SleepPPG-Net: A Deep Learning Algorithm for Robust Sleep Staging From Continuous Photoplethysmography

Kevin Kotzen, Peter H. Charlton, Sharon Salabi, Lea Amar, Amir Landesberg, and Joachim A. Behar, Senior Member, IEEE

Abstract—Sleep staging is an essential component in the diagnosis of sleep disorders and management of sleep health. Sleep is traditionally measured in a clinical setting and requires a labor-intensive labeling process. We hypothesize that it is possible to perform automated robust 4-class sleep staging using the raw photoplethysmography (PPG) time series and modern advances in deep learning (DL). We used two publicly available sleep databases that included raw PPG recordings, totalling 2,374 patients and 23,055 h: m: s of continuous data. We developed SleepPPG-Net, a DL model for 4-class sleep staging from the raw PPG time series. SleepPPG-Net was trained end-to-end and consists of a residual convolutional network for automatic feature extraction and a temporal convolutional network to capture long-range contextual information. We benchmarked the performance of SleepPPG-Net against models based on the best-reported state-of-the-art (SOTA) algorithms. When benchmarked on a held-out test set, SleepPPG-Net obtained a median Cohen’s Kappa (κ) score of 0.75 against 0.69 for the best SOTA approach. SleepPPG-Net showed good generalization performance to an external database, obtaining a κ score of 0.74 after transfer learning. Overall, SleepPPG-Net provides new SOTA performance. In addition, performance is high enough to open the path to the development of wearables that meet the requirements for usage in clinical applications such as the diagnosis and monitoring of obstructive sleep apnea.

Index Terms—Deep Learning, Photoplethysmography, Remote Health, Sleep staging.

I. INTRODUCTION

SLEEP is essential for human health and general well-being. Insufficient sleep and poor sleep quality are known to cause a myriad of physical and mental diseases such as cardiovascular disease, obesity, and depression [1]. Sleep disorders such as obstructive sleep apnea (OSA) are highly prevalent, affecting up to one-sixth of the global adult population [2]. Despite the impact on quality of life, many people with sleep disorders are unaware of their condition and remain undiagnosed [2].

Sleep disorders are traditionally diagnosed with a sleep study called polysomnography (PSG). During a PSG study, the patient is monitored and observed overnight, usually in a sleep laboratory. The patient is connected to sensors that measure and record several neurophysiological and cardiorespiratory variables [1]. PSG data is primarily labeled using Electroencephalography (EEG) in a manual or semi-manual manner by a technician trained in sleep scoring. Labels are assigned for each successive 30 s window called sleep epochs, henceforth referred to as “sleep-windows”. The PSG process is uncomfortable for the patient, who has to spend a night in a clinical environment, and labor-intensive, requiring a technician to monitor the patient overnight and another technician to perform manual sleep stage labeling. Furthermore, the number of clinics that perform PSG are limited and most clinics have long waiting times [3]. For example, in Australia patients wait an average of over 100 days for a PSG examination [3]. The limited availability of PSG make
repeated studies unfeasible and long-term monitoring of disease progression is currently not an option.

With the recent proliferation of wearable sensors and mobile health applications, there has been a rapid increase in the number of devices that aim to assess sleep quality and disorders more objectively and frequently, particularly targeting the monitoring of the individual in their home environment [4], [5], [6], [7]. The most common sensing technologies for at home sleep staging include portable EEG, wrist photoplethysmography (PPG), and mattress ballistocardiography (BCG).

Portable EEG has shown significant promise in terms of accuracy and clinical diagnostic potential [8]. However, its widespread use, beyond single night clinical evaluations, is limited since users must attach multiple electrodes to specific locations on their head or wear a tight-fitting headband in order to obtain EEG of sufficient quality. This is often uncomfortable with up to 60% of users reporting atypical sleep due to attached EEG devices [8]. In contrast, PPG acquisition simply requires that the user wear a watch or ring like device with embedded PPG. Most modern wearables already incorporate PPG [9]. However, the sleep staging accuracy of these devices has been shown to be limited and they have not yet met clinical requirements [10]. There is therefore great value in improving the accuracy of sleep staging from PPG to impact millions of daily users.

Sleep and the autonomous nervous system (ANS) are regulated by the same major central nervous system mechanisms resulting in a strong connection between sleep stage and ANS activity [11]. The ANS in turn regulates the cardiovascular and respiratory systems which makes these systems a good proxy for sleep measurement [11]. As reviewed by Ebrahimi et al. [12], research efforts to improve the clinical accuracy of sleep staging from cardiorespiratory waveforms have thus far mostly focused on the development of algorithms that perform sleep staging from the electrocardiogram (ECG). A vast majority of these works used feature engineering (FE) and recurrent neural network (RNN) for automated sleep staging [13], [14], [15]. Cohen’s Kappa (κ) performance for this FE-based approach has reached 0.60 [15]. More recently, Sridhar et al. [16] developed a deep learning (DL) model taking as input the instantaneous heart rate (IHR), i.e. a time series derived from the interbeat intervals (IBIs) computed from the ECG. Their DL model consists of a residual convolutional network (ResNet) followed by a temporal convolutional network (TCN). They reported in-domain test κ performance of 0.67 for the Sleep Heart Health Study (SHHS) and 0.69 for the Multi-Ethnic Study of Atherosclerosis (MESA), and out-of-domain generalization performance of 0.55 for the PhysioNet/Computing in Cardiology database.

