SleepPPG-Net: A deep learning algorithm for robust sleep staging from continuous photoplethysmography

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

SleepPPG-Net: A deep learning algorithm for robust sleep staging from continuous photoplethysmography

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

***Abstract*—Sleep staging is an essential component in the diagnosis of sleep disorders and management of sleep health. It 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 hours 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*—Sleep staging, Deep Learning, Photoplethysmography, Remote Health**

I. Introduction

Sleep is essential for human health, well-being, and longevity [1]. 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 [3]. PSG data is labeled using the electroencephalogram in a manual or semi-manual manner by a technician trained in sleep scoring. Labels are assigned for each successive 30s 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 [4]. For example, in Australia and Singapore patients wait an average of 80 days for a PSG examination [4], [5]. 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 i.e. outside of the traditional clinical setting [6]–[10]. However, the accuracy of new wearablesis limited and these devices do not yet meet clinical requirements [11], [12].

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 [13]. The ANS in turn regulates the cardiovascular and respiratory systems which makes these systems a good proxy for sleep measurement [14]. As reviewed by Ebrahimi et al. [15], 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 [16]–[19]. Cohen's Kappa (κ) performance for this FE-based approach has reached 0.60 [19]. More recently, Sridhar et al. [20] 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 [21].

Most novel wearable sensors are capable of recording continuous photoplethysmography (PPG). There is however 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 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 [22], [23]. Sleep staging from the raw PPG is a relatively novel approach. In 2020 Korkalainen et al. [24] used the PPG as input to a convolutional neural network followed by RNN to obtain κ performance of 0.55. Most recently Hutten et al. [25], under the supervision of Korkalainen, updated these results to a κ of 0.64.

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 deep learning algorithm for sleep staging working from the raw PPG signal.

P. H. Charlton receives funding from the British Heart Foundation grant FS/20/20/34626.

K. Kotzen, A. Landesberg and J. A. Behar are with the Biomedical Engineering Faculty at Technion-Israel Institute of Technology, Haifa, Israel. (e-mail: kevinkotzen@gmail.com, jbehar@technion.ac.il)

P. H. Charlton is with the Department of Public Health and Primary Care at University of Cambridge, Cambridge, UK

P. H. Charlton is with the Research Centre for Biomedical Engineering, University of London, London, UK

S. Salabi is with the Computer Science Faculty at Technion-Israel Institute of Technology, Haifa, Israel

L. Amar is with Mines ParisTech - PSL University, Paris, France

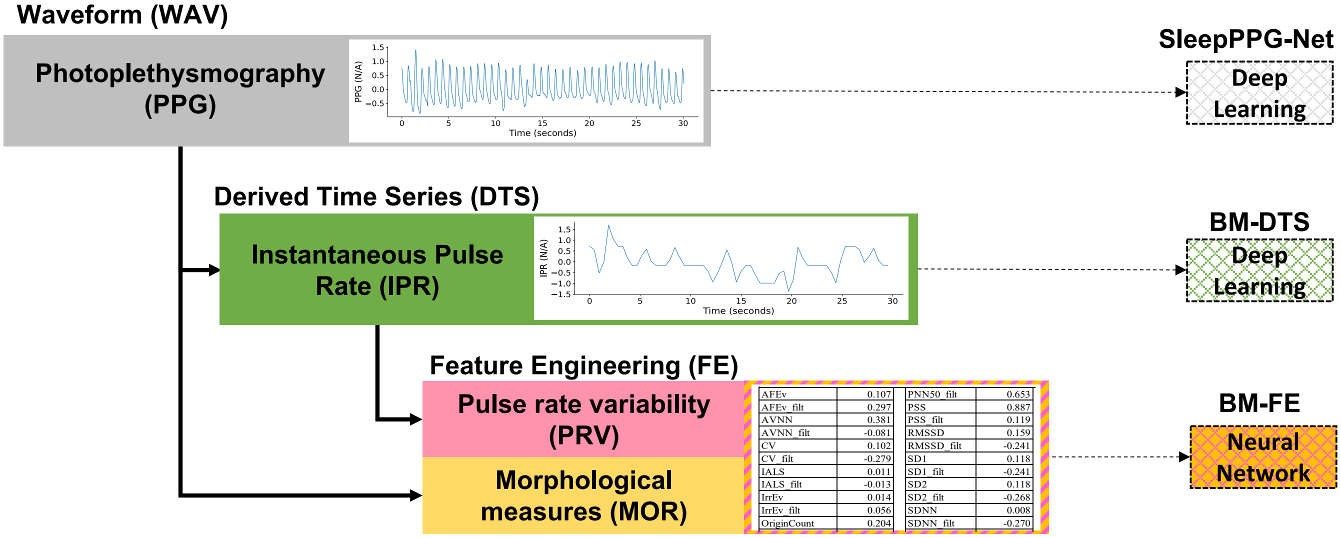

Fig. 1: Three approaches to machine learning (ML) for sleep staging from PPG. Our new model, SleepPPG-Net takes as input the PPG waveform (WAV).The derived time series (DTS) and feature engineering (FE) approaches allow comparison with SOTA algorithms described in the literature.

