc., Newtown, PA, USA, and was funded by Sleep Number Corporation.

Data availability statement

Datasets D1 and D2 used in this study are publicly available at: Physionet.org (Goldberger et al 2000), under the names ‘You Snooze You Win – The PhysioNet Computing in Cardiology Challenge 2018’ dataset (https://physionet.org/content/challenge-2018/1.0.0/#files) and ‘CAP sleep’ dataset (https://www.physionet.org/content/capslpdb/1.0.0/), respectively.

Dataset D3 is not publicly available due to limitations in the consent form.

Ethics statement

This report used 3 sources of data (D1, D2, and D3). D1 and D2 are publicly available and contain only anonymized dataset information. The dataset that generated Dataset D3 was submitted to the Institutional Review Board of the University of Chicago (IRB00000331, IRB00000735, IRB00002169) and received ‘exempt status.’

References

Barone D A et al 2017 Heart rate variability in restless legs syndrome and periodic limb movements of Sleep Sleep. Sci. 10 80–6
Bellesi M et al 2014 Enhancement of sleep slow waves: underlying mechanisms and practical consequences Front. Syst. Neurosci. 8 1–17
Bresch E, Grossekathofer U and Garcia-Molina G 2018 Recurrent deep neural networks for real-time sleep stage classification from single channel EEG Front. Comput. Neurosci. 12 1–12
Carskadon M and Dement W 2011 Principles and Practice of Sleep Medicine ed ed M Kryger et al (St Louis: Elsevier Saunders) pp 16–26
Chollet F 2015 Keras https://github.com/fchollet/keras
Chouchou F and Desseilles M 2014 Heart rate variability: a tool to explore the sleeping brain? Front Neurosci 8 1–9
Chriskos P et al 2021 A review on current trends in automatic sleep staging through bio-signal recordings and future challenges Sleep Med. Rev. 55 101377
Cohen J 2016 A coefficient of agreement for nominal scales Educ. Psychol. Meas. 20 37–46
Cui Y et al 2019 Class-balanced loss based on effective number of samples 2019 IEEE/CVF Conf. Comp. Vision. Patt. Recog. pp 9260–9
Dobellemaniere E et al 2018 Performance of an ambulatory Dry-EEG device for auditory closed-loop stimulation of sleep slow oscillations in the home environment Front. Hum. Neurosci. 12 1–15
Diep C et al 2020 Acoustic slow wave sleep enhancement via a novel, automated device improves executive function in middle-aged men Sleep 43 1–11
Dorffner G, Vitt M and Anderer P 2015 The effects of aging on sleep architecture in healthy subjects Adv. Exp. Med. Biol. 821 93–100
Elschenbruch S, Harnish M and Orr W 1999 Heart rate variability during waking and sleep in healthy males and females Sleep 22 1067–71
Fonseca P et al 2015 Sleep stage classification with ECG and respiratory effort Physiol. Meas. 36 2027–40
Fonseca P et al 2017 Validation of photoplethysmography–based sleep staging compared with polysomnography in healthy middle-aged adults Sleep 40 1–10
Franzen P and Buyse D 2008 Sleep disturbances and depression: risk relationships for subsequent depression and therapeutic implications Dialogue Clin Neurosci 10 473–81
Garcia-Molina G et al 2018 Closed-loop system to enhance slow wave activity J. Neural Eng. 15 066018
Garcia-Molina G et al 2019 Hybrid in-phase and continuous auditory stimulation significantly enhances slow wave activity during sleep Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. pp 4052–5
Geng D-Y et al 2020 A decision support system for automatic sleep staging from HRV using wavelet packet decomposition and energy features Bioméd. Signal Process. Control 56 1–9
Ghassemi M et al 2018 The physionet/computing in cardiology challenge 2018 2018 Computing in Cardiology Conf. (CinC), I–4
Glorot X and Bengio Y 2010 Understanding the difficulty of training deep feedforward neural networks Xavier Proc. of the Thirteenth Int. Conf. on Artificial Intelligence and Statistics 9 249–256
Goldberger A L et al 2000 Physio bank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals Circulation 101 e215–20
Goodfellow I, Bengio Y and Courville A 2016 Deep Learning (Cambridge, MA: MIT Press) pp 163–220
Haba-Rubio J et al 2018 Prevalence and determinants of rapid eye movement sleep behavior disorder in the general population Sleep 41 1–8
Hochreiter S and Schmidhuber J 1997 Long short-term memory Neuronal Comput. 9 1735–80
Javaheri S and Redline S 2012 Sleep, slow-wave sleep, and blood pressure Curr. Hypertens. Rep. 14 442–8
Kemper K, Hamilton C and Atkinson M 2007 Heart rate variability: impact of differences in outlier identification and management strategies on common measures in three clinical populations Pediatr. Res. 62 337–62
Kingma D and Ba J 2015 ADAM: a method for stochastic optimization In: Proceedings of the International Conference on Learning Representations (San Diego, CA)
Kiranyaz S et al 2021 1D convolutional neural networks and applications: a survey Mech. Syst. Signal. Process. 151 1–21
Kumar P et al 2021 Classification of imbalanced data: review of methods and applications IOP Conf. Ser.: Mater. Sci. Eng. 1099 012077
Landis JR and Koch G G 1977 The measurement of observer agreement for categorical data Biometrics 33 159–74