There is significantly less work published on sleep staging from PPG than there is for ECG. Most works that use PPG usually do so in the context of transfer learning (TL), where models are trained on a large database of ECG derived heart rate variability (HRV) measures and then fine-tuned to a smaller database of pulse rate variability (PRV) measures derived from the IBIs detected on the PPG. These works report κ performance approaching 0.66 [17], [18], [19].

This research aims to demonstrate that sleep staging from the raw PPG, using an advanced DL approach, is superior to sleep staging approaches that use features or time series extracted from the IBIs of the PPG. Specifically, our main contributions are:

- We create SleepPPG-Net, a novel end-to-end DL algorithm for sleep staging working from the raw PPG signal.
- We benchmark the performance of SleepPPG-Net against state-of-the-art (SOTA) algorithms. When benchmarked on a held-out test set, SleepPPG-Net obtained a median κ of 0.75 against 0.69 for the best SOTA approach.
- We evaluate the generalization of SleepPPG-Net on an external test set and demonstrate good generalization performance (κ of 0.74) following a transfer learning re-calibration step.
- We validate the clinical usability of SleepPPG-Net by demonstrating it is possible to estimate important sleep metrics such as total sleep time, sleep efficiency, sleep-stage fractions and sleep stage transitions.

II. METHODS

We considered three machine learning (ML) approaches to sleep staging from PPG (Fig. 1). The first approach used handcrafted features engineered from the PPG and a neural network (NN) classifier, the second approach used derived time series (DTS) extracted from IBIs of the PPG as input to a DL classifier, and the third approach used the minimally-preprocessed PPG and DL. Models for the approaches are named Benchmark-FE (BM-FE), Benchmark-DTS (BM-DTS), and our new algorithm SleepPPG-Net. All models used a sequence-to-sequence architecture and were trained end-to-end.

A. Databases

Permission to use retrospective medical databases was granted following internal review board approval 62-2019. We used three labeled PSG databases in our experiments. SHHS Visit 1 [20], [21], totaling 5,758 unique patients, was used for model pretraining. MESA [21], [22], totaling 2,056 unique patients, was used for training and testing, and the Cleveland
Family Study (CFS) Visit-5 v1 [21], [23], totaling 324 unique patients, was used to evaluate generalization performance both with and without TL. Patients from MESA were randomly allocated to mutually exclusive train and tests sets, stratifying by age, gender, and apnea-hypopnea index (AHI). MESA-train contains 1,850 patients and MESA-test 204 patients. Patients from CFS were allocated into folds to support evaluation with TL. CFS-train consists of 4 overlapping folds with 240 patients each and CFS-test has 4 non-overlapping folds of 80 patients each. Performance was evaluated on MESA-test and CFS-test. Databases are described in more detail in Table I.

All databases were downloaded from the National Sleep Resource Center [21] and came with sleep stage labels that were manually assigned by sleep experts from the full PSG [20], [21], [22], [23]. Each PSG was labeled only once [20], [21], [22], [23]. The MESA PSG study was done in-home whereas the CFS PSG study was done in-laboratory. PSG data for MESA and CFS was acquired from the fingertip using Nonin 8000 series pulse oximeters. PPG sampling rates were 256 Hz in MESA, and 8 Hz and a stop-band attenuation of 40 dB. The filtered PPG was downsampled to form WAV_{PPG}. Low-pass filtering removes high-frequency noise and prevents aliasing during down-sampling. We specifically used a low-pass filter as we wished to keep lower frequency components such as breathing and capillary modulation intact. The filter was built using a zero-phase 8th order low-pass Chebyshev Type II filter with a cutoff frequency of 8 Hz and a stop-band attenuation of 40 dB. The filtered PPG was downsampled to 34.17 Hz using linear interpolation, reducing the computational and memory requirements for ML. We choose a sampling rate of 34.17 Hz as this resulted in 1024 (2^{10}) samples per 30 s sleep-window. By using a 2^{th} number we could maintain full temporal alignment of data with sleep-windows during ML testing.

### B. Sleep Stages

Modern sleep scoring follows guidelines maintained by the American Academy of Sleep Medicine (AASM) [24]. AASM sleep stages include wake, rapid eye movement (REM), and three non-rapid eye movement (NREM) stages denoted N1, N2, and N3. In this work we consider 4-class sleep staging with classes: wake, light (N1/N2), deep (N3), and REM. CFS and SHHS were not labeled using AASM but rather an older set of guidelines called Rechtschaffen and Kales (R&K). The major difference between R&K and AASM is that R&K contains an additional NREM stage. R&K NREM stages are denoted S1, S2, S3, and S4. We assign R&K labels to our 4-classes as follows: wake, light (S1/S2), deep (S3/S4), and REM.

### C. Data Preparation

#### 1) PPG Preprocessing

The PPG was filtered and downsampled to form WAV_{PPG}. Low-pass filtering removes high-frequency noise and prevents aliasing during down-sampling. We specifically used a low-pass filter as we wished to keep lower frequency components such as breathing and capillary modulation intact. The filter was built using a zero-phase 8th order low-pass Chebyshev Type II filter with a cutoff frequency of 8 Hz and a stop-band attenuation of 40 dB. The filtered PPG was downsampled to 34.17 Hz using linear interpolation, reducing the computational and memory requirements for ML. We choose a sampling rate of 34.17 Hz as this resulted in 1024 (2^{10}) samples per 30 s sleep-window. By using a 2^{th} number we could maintain full temporal alignment of data with sleep-windows during ML testing.