- We benchmark the performance of SleepPPG-Net against 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 (k 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 (TST), sleep efficiency (SE), sleep- stage fractions and sleep stage transitions.

II. Methods

As depicted in Figure 1, we considered three machine learning (ML) approaches to sleep staging from PPG. 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 Technion-IIT Rappaport Faculty of Medicine institutional review board approval 62-2019. We used three labeled PSG databases in our experiments. SHHS Visit 1 [26], [27], totaling 5,758 unique patients, was used for model pretraining. MESA [27], [28], totaling 2,056 unique patients, was used for training and testing, and the Cleveland Family Study (CFS) Visit-5 v1 [27], [29], 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 [27] and came with sleep stage labels that were manually assigned by sleep experts from the full PSG [26]–[29]. Each PSG was labeled only once [26]–[29]. The MESA PSG study was done in-home whereas the CFS PSG study was done in-laboratory. PPG data for MESA and CFS was acquired from the fingertip using Nonin 8000 series pulse oximeters. PPG sampling rates were 256Hz in MESA, and 128Hz in CFS. SHHS did not contain PPG, so we used ECG from this database to pretrain our models. SHHS ECG sampling rate was 512Hz.All PSG recordings were padded or truncated to a constant length of 10 hours to simplify the design of the DL model. Truncation or zero padding was only applied to the end of each recording. We used all patients in our databases, only removing patients where a valid PPG or ECG waveform could not be loaded. In total, two patients were removed from MESA and four patients were removed from CFS. In the interest of improving reproducibility and comparability, the list of patients assigned to MESA-test is included in Supplement I.

TABLE I: Description of PSG databases used in the experiments. Statistics are presented with their median and interquartile-range (IQR)

	SHHS	MESA	CFS
Patients	5,767	2,054	320
Gender (M:F)	1:1.1	1:1.2	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]

B. Sleep stages

Modern sleep scoring follows guidelines maintained by the American Academy of Sleep Medicine (AASM) [30]. 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 as follows: 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 $8Hz$ and a stop-band attenuation of 40dB. The filtered PPG was downsampled to $34.1\overline{3}Hz$ using linear interpolation, reducing the computational and memory requirements for ML. We choose a sampling rate of $34.1\overline{3}Hz$ as this resulted in 1024 (2^{10}) samples per 30s sleep-window. By using a 2^n number we could maintain full temporal alignment of data with sleep-windows during ML pooling operations. WAV$_{PPG}$ was cleaned by clipping values to three standard deviations and then standardized by subtracting the mean and dividing by the standard deviation.

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 40-100bpm (0.66Hz–1.66Hz) and based on a review of the literature around the optimal filtering of PPGs [31]–[35], we used a band-pass filter with a pass-band of 0.4-8Hz. The filter was built using a zero-phase 8th order band-pass Chebyshev Type II filter with a pass-band of 0.4-8Hz and stop-band attenuation of 40dB. PPG peaks were detected from the filtered time series using an automatic beat detection algorithm developed by Aboy et. al. [36] and implemented in the PulseAnalyse toolbox [37]. This PPG peak detector was chosen because it demonstrated the highest peak detection performance when evaluated on PPGs recorded during PSG [38]. For SHHS, the ECG peaks were detected using epltd0 [39] a SOTA ECG peak detection algorithm.

PRV and HRV measures were extracted using the Python HRV features implemented in [40]. 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 30s.

We standardized MOR and PRV features on a per-patient basis. The mean and standard deviation of each feature for each patient were standardized to 0 and 1 respectively. 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. [20]. The only modification made was that we used a re-sampling rate of $2.1\overline{3}Hz$, as opposed to $2.0Hz$ as this yielded 64 (2^6) samples per 30s sleep-window. 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 $F(S) \mapsto P$. Sleep-windows are indexed with subscript l, where $\{l : 1...L\}$ 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 $F_E(S) \mapsto X$. 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 $X_{l-i}...X_{l+i}$, where i is the receptive-field of F_S, resulting in a richer representation Z_l such that $F_S(X) \mapsto Z$. Finally, F_C computes a probability prediction of each sleep-stage at each sleep-window P_l from Z_l such that $F_E(Z) \mapsto P$.

We further define M as demographic data, n_x as the size of S_l, n_e as the size of X_l, n_z 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, 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. [22]. 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. $n_e = 16$, $n_h = 128$, 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. [20] 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 [41], Wave2Vec [42] and Sridhar et al [20]. SleepPPG-Net architecture is shown in Figure 2. Using $'$ to denote continuous data. The input S' is the continuous WAV$_{PPG}$ time series with $1,228,800$ samples formed by flattening $L = 1200$ and $n_x = 1024$. F_E extracts continuous embeddings X' from S'. F_E consists of 8 stacked ResConvs. Each ResConv contains 3 1D-convolutions followed by max pooling and residual addition. The ResConvs have a kernel of size 3 and max pooling rate of 2. The number of filters in the