Lim Y G, Kim K K and Park K S 2007 ECG recording on a bed during sleep without direct skin-contact IEEE Trans. Biomed. Eng. 54 718–25
Lippman N, Stein K M and Lerman B B 1994 Comparison of methods for removal of ectopy in measurement of heart rate variability Am. J. Physiol. Heart Circ. Physiol. 267 H411–8
Liu G et al 2014 Robustness evaluation of heart rate variability measures for age gender related autonomic changes in healthy volunteers Australas. Phys. Eng. Sci. Med. 37 567–74
Maquet P 2000 Sleep on it! Nat. Neurosci. 3 1235–6
Minis-tieri C M et al 2016 50. Decreased heart rate variability in REM sleep behavior disorder Clin. Neurophysiol. 127 e144
Muzet A et al 2016 Assessing sleep architecture and continuity measures through the analysis of heart rate and wrist movement recordings in healthy subjects: comparison with results based on polysomnography Sleep Med. 21 47–56
Ohayon M M et al 2004 Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: developing normative sleep values across the human lifespan Sleep 27 1255–73
Ohayon M M and Roth T 2001 What are the contributing factors for insomnia in the general population? J. Psychosom. Res. 51 745–55
Ohayon M M and Roth T 2002 Prevalence of restless legs syndrome and periodic limb movement in the general population J. Psychosom. Res. 53 547–54
Panagriha S, Nanda A and Swarnkar T 2021 Intelligent and Cloud Computing, ed D Misha et al (Singapore: Springer) pp 781–9
Papalambros N A et al 2017 Acoustic enhancement of sleep slow oscillations and concomitant memory improvement in older adults Front Hum Neurosci. 11 1–14
Patel A, Reddy V and Araujo J 2021 StatPears [Internet] (Florida: StatPears Publishing)
Radha M et al 2019 Sleep stage classification from heart-rate variability using long short-term memory neural networks Sci. Rep. 9 14149
Raymann R J, Swaab D F and Van Someren E J 2008 Skin deep: enhanced sleep depth by cutaneous temperature manipulation Brain 131 500–13
Russo M A, Santarelli D M and O’Rourke D 2017 The physiological effects of slow breathing in the healthy human Breathe (Sheff) 13 298–309
Shaffer F and Ginsberg J P 2017 An overview of heart rate variability metrics and norms Front. Public Health 5 1–17
Shen Q et al 2021 Multiscale deep neural network for obstructive sleep apnea detection using RR interval from single-lead ECG signal IEEE Trans. Instrum. Meas. 70 1–13
Slater G and Steier J 2012 Excessive daytime sleepiness in sleep disorders J. Thorac. Dis. 4 608–16
Spiegelhalder K et al 2011 Heart rate and heart rate variability in subjectively reported insomnia J. Sleep. Res. 20 137–45
Sridhar N et al 2020 Deep learning for automated sleep staging using instantaneous heart rate NPJ Digit. Med. 3 1–10
Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology (TFESCNASPE) 1996 Heart rate variability Circulation 93 1043–65
Tautan A M et al 2020 Automatic sleep stage detection A Study on the Influence of Various PSG Input Signals. (Annual International Conf. of the IEEE Engineering in Medicine and Biology Society pp 5330–4
Tirranno M et al 2002 Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep Sleep Med. 3 187–99
Tobaldini E et al 2013 Heart rate variability in normal and pathological sleep Front Physiol. 4 1–11
de la Torre J, Puig D and Valls A 2018 Weighted kappa loss function for multi-class classification of ordinal data in deep learning Pattern Recognit. Lett. 105 144–54
Usher F et al 2012 Machine learning of human sleep patterns based on stage bout durations Int. Conf. on Health Informatics - HEALINF, (Vilamoura, Algarve, Portugal pp 71–80
Vollmer M 2014 Robust detection of heart beats using dynamic thresholds and movingwindows Comput. Cardiol. 41 569–72
Van de Water A T, Holmes A and Hurley D A 2011 Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography—a systematic review J. Sleep. Res. 20 183–200
Wei R et al 2018 The research of sleep staging based on single-lead electrocardiogram and deep neural network Biomed. Eng. Lett. 8 87–93
Wei Y et al 2019 A multi-class automatic sleep staging method based on long short-term memory network using single-lead electrocardiogram signals IEEE Access 7 85959–70
Ye R et al 2019 Non-invasive in-home sleep stage classification using a ballistocardiography bed sensor IEEE EMBS Int. Conf. on Biomedical & Health Informatics (BHI), Chicago, IL USA
Yordanova J et al 2010 Differential associations of early- and late-night sleep with functional brain states promoting insight to abstract task regularity PLoS One 5 e9442
Yüzelbaş Ş et al 2018 Automatic sleep staging based on SVD, VMD, HHT and morphological features of single-lead ECG signal Expert Syst. Appl. 102 193–206
de Zambotti M et al 2018 Dynamic coupling between the central and autonomic nervous systems during sleep: a review Neurosci. Biobehav. Rev. 90 84–103
Zhao X and Sun G 2021 A multi-class automatic sleep staging method based on photoplethysmography signals Entropy 23 1–12
Zhou J et al 2017 Excessive daytime sleepiness predicts neurodegeneration in idiopathic REM sleep behavior disorder Sleep 40 1–7