#### 2) Feature Engineering

Our FE and DTS approaches rely on robust detection of peaks on the PPG. We used a band-pass filter to remove noise from the PPG that would otherwise affect peak detection. This filtering stage was independent of WAV_{PPG} preprocessing. The band-pass filter was designed to have a minimal impact on the morphology of the PPG. Given that the heart beats in a range of around 40-100 bpm (0.66 Hz–1.66 Hz) and based on a review of the literature around the optimal filtering of PPGs [25], [26], we used a band-pass filter with a pass-band of 0.4-8 Hz. The filter was built using a zero-phase 8th order band-pass Chebyshev Type II filter with a pass-band of 0.4-8 Hz and stop-band attenuation of 40 dB. PPG peaks were detected from the filtered time series using an automatic beat detection algorithm developed by Aboy et al. [27] and implemented in the PulseAnalyse toolbox [28]. This PPG peak detector was chosen because it demonstrated the highest peak detection performance when evaluated on PPGs recorded during PSG [29]. For SHHS, the ECG peaks were detected using epltd0 [30] a SOTA ECG peak detection algorithm.

PRV and HRV measures were extracted using the Python HRV features implemented in [31]. This library calculates 21 HRV measures per set of IBIs. Morphological measures (MOR) were extracted from the time domain, first and second order derivatives, and the frequency domain of the PPG. A total of 41 features were extracted using a MOR toolbox developed within the context of this research. We calculated measures for each sleep-window twice. First only for the current sleep-window and then again with the two preceding and proceeding windows included. We did this because HRV measures should be calculated with a time span of at least two and a half minutes, but sleep-windows are only 30 s. We standardized MOR and PRV features on a per-patient basis. This per-patient standardization acts as a form of personalization and eliminates differences in baseline values between patients.

#### 3) Instantaneous Pulse Rate

The IHR and instantaneous pulse rate (IPR) were extracted from the IBIs according to the methods described by Sridhar et al. [16]. The only modification made was that we used a re-sampling rate of 2.17 Hz, as opposed to 2.0 Hz as this yielded 64 (2^{6}) samples per 30 s sleep-window.

### Table I

| Patients | SHHS | MESA | CFS |
|----------|------|------|-----|
| Gender (M:F) | 1:1.1 | 1:1 | 1:1.2 |
| Total Windows | 5.83M | 2.35M | 0.37M |
| Duration (hrs) | 9 [8-9] | 10 [9-10] | 10 [9-10] |
| Age (yrs) | 63 [55-72] | 68 [62-76] | 42 [21-54] |
| Wake (%) | 27 [19-35] | 37 [30-47] | 34 [27-44] |
| Light (%) | 44 [36-52] | 43 [36-50] | 39 [29-46] |
| Deep (%) | 12 [6-18] | 5 [1-10] | 12 [7-19] |
| REM (%) | 14 [10-17] | 11 [7-14] | 11 [8-14] |
By using a $2^n$ number we could maintain full temporal alignment of data with sleep-windows during ML pooling operations.

### D. Machine Learning

We define our problem as follows: given a sequence of $L$ ordered sleep-windows, with input signal $S$ and labels $P$, map the input signal to the labels using network $F$ such that $F(S) \rightarrow P$. Sleep-windows are indexed with subscript $l$, where $\{1 : \ldots : L\}$ refers to the $l$-th sleep-window in the sequence. In line with other sequence-to-sequence models, we break $F$ into parts, namely, a sleep-window encoder $F_E$, a sequence encoder $F_S$, and a classifier $F_C$. The $F_E$ extracts information from each individual $S_l$, translating the high dimensionality inputs into a lower-dimensional space called an embedding $X_l$ such that $F_E(S_l) \rightarrow X_l$. $F_S$ then exploits the cyclic and phasic nature of sleep and considers the sequence as a whole, adding contextual information to each $X_l$, by looking at neighboring embeddings $X_{l-\ldots\ldots\cdot l_{l+i}}$, where $i$ is the receptive-field of $F_S$, resulting in a richer representation $Z_l$ such that $F_S(X_l) \rightarrow Z_l$. Finally, $F_C$ computes a probability prediction of each sleep-stage at each sleep-window $P_l$ from $Z_l$ such that $F_C(Z_l) \rightarrow P_l$. We further define $M$ as demographic data, $n_x$ as the size of $S_l$, $n_z$ as the size of $X_l$, $n_h$ as the size of $Z_l$, $n_h$ as the number of hidden units in a RNN, and $C$ as the number of output classes. For 4-class sleep staging $C = 4$. We feed $M, E$ to each model by concatenating $M$ to each $X_l$.

1) **BM-FE Model**: BM-FE model architecture is similar to the model developed by Radha et al. [17]. The input $S$ consists of a sequence of PRV and MOR features. $L = 1, 200$ and $n_x = 126$. The $F_E$ consists of a 5-layer time-distributed deep neural network (DNN). Time-distribution applies the same encapsulated layer to each temporal slice. The $F_S$ consists of 2-stacked bidirectional long short-term memory (LSTM) layers each with a temporal-reach of 100 sleep-windows, and the $F_C$ is a 4-layer time-distributed DNN. Dropout is used in the $F_C$ for regularization. $n_x = 16, n_h = 128$, and $n_z = 256$. A full description of the model and hyperparameters are presented in Supplement II.