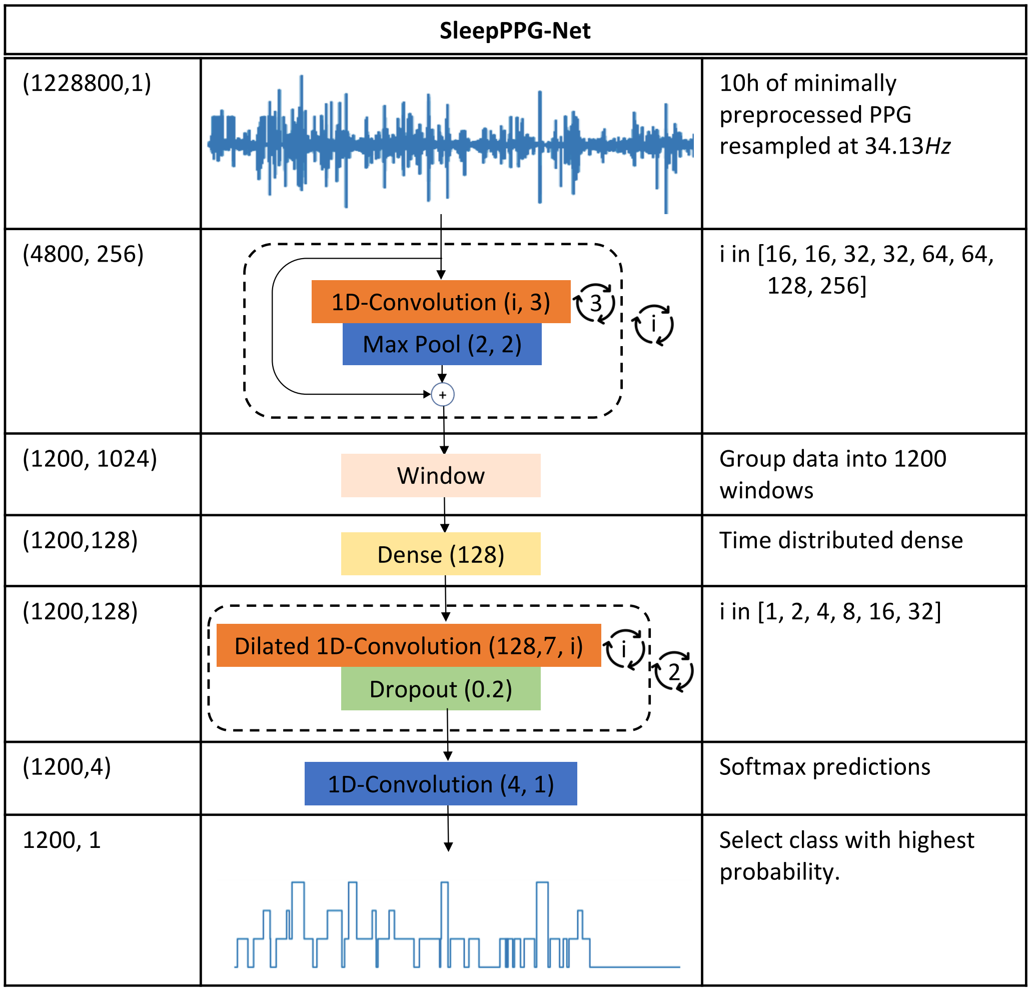
SleepPPG-Net
(1228800,1)
10h of minimally preprocessed PPG resampled at 34.13Hz
(4800, 256)
1D-Convolution (i, 3)
Max Pool (2, 2)
3
i
+
i in [16, 16, 32, 32, 64, 64, 128, 256]
(1200, 1024)
Window
Group data into 1200 windows
(1200,128)
Dense (128)
Time distributed dense
(1200,128)
Dilated 1D-Convolution (128,7, i)
Dropout (0.2)
i
2
i in [1, 2, 4, 8, 16, 32]
(1200,4)
1D-Convolution (4, 1)
Softmax predictions
1200, 1
Select class with highest probability.

Fig. 2: Model architecture of SleepPPG-Net, a novel algorithm for 4-class sleep staging from the raw PPG time series. The input to the network is a preprocessed PPG with a sampling rate of $34.1\overline{3}Hz$. An 8-layer ResConv is used to extract increasingly complex features from the input. A TCN with a wide receptive field adds contextual information to the embeddings. This is followed by classification using a 1D-convolution. The classes with the highest probabilities are used to form the predicted hypnogram.

ResConvs are 16, 16, 32, 32, 64, 64, 128, 256. The max pooling layer in each ResConv reduces temporal dimensionality by a factor of 2, resulting in X' with a dimension of 4800×256. A temporal windowing layer reestablishes X with $L = 1200$ by dividing X' into 1200 windows. At this point, X has a dimension of 1200×1024. A time-distributed DNN then compresses each X_l so that $n_e = 128$. The F_S consists of 2 stacked TCN blocks which add long-range temporal information to X forming Z. Each TCN consists of 5 dilated 1D-convolutions followed by residual addition and dropout. The dilated 1D-convolutions have a kernel size of 7, dilation rates of 1, 2, 4, 8, 16, 32 and filters of size 128 such that $n_z = 128$. Finally, F_C uses a 1D-convolution with kernel of size 1 and 4 filters to make predictions P with a shape of 1200×4. 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 [43]. 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. TL is widely used in medical image analysis, where it has been shown to improve performance when training on small datasets [43]. 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 hyperparameters including; the number of ResConv blocks, ResConv kernel size, ResConv filter size, embedding size, number of TCN blocks, TCN block kernel size and dilation rate, dropout rate, batch size, learning rate and number of training epochs. 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 2.5×10^{-4} and trained for 30 epochs. When training models for the With-pretrain and With-TL training schemes we used a learning rate of 1.0×10^{-4} and trained for 5 epochs. A batch size of 8 was used in all experiments. The remaining hyperparameters used are shown in Figure 2.