2) **BM-DTS Model**: The BM-DTS model architecture was based on Sridhar et al. [16] with some minor modifications. The input $S$ is the continuous IPR time series. $F_E$ consists of 3 time-distributed residual convolution (ResConv) blocks followed by a time distributed DNN. Each ResConv has 3 1D-convolutions followed by pooling layer and residual addition. $F_S$ uses 2 stacked TCNs. Each TCN consists of 5 dilated 1D-convolutions followed by residual addition and dropout. $F_C$ is simply a 1D-convolution. $L = 1200, n_x = 256, n_e = 128$ and $n_z = 128$. A full description of the model including parameters is presented in Supplement III.

3) **SleepPPG-Net**: SleepPPG-Net was inspired by WaveNet [32], Wave2Vec [33] and Sridhar et al. [16]. SleepPPG-Net architecture is shown in Fig. 2. The input is the continuous PPG time series. $F_E$ extracts continuous embeddings directly from the input. $F_E$ consists of 8 stacked ResConvs, with each ResConv containing 3 1D-convolutions followed by max pooling and residual addition. A windowing layer follows $F_E$ to reestablish temporal windows. Temporal windows are then further compressed using a time-distributed DNN. Long-range temporal information is added to each temporal window using $F_E$ which consists of 2 stacked TCN blocks. Each TCN contains 5 dilated 1D-convolutions followed by residual addition and dropout. Finally, $F_C$ uses a 1D-convolution to make predictions. The Leaky ReLU activation function was used in all layers except the output layer which uses the Softmax activation function.

4) **Pretraining**: Pretraining is the process of specifically training a ML model with the intention of using the pretrained model as a starting point for solving other problems. Pretraining improves training convergence times and sometimes improves model performance [34]. We pretrained our models on ECG data from the SHHS database. BM-FE was pretrained on HRV measures derived from the ECG. BM-DTS was pretrained on the IHR derived from the IBIs of the ECG and SleepPPG-Net was pretrained on the raw ECG.

5) **Transfer Learning**: TL is an effective means of adapting a ML model from one domain to another. In our work, we use the term TL to denote adaption to a specific external database. We applied TL to our external database using 4-folds. The pretrained model was used as a starting point and each fold was trained and
evaluated independently, before all results were brought together and analyzed as a whole.

6) **Training**: Models were built using Keras 2.6 and trained using a single NVIDIA A100 GPU. Loss was calculated using the categorical cross-entropy loss. Temporal sample weighting was used to address class imbalance and remove padded regions from loss calculations. Temporal weighting was applied by allocating a sample weight to each sleep-window according to its overall class prevalence in the training set. Padded regions were negated with a sample weight allocation of 0, effectively removing them from all loss calculations. The Adam optimizer was used. Hyperparameters for the BM-FE model were selected through manual experimentation. For BM-DTS and SleepPPG-Net we used 100 Bayesian optimization iterations to tune the model hyperparameters. Initial weights for convolutional neural network, DNN, and LSTM layers were set with Xavier uniform initialization. When training SleepPPG-Net models from scratch, we used a learning rate of $2.5 \times 10^{-4}$ and trained for 30 epochs. When training models for the With-pretrain and With-TL training schemes we used a learning rate of $1.0 \times 10^{-4}$ and trained for 5 epochs. A batch size of 8 was used in all experiments.

### E. Performance Measures

The models output a probability prediction for each of the 4 sleep stages at each sleep-window in the full sequence. Probabilities were converted into predictions by selecting the class with the highest probability. All padded regions were removed before calculating performance measures. Performance was evaluated using $\kappa$ and $Ac$. We calculate the $\kappa$ and $Ac$ per patient. The hypnogram labels assigned by sleep experts during PSG scoring are considered to be the ground truth. The final reported scores represent the median $\kappa$ and median $Ac$ for all patients in the test set. Significance of results was computed using the Kolmogorov-Smirnov test for continuous distributions, and the Student’s $t$-test when only the mean and standard deviation were known. We evaluated performance across calculated performance metrics per patient population groups to including: age, sex, race, smoking status, apnea severity, hypertension diagnosis, diabetes diagnosis, and beta blocker usage.

Common sleep metrics obtainable from the PSG include; total sleep time ($TST$), sleep efficiency (SE), sleep-stage fractions ($FR_{Light}$, $FR_{Deep}$, $FR_{REM}$), and sleep stage transitions (Transitions). The formulae used are shown in (1)-4. We evaluated the degree to which the sleep metrics calculated from the sleep stages predicted by our models matched the ones calculated from the ground truth. The degree of agreement was quantified using the mean square error (MSE) and R-Squared errors ($R^2$).

\[
TST = \sum Light + \sum Deep + \sum REM
\]

\[
SE = \frac{TST}{TST + \sum Wake} \times 100
\]

\[
FR_{Stage} = \frac{\sum Stage}{TST} \times 100
\]

\[
Transitions = \frac{\sum \text{Transition Between Stages}}{\text{Total Recording Time}}
\]

### III. Results

Models without pretraining (No-pretrain) and without transfer learning (No-TL) were trained from Xavier initialization. Models with pretraining (With-pretrain) and with transfer learning (With-TL) were trained with the weights obtained during initial training. Models are evaluated on a held-out subset of MESA denoted as MESA-test, and on 4 non-overlapping folds of the Cleveland Family Study (CFS) denoted as CFS-test.