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 κ and Ac. We calculate the κ and Ac per patient. The hypnogram labels assigned by sleep experts during PSG scoring are considered to be the ground truth and the probabilities obtained from the model are denoted model predictions. The final reported scores represent the median κ and median Ac, calculated from all patients in the test set. Ac is the observed agreement over all the examples and is calculated according to Equation 1. κ is calculated according to Equation 2 where $Q\equiv Ac$ and Q_e is the chance agreement. L is the total number of samples, C is the number of categories, and n_{cr} is the total number of samples of class c counted by rater r. 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.

$$Ac \equiv Q = \frac{P_{correct}}{L} \tag{1}$$

$$\kappa = \frac{Q-Q_e}{1-Q_e} \quad where \quad Q_e = \frac{1}{L^2}\sum_{c=1}^{C} n_{c1}n_{c2} \tag{2}$$

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.

F. Sleep metrics

Common sleep metrics obtainable from the polysomnography 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 to derive each metric are shown in Equations 3-6. 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).

$$TST = \sum Light + \sum Deep + \sum REM \tag{3}$$

$$SE = \frac{TST}{TST+\sum Wake} \times 100 \tag{4}$$

$$FR_{Stage} = \frac{\sum Stage}{TST} \times 100 \tag{5}$$

$$Transitions = \frac{\sum Transition \quad Between \quad Stages}{Total \quad Recording \quad Time} \tag{6}$$

III. Results

We used four different training schemes to train and evaluate the models as depicted in Figure 3. Training schemes are: No-pretrain, With-pretrain, No-TL and With-TL. The No-pretrain and With-TL models were trained from Xavier initialization, while the With-pretrain and With-TL models were trained with the weights obtained during pretraining.

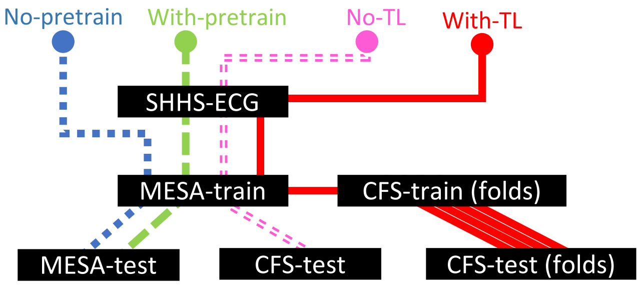

Fig. 3: Chart depicting the four training schemes used in our experiments.

Models are evaluated on a held-out subset of MESA denoted MESA-test, and on 4 non-overlapping folds of the Cleveland Family Study (CFS) denoted CFS-test.

A. Performance measures

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 κ of 0.64 (0.56 to 0.72) and accuracy (Ac) of 76% to a κ of 0.69 (0.62 to 0.77) and Ac of 80%. The best performing model was the pretrained SleepPPG-Net which scored a κ of 0.75 (0.69 to 0.81) and Ac of 84%. The confusion matrix for the pretrained SleepPPG-Net is presented in Figure 4. The κ distribution for the pretrained BM-FE, BM-DTS, and SleepPPG-Net models are compared in Figure 5.

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 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 κ 0.67 (0.55 to 0.74) and Ac of 76%. With TL SleepPPG-Net scored a κ of 0.74 (0.66 to 0.79) and Ac of 82% which is significantly better ($p = 0.0005$, Kolmogorov-Smirnov test). The confusion matrix of SleepPPG-Net with TL is presented in Figure 6. To determine the number of patients needed for effective TL, we evaluate the κ performance on CFS-test as a function of the number of patients used for TL as depicted in Figure 7. Performance for SleepPPG-Net with TL improved from a κ 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 κ performance is presented in Figure 8. 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.

TABLE II: Results for models evaluated on MESA-test (n=204). When performed, pretraining used ECG from SHHS.

	No-pretrain		With-pretrain	
Model	κ	Ac	κ	Ac
BM-FE	0.66	78	0.66	78
BM-DTS	0.64	76	0.69	80
SleepPPG-Net	0.74	83	**0.75**	**84**

TABLE III: Results for models evaluated on CFS-test (n=320). When performed, TL used 4 non-overlapping folds.

	No-TL		With-TL	
Model	κ	Ac	κ	Ac
BM-FE	0.47	63	0.64	76
BM-DTS	0.53	69	0.70	79
SleepPPG-Net	0.67	76	**0.74**	**82**

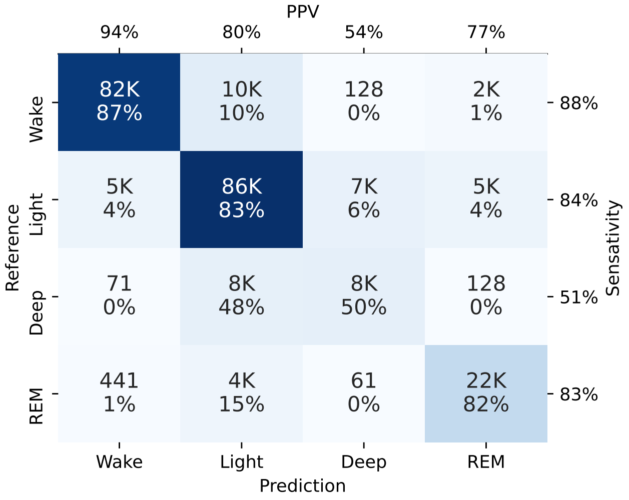

Fig. 4: Confusion matrix for pretrained SleepPPG-Net, evaluated on MESA-test (n=204).

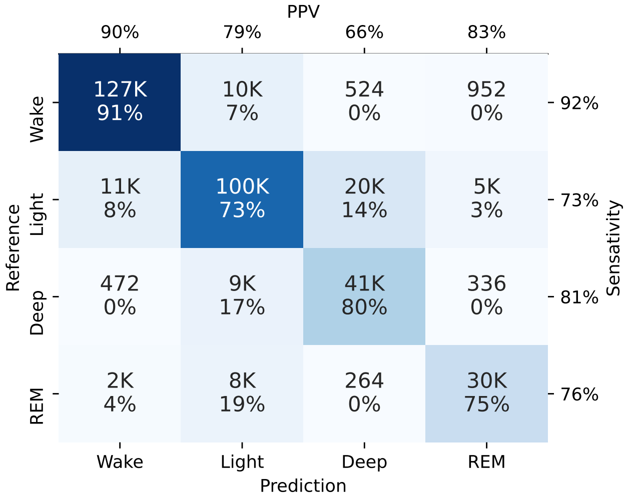

Fig. 6: Confusion matrix for SleepPPG-Net with TL, evaluated on CFS-test (n=320).

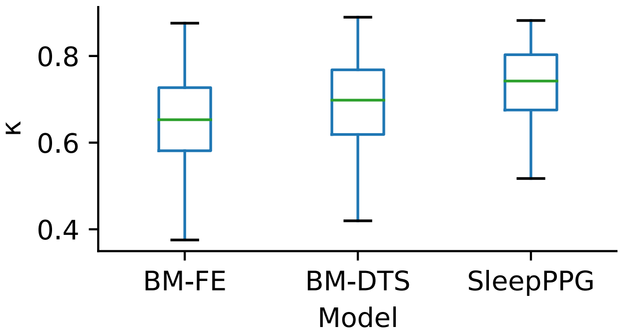

Fig. 5: Distribution of κ performance for BM-FE, BM-DTS and SleepPPG on MESA-test (n=204). All models were pretrained on ECG from SHHS and then trained on MESA-train.

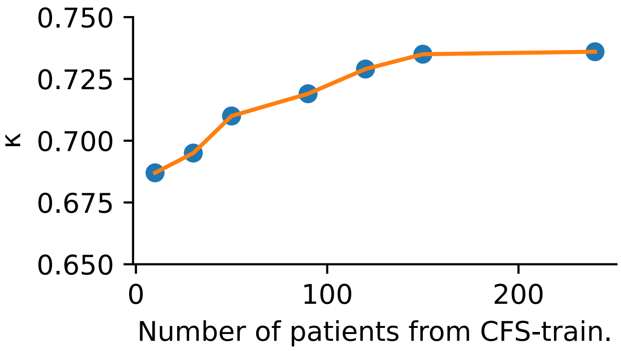

Fig. 7: TL performance on CFS-test by number of patients from CFS-train used for TL. A random subset of patients is taken from each CFS-train fold for training. Experiments were run multiple times and the average is shown.

B. Sleep metrics

We evaluated sleep metrics for MESA-test (n=204) using the pretrained SleepPPG-Net. In Figure 9 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, 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

When interpreting the performance of automated sleep staging algorithms it is important to keep in mind that manual scoring by humans is highly subjective [44]. Inter-rater agreement for 5-class PSG labeled by human scorers is reported as a κ of 0.76 (95% confidence interval, 0.71–0.81) [45]. Common mistakes between human scorers during PSG include confusion between wake and light sleep and light sleep and deep sleep [46]. 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 ($p < 0.0004$, 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 κ of 0.75 against 0.66 for BM-FE and 0.69 for BM-DTS approaches. SleepPPG-Net performance is also significantly ($p < 0.001$, two-sample t-test) higher than the current published SOTA results for sleep staging from PPG which stand at a κ of 0.66 [22], [23], and significantly ($p = 0.02$, two-sample t-test) higher than the current SOTA results for sleep-sleep staging from ECG which are reported at κ of 0.69 [20]. While reporting performances from other studies, it is important to keep in mind that these may be for a different cohorts/different