Table II presents evaluation results for MESA-test. We show performance for models trained with Xavier initialization and on MESA-train (i.e. No-pretrain) and then for models pretrained on ECG from SHHS (i.e. With-pretrain). Pretraining on SHHS did not have an important effect on the performance of BM-FE or SleepPPG-Net models but significantly improved ($p = 0.0002$, Kolmogorov-Smirnov test) BM-DTS performance from a $\kappa$ of 0.64 (0.56 to 0.72) and of 76% to a $\kappa$ of 0.69 (0.62 to 0.77) and $Ac$ of 80%. The best performing model was the pretrained SleepPPG-Net which scored a $\kappa$ of 0.75 (0.69 to 0.81) and $Ac$ of 84%. The confusion matrix for the pretrained SleepPPG-Net is presented in Fig. 3. The $\kappa$ distribution for the pretrained BM-FE, BM-DTS, and SleepPPG-Net models are compared in Fig. 4.

Table III presents generalization to external test set results using CFS-test for evaluation. We first show performance for models with pretraining on ECG from SHHS and MESA-train.
Fig. 4. Distribution of \( \kappa \) performance for BM-FE, BM-DTS and SleepPPG-Net on MESA-test \((n = 204)\). All models were pretrained on ECG from SHHS and then trained on MESA-train.

Fig. 5. Confusion matrix for SleepPPG-Net with TL, evaluated on CFS-test \((n = 320)\).

Fig. 6. TL performance on CFS-test by number of patients from CFS-train used for TL. A random subset of patients is taken from each CFS-train fold for training. Experiments were run multiple times and the average is shown.

but before TL (i.e. No-TL), and then show performance for the same models after applying TL using CFS-train (i.e. With-TL). Before TL SleepPPG-Net scored a \( \kappa \) of 0.67 (0.55 to 0.74) and \( Ac \) of 76%. With TL SleepPPG-Net scored a \( \kappa \) of 0.74 (0.66 to 0.79) and \( Ac \) of 82% which is significantly better \((p = 0.0005, \text{Kolmogorov-Smirnov test})\). The confusion matrix of SleepPPG-Net with TL is presented in Fig. 5. To determine the number of patients needed for effective TL, we evaluate the \( \kappa \) performance on CFS-test as a function of the number of patients used for TL as depicted in Fig. 6. Performance for SleepPPG-Net with TL improved from a \( \kappa \) of 0.68 (0.56 to 0.75) when using only 10 patients to 0.73 (0.64 to 0.78) when using 120 patients.

The per group \( \kappa \) performance is presented in Fig. 7. Performance is not affected by gender, race, or presence of diabetes. Performance is lower in patient groups with higher apnea severity, older age, hypertension diagnosis and beta blocker usage. Performance is higher for patients that smoke.

We evaluated sleep metrics for MESA-test \((n = 204)\) using the pretrained SleepPPG-Net. In Fig. 8 we compare the predicted sleep metrics to those calculated from the ground truth. The pretrained SleepPPG-Net scored a MSE of 0.39 hours for total sleep time, 7.87% for Light fraction, 6.55% for Deep fraction, 4.08% for REM fraction, 4.1% for Sleep Efficiency, and 4.2 transitions/hour for Transitions.

IV. DISCUSSION AND CONCLUSION

When interpreting the performance of automated sleep staging algorithms it is important to keep in mind that manual scoring by humans is highly subjective \([35]\). Inter-rater agreement for 5-class PSG labeled by human scorers is reported as a \( \kappa \) of 0.76 \((95\% \text{ confidence interval, 0.71–0.81})\) \([36]\). Common mistakes between human scorers during PSG include confusion between wake and light sleep and light sleep and deep sleep \([37]\). While our problem is somewhat simplified in that we consider 4-class sleep staging, human inter-rater scores provide a sense of the highest performance that may be reached by data-driven algorithms.