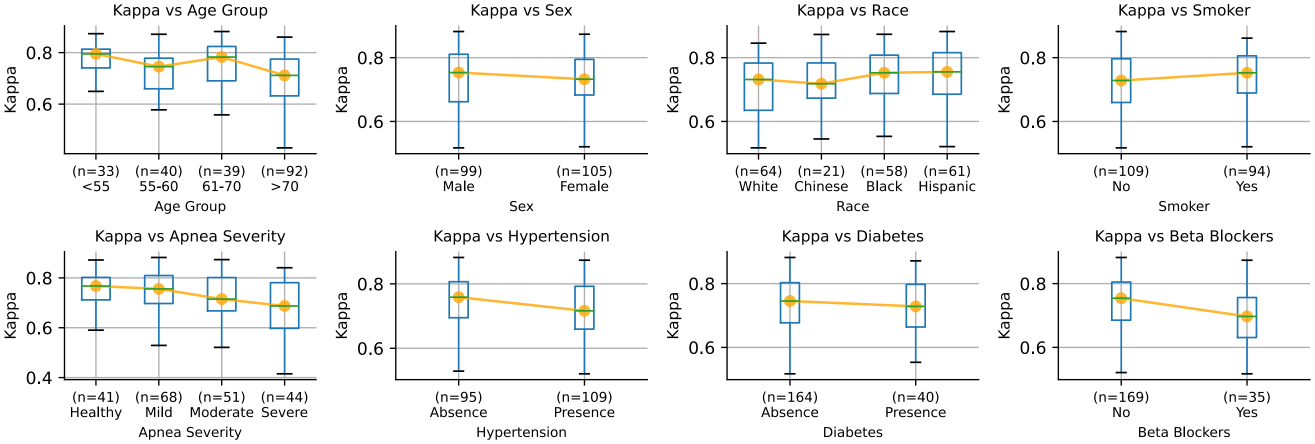

Fig. 8: SleepPPG-Net With-pretrain performance per clinical group evaluated on MESA-test (n=204).

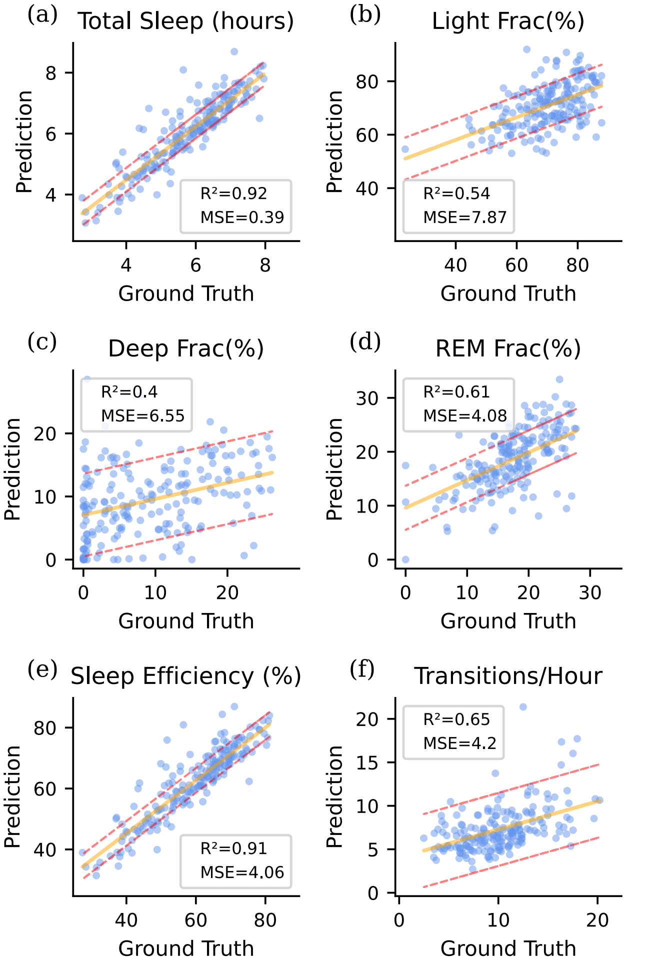

Fig. 9: Comparison between sleep metrics of MESA-test (n=204) calculated from pretrained SleepPPG-Net to those calculated using ground truth. Dotted line shows the MSE.

train-test split and thus a direct comparison may not possible.

Figure 10 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 [47], [48]. 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. Similar performance was maintained for all clinical groups with no important outliers. The small differences that are observed between groups are likely due to the model's inability to accurately distinguish between light and deep sleep and detect short awakenings.

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 κ of 0.67 (0.55 to 0.74) on CFS-test with no TL step. This is markedly higher

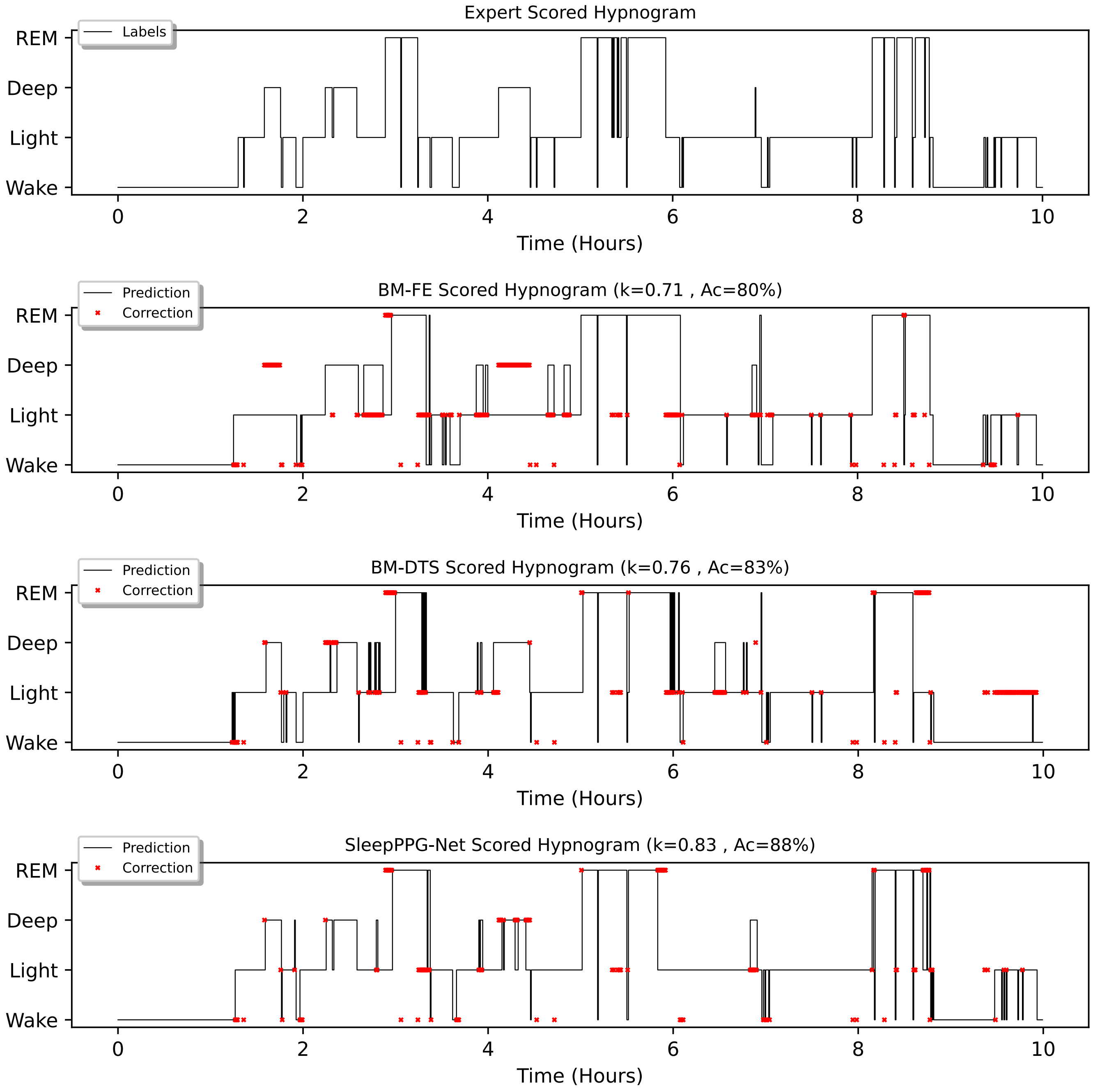

Fig. 10: Hypnogram showing discrepancies between sleep stage predictions and the hypnogram labels assigned by human scorers during PSG. Hypnogram shown is from MESA patient (ID-0310) who is in MESA-test.

than the generalization performance reported by Sridhar et al. [20], whose model scored a κ of 0.55 on the PhysioNet/Computing in Cardiology Sleep database [21]. With TL, performance of SleepPPG-Net increased significantly reaching a κ 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 κ 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 pretraining 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 [49]. However, for disorders such as night terrors or sleepwalking and it is important to distinguish between light and deep sleep [49]. 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 (Figure 8). 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 [50] 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 beating 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 may also be experimented in order to improve performance. In particular, models developed for sleep staging from the electroencephalography (EEG) were developed over the past few years and showed high performance. These include the work of Phan et al. [51], with SOTA models such as SalientSleepNet [52] and SleepTransformer [53] achieving k 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 [54], may be valuable in order to provide the human interpreter with some reliability measures on the produced hypnogram.

V. Conclusion

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 that are accurate enough for clinical diagnosis. This will allow for improved detection, monitoring, and treatment of sleep disorders in the general population.

References

[1] M. Walker, Why we sleep. London: Penguin Books, 2018.

[2] A. V. Benjafield, N. T. Ayas, P. R. Eastwood, R. Heinzer, M. S. Ip, M. J. Morrell, C. M. Nunez, S. R. Patel, T. Penzel, J. L. D. Pépin, P. E. Peppard, S. Sinha, S. Tufik, K. Valentine, and A. Malhotra, "Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis," The Lancet Respiratory Medicine, vol. 7, no. 8, pp. 687–698, 8 2019.

[3] V. K. Kapur, D. H. Auckley, S. Chowdhuri, D. C. Kuhlmann, R. Mehra, K. Ramar, and C. G. Harrod, "Clinical Practice Guideline for Diagnostic Testing for Adult Obstructive Sleep Apnea: An American Academy of Sleep Medicine Clinical Practice Guideline," Journal of Clinical Sleep Medicine, vol. 13, no. 3, pp. 479–504, 2017.