The first important contribution of this research is the novel SleepPPG-Net algorithm. SleepPPG-Net was demonstrated to significantly \((p < 0.0004, \text{Kolmogorov-Smirnov test})\) outperform SOTA algorithms including BM-FE and BM-DTS. On the held out test set of 204 MESA patients, SleepPPG-Net scored a \( \kappa \) of 0.75 against 0.66 for BM-FE and 0.69 for BM-DTS approaches. SleepPPG-Net performance is also significantly \((p < 0.001, \text{two-sample t-test})\) higher than the current published SOTA results for sleep staging from PPG which stand at a \( \kappa \) of 0.66 \([17]\), \([18]\), and significantly \((p = 0.02, \text{two-sample t-test})\) higher than the current SOTA results for sleep-sleep staging from ECG which are reported at \( \kappa \) of 0.69 \([16]\). While reporting performances from other studies, it is important to keep in mind that these may be for a different cohorts/different train-test split and thus a direct comparison may not possible. Fig. 9 presents an example of the hypnograms generated by BM-FE, DB-DTS and SleepPPG-Net for a single patient. Performance for this patient is best for the SleepPPG-Net model which accurately detects all important sleep structures. We believe that the improved performance achieved by SleepPPG-Net over other approaches can be attributed to several factors. First, SleepPPG-Net does not require the annotation of fiduciaries using a PPG peak detector. PPG peak detectors are sensitive to noise and are often unable to handle irregular health rhythms. This may result in noisy and inaccurate IBIs which are relied upon by FE and DTS approaches. Second, SleepPPG-Net extracts relevant features from the data automatically thus going beyond domain knowledge. FE approaches use PRV and MOR measures which have been developed as measures of general cardiovascular functioning and may not be optimized for sleep staging. Third, in using only IBI data, any information contained within the PPG that is not directly related to the heart rate is lost. We included MOR measures in an attempt to include some of this information in our
BM-FE model, but as previously stated, these measures are not optimized to sleep staging. Additional information embedded in the raw PPG may include respiration rate, blood pressure, stroke volume, cardiac contractility, peripheral vascular tone and pulse-transit time which are all regulated by the ANS [38]. Finally, the choice of sequence encoder used in SleepPPG-Net is important. The TCN is likely better suited to extract the long-term contextual information than the RNN used in the BM-FE model.

Fig. 7. SleepPPG-Net with pretrain performance per clinical group evaluated on MESA-test (n = 204). We evaluated performance based on age group, sex, race, smoking status, apnea severity, hypertension diagnosis, diabetes diagnosis and beta blocker usage.

Fig. 8. Comparison between sleep metrics of MESA-test (n = 204) calculated from pretrained SleepPPG-Net to those calculated using ground truth. Dotted line shows the MSE.

The second important finding of the research is that pretraining SleepPPG-Net on a PSG database with ECG (thus pretraining in another domain) proved to be an effective means of speeding up training convergence. When trained from scratch, SleepPPG-Net needs to be trained for 30 epochs, whereas when trained from the ECG domain pretrained model, convergence was reached after only 5 epochs. Given the ease that SleepPPG-Net adapts from ECG to PPG, we expect that our pretrained SleepPPG-Net model can be leveraged to develop models with new signal-domains such as wrist-based PPG used in wearables.

The third important finding of this research is that SleepPPG-Net demonstrates good generalizability, scoring a $\kappa$ of 0.67 (0.55 to 0.74) on CFS-test with no TL step. This is markedly higher than the generalization performance reported by Sridhar et al. [16], whose model scored a $\kappa$ of 0.55 on the PhysioNet/Computing in Cardiology Sleep database [39]. With TL,
performance of SleepPPG-Net increased significantly reaching a $\kappa$ of 0.74 (0.66 to 0.79). The number of CFS patients needed for effective adaption to CFS is shown to be 120. However, even with only 50 patients performance reaches a $\kappa$ of 0.71 (0.63 to 0.79). These results are promising as they indicate that SleepPPG-Net can effectively be fine-tuned to a new population sample using significantly fewer patients than was required for its original training. This will reduce the time and cost involved in the development of new sleep staging devices.

**Recommendation:** We recommend that sleep staging from PPG be performed using the raw PPG time series and SleepPPG-Net architecture. To obtain optimal results we suggest pre-training SleepPPG-Net with ECG from a large sleep databases such as SHHS before training on PPG from MESA or another sleep dataset. For optimal generalization performance to a new database, transfer learning with at least 120 patients should be used.

**Limitations:** An analysis of per class performance shows that SleepPPG-Net struggles in some areas. Deep sleep is consistently underestimated and is often confused with light sleep. This is likely due to the similarity of the cardiovascular and pulmonary characteristics expressed during deep and light sleep. For applications such as the general detection of OSA, this may not be a problem as light and deep sleep can be grouped without affecting diagnosis [40]. However, for disorders such as night terrors or sleepwalking and it is important to distinguish between light and deep sleep [40]. The detection of sleep fragmentation is another issue. Our model fails to reliably detect very short awakenings. It is possible that while these changes are visible in the EEG, they are too rapid to be reflected by the cardiovascular activity. Wake periods that are longer than 1.5 minutes are accurately detected. SleepPPG-Net performance varied across patient demographic and clinical subgroups (Fig. 7). Lower performance were obtained for a high OSA severity which may be attributed to SleepPPG-Net limited ability in accurately detecting the repeated short awakening due to arousals which affect the sleep structure. Older patients are known to have disrupted sleep patterns [41] which may lead the model performance to be impaired. Finally, patients diagnosed with hypertension are likely taking medications, including beta-blockers, that can affect ANS modulation of the heart rate and the PPG morphology and thus affect the model performance. The performance for these subgroups may be improved by increasing the number of individuals from these subgroups in the training set in order for the model to better learn their characteristic patterns. While the PPG is sensitive to movement, the incorporation of accelerometer data may further improve performance. Alternative DL architectures should also be experimented in order to improve performance. In particular, models developed for sleep staging from EEG may be good candidates for further experimentation. SOTA EEG based sleep staging models such as SalientSleepNet [42] and SleepTransformer [43] have demonstrated $\kappa$ scores in excess of 0.80. Evaluating these architectures on PPG (versus EEG) will be valuable in future work. Finally, providing an estimate of prediction uncertainty, such as using Monte Carlo Dropout [44], may be valuable in order to provide the human interpreter with some reliability measures on the produced hypnogram.

Overall, SleepPPG-Net demonstrates SOTA performance for sleep staging from PPG. SleepPPG-Net is shown to perform well across patient groups and is easily adapted to new databases and measurement settings. As such, SleepPPG-Net paves the way for the development of sleep staging applications from wearable devices that are accurate enough for clinical diagnosis. This will allow for improved detection, monitoring, and treatment of sleep disorders in the general population.

**ACKNOWLEDGMENT**

Contributions: KK and JB conceptualized the new methods (experimental setting and algorithms). KK performed the formal analysis, software development (feature engineering, statistical analysis and machine learning models, developments), methodology and writing of the original draft under the supervision of JB. SS and LA contributed some of the features used for the benchmark FE model. PHC contributed an implementation of the PPG peak detection algorithm. AL contributed to the discussion in interpreting the results, in particular relating them to the underlying physiology. All authors reviewed and commented the manuscript and approved it.

**REFERENCES**

[1] S. L. Worley, “The extraordinary importance of sleep: The detrimental effects of inadequate sleep on health and public safety drive an explosion of sleep research,” *Pharm. Therapeutics*, vol. 43, 2018, Art. no. 758.

[2] A. V. Benjafeld et al., “Estimation of the global prevalence and burden of obstructive sleep apnoea: A literature-based analysis,” *Lancet Respir. Med.*, vol. 7, no. 8, pp. 687–698, 2019.

[3] B. Sriram, V. Singh, S. Bandaralage, and I. Bashford, “P135 an audit investigating the length of time from GP referral to diagnostic polysomnography testing in an australian tertiary center.” *SLEEP Adv.*, vol. 2, no. Supplement_1, pp. A65–A65, 2021.

[4] J. Behar et al., “SleepAp: An automated obstructive sleep apnoea screening application for smartphones,” *IEEE J. Biomed. Health Inform.*, vol. 19, no. 1, pp. 325–331, Jan. 2015.

[5] J. A. Behar et al., “PhysioZoo: A novel open access platform for heart rate variability analysis of mammalian electrocardiographic data,” *Front. Physiol.*, vol. 9, 2018, Art. no. 1390.

[6] J. A. Behar et al., “Feasibility of single channel oximetry for mass screening of obstructive sleep apnea,” *Clinical Medicine*, vol. 11, pp. 81–88, 2019.

[7] S. A. Imtiaz, “A systematic review of sensing technologies for wearable sleep staging,” *Sensors*, vol. 21, no. 5, 2021, Art. no. 1562.

[8] J. R. Lunsford-Avery, C. Keller, S. H. Kollins, A. D. Krystal, L. Jackson, and M. M. Engelhard, “Feasibility and acceptability of wearable sleep electroencephalogram device use in adolescents: Observational study,” *JMIR mHealth uHealth*, vol. 8, 2020, Art. no. e20590.

[9] B. Bent, B. A. Goldstein, W. A. Kibbe, and J. P. Dunn, “Investigating sources of inaccuracy in wearable optical heart rate sensors,” *NPJ Digit. Med.*, vol. 3, pp. 1–9, 2020.

[10] E. D. Chinoy et al., “Performance of seven consumer sleep-tracking devices compared with polysomnography,” *Sleep*, vol. 44, no. 5, 2021, Art. no. zsa291.

[11] R. Cabiddu, S. Cerutti, G. Viardot, S. Werner, and A. M. Bianchi, “Modulation of the sympatho-vagal balance during sleep: Frequency domain analysis of obstructive sleep apnoea,” *Front. Physiol.*, vol. 4, 2013, Art. no. 45.

[12] F. Ebrahimi and I. Alizadeh, “Automatic sleep staging by cardiorespiratory signals: A systematic review,” *Sleep Breathing*, vol. 26, no. 2, pp. 965–981, Jun. 2022.

[13] R. Wei, X. Zhang, J. Wang, and X. Dang, “The research of sleep staging based on single-lead electrocardiogram and deep neural network,” *Biomed. Eng. Lett.*, vol. 8, no. 1, 2018, Art. no. 87.

[14] H. Sun et al., “Sleep staging from electrocardiography and respiration with deep learning,” *Sleep*, vol. 43, no. 7, 2020, Art. no. zsz306.
[15] P. Fonseca et al., “Automatic sleep staging using heart rate variability, body movements, and recurrent neural networks in a sleep disordered population,” Sleep, vol. 43, no. 9, pp. 1–10, 2020.

[16] N. Sridhar et al., “Deep learning for automated sleep staging using instantaneous heart rate,” NPJ Digit. Med., vol. 3, no. 1, pp. 1–10, 2020.

[17] M. Radha et al., “A deep transfer learning approach for wearable sleep stage classification with photoplethysmography,” NPJ Digit. Med., vol. 4, no. 1, 2021, Art. no. 135.

[18] B. M. Wulkerken et al., “It is all in the wrist: Wearable sleep staging in a clinical population versus reference polysomnography,” Nature Sci. Sleep, vol. 13, 2021, Art. no. 885.

[19] R. Huttunen et al., “Assessment of obstructive sleep apnea-related sleep fragmentation utilizing deep learning-based sleep staging from photoplethysmography,” Sleep, vol. 44, no. 10, pp. 1–10, 2021.

[20] S. F. Quan et al., “The sleep heart health study: Design, rationale, and methods,” Sleep, vol. 20, no. 12, pp. 1077–1085, 1997.

[21] G.-Q. Zhang et al., “The National sleep research resource: Towards a sleep data commons,” J. Am. Med. Inform. Assoc., vol. 25, no. 10, pp. 1351–1358, 2018.

[22] X. Chen et al., “Racial/ethnic differences in sleep disturbances: The multi-ethnic study of Atherosclerosis (MESA),” SLEEP, vol. 38, no. 6, pp. 877–888, 2015.

[23] S. Redline et al., “The familial aggregation of obstructive sleep apnea,” Amer. J. Respir. Crit. Care Med., vol. 151, no. 3, pp. 682–687, 1995.

[24] R. B. Berry et al., “AASM Scoring Manual Updates for 2017 (Version 2.4),” J. Clin. Sleep Med.: JCSM: Official Pub. Amer. Acad. Sleep Med., vol. 13, no. 5, 2017, Art. no. 665.

[25] S. Hara et al., “Parameter optimization of motion artifact canceling PPG-based heart rate sensor by means of cross validation,” in Proc. IEEE 11th Int. Symp. Med. Inf. Commun. Technol., 2017, pp. 73–76.

[26] Y. Liang, M. Elgendi, Z. Chen, and R. Ward, “An optimal filter for short photoplethysmogram signals,” in Proc. IEEE Comput. Cardiol. (CinC), 2021, pp. 1–4.

[27] M. Aboy, J. McNames, T. Thong, M. Ellenby, and B. Goldstein, “An automatic beat detection algorithm for pressure signals,” IEEE Trans. Biomed. Eng., vol. 52, no. 10, pp. 1662–1670, Oct. 2005.

[28] P. H. Charlton, J. Mariscal Harana, S. Vennin, Y. Li, P. Chowienczyk, and J. Alastruey, “Modeling arterial pulse waves in healthy aging: A database for in silico evaluation of hemodynamics and pulse wave indexes,” Amer. J. Physiol.-Heart Circulatory Physiol., vol. 317, no. 5, pp. H1062–H1085, 2019.

[29] K. Kotzen, P. H. Charlton, A. Landesberg, and J. A. Behar, “Benchmarking photoplethysmography peak detection algorithms using the electrocardiogram signal as a reference,” in Proc. IEEE Comput. Cardiol. (CinC), 2021, pp. 1–4.

[30] P. Hamilton, “Open source ECG analysis,” in Proc. IEEE Comput. Cardiol., 2002, pp. 101–104.

[31] A. Chocron et al., “Machine learning for nocturnal mass diagnosis of atrial fibrillation in a population at risk of sleep-disordered breathing,” Physiol. Meas., vol. 41, no. 10, Nov. 2020, Art. no. 104001.

[32] A. van den Oord et al., “WaveNet: A generative model for raw audio,” in Proc. 9th ISCA Workshop Speech Synth. Workshop (SSW 9), 2016, p. 125.

[33] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition,” in Proc. Interspeech, 2019, pp. 3465–3469.

[34] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, “Transfusion: Understanding transfer learning for medical imaging,” in Proc. Adv. Neural Inf. Process. Syst., vol. 32, 2019, pp. 3347–3357.

[35] H. Phan and K. Mikkelsen, “Automatic sleep staging of EEG signals: Recent development, challenges, and future directions,” Physiol. Meas., vol. 43, no. 4, Apr. 2022.

[36] Y. J. Lee, J. Y. Lee, J. H. Cho, and J. H. Choi, “Inter-rater reliability of sleep stage scoring: A meta-analysis,” J. Clin. Sleep Med., vol. 18, no. 1, pp. 193–202, 2021.

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

[38] M. Elgendi, “On the analysis of fingertip photoplethysmogram signals,” Curr. Cardiol. Rev., vol. 8, no. 1, pp. 14–25, 2012.

[39] M. M. Ghassemi et al., “You snooze, you win: The physionet/computing in cardiology challenge 2018,” in Proc. Comput. Cardiol. Conf., 2018, vol. 45, pp. 1–4.

[40] Teofilø L. Lee-Chiong, Ed., Sleep: A Comprehensive Handbook, Hoboken, NJ, USA: Wiley, 2005.

[41] M. A. Carskadon, E. D. Brown, and W. C. Dement, “Sleep fragmentation in the elderly: Relationship to daytime sleep tendency,” Neurobiol. Aging, vol. 3, pp. 321–327, 1982.

[42] Z. Jia, Y. Lin, J. Wang, X. Wang, P. Xie, and Y. Zhang, “SalientSleep-Net: Multimodal salient wave detection network for sleep staging,” in Proc. 13th Int. Joint Conf. Artif. Intell., IJCAI-21, vol. 8, 2021, pp. 2614–2620.

[43] H. Phan, K. Mikkelsen, O. Y. Chén, P. Koch, A. Mertins, and M. D. Vos, “SleepTransformer: Automatic sleep staging with interpretability and uncertainty quantification,” IEEE Trans. Biomed. Eng., vol. 69, no. 8, pp. 2456–2467, Aug. 2022.

[44] Y. Gal and Z. A. Uk, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” in Proc. Int. Conf. Mach. Learn., 2016, pp. 1050–1059.