[4] B. Sriram, V. Singh, S. Bandaralage, and J. Bashford, "P135 An audit investigating the length of time from GP referral to diagnostic polysomnography testing in an Australian tertiary center," SLEEP Advances, vol. 2, no. Supplement_1, pp. A65–A65, 10 2021.

[5] C. Q. Phua, I. J. Jang, K. B. Tan, Y. Hao, S. R. B. Senin, P. R. Song, R. Y. Soh, and S. T. Toh, "Reducing cost and time to diagnosis and treatment of obstructive sleep apnea using ambulatory sleep study: a Singapore sleep centre experience," Sleep and Breathing, vol. 25, no. 1, pp. 281–288, 3 2021.

[6] J. Behar, A. Roebuck, M. Shahid, J. Daly, A. Hallack, N. Palmius, J. Stradling, and G. D. Clifford, "SleepAp: An Automated Obstructive Sleep Apnoea Screening Application for Smartphones," IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 1, pp. 325–331, 1 2015.

[7] X. Tan, J. D. Cook, J. Cedernaes, and C. Benedict, "Consumer sleep trackers: a new tool to fight the hidden epidemic of obstructive sleep apnoea?" The Lancet Respiratory Medicine, vol. 7, no. 12, p. 1012, 12 2019.

[8] J. A. Behar, A. A. Rosenberg, I. Weiser-Bitoun, O. Shemla, A. Alexandrovich, E. Konyukhov, and Y. Yaniv, "PhysioZoo: a novel open access platform for heart rate variability analysis of mammalian electrocardiographic data," Frontiers in physiology, vol. 9, p. 1390, 2018.

[9] J. A. Behar, N. Palmius, Q. Li, S. Garbuio, F. P. Rizzatti, L. Bittencourt, S. Tufik, and G. D. Clifford, "Feasibility of Single Channel Oximetry for Mass Screening of Obstructive Sleep Apnea," EClinicalMedicine, vol. 11, pp. 81–88, 5 2019.

[10] S. A. Imtiaz, "A Systematic Review of Sensing Technologies for Wearable Sleep Staging," Sensors 2021, Vol. 21, Page 1562, vol. 21, no. 5, p. 1562, 2 2021.

[11] F. Moreno-Pino, A. Porras-Segovia, P. López-Esteban, A. Artés, and E. Baca-García, "Validation of Fitbit Charge 2 and Fitbit Alta HR Against Polysomnography for Assessing Sleep in Adults With Obstructive Sleep Apnea," Journal of Clinical Sleep Medicine, vol. 15, no. 11, pp. 1645–1653, 11 2019.

[12] E. D. Chinoy, J. A. Cuellar, K. E. Huwa, J. T. Jameson, C. H. Watson, S. C. Bessman, D. A. Hirsch, A. D. Cooper, S. P. Drummond, and R. R. Markwald, "Performance of seven consumer sleep-tracking devices compared with polysomnography," Sleep, vol. 44, no. 5, 5 2021.

[13] A. M. Fink, U. G. Bronas, and M. W. Calik, "Autonomic regulation during sleep and wakefulness: a review with implications for defining the pathophysiology of neurological disorders," Clinical autonomic research : official journal of the Clinical Autonomic Research Society, vol. 28, no. 6, p. 509, 12 2018.

[14] R. Cabiddu, S. Cerutti, G. Viardot, S. Werner, and A. M. Bianchi, "Modulation of the sympatho-vagal balance during sleep: Frequency domain study of heart rate variability and respiration," Frontiers in Physiology, vol. 3 MAR, p. 45, 2012.

[15] F. Ebrahimi and I. Alizadeh, "Automatic sleep staging by cardiorespiratory signals: a systematic review," Sleep and Breathing, 2021.

[16] "Validation of Photoplethysmography-Based Sleep Staging Compared With Polysomnography in Healthy Middle-Aged Adults," Sleep, vol. 40, no. 7, 7 2017.

[17] R. Wei, X. Zhang, J. Wang, and X. Dang, "The research of sleep staging based on single-lead electrocardiogram and deep neural network," Biomedical Engineering Letters, vol. 8, no. 1, p. 87, 2 2018.

[18] H. Sun, W. Ganglberger, E. Panneerselvam, M. J. Leone, S. A. Quadri, B. Goparaju, R. A. Tesh, O. Akeju, R. J. Thomas, and M. B. Westover, "Sleep staging from electrocardiography and respiration with deep learning," Sleep, vol. 43, no. 7, 2020.

[19] P. Fonseca, M. M. van Gilst, M. Radha, M. Ross, A. Moreau, A. Cerny, P. Anderer, X. Long, J. P. van Dijk, and S. Overeem, "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.

[20] N. Sridhar, A. Shoeb, P. Stephens, A. Kharbouch, D. B. Shimol, J. Burkart, A. Ghoreyshi, and L. Myers, "Deep learning for automated sleep staging using instantaneous heart rate," npj Digital Medicine, vol. 3, no. 1, 2020.

[21] M. M. Ghassemi, B. E. Moody, L.-W. H. Lehman, C. Song, Q. Li, H. Sun, R. G. Mark, M. B. Westover, and G. D. Clifford, "You Snooze, You Win: the PhysioNet/Computing in Cardiology Challenge 2018."

[22] M. Radha, P. Fonseca, A. Moreau, M. Ross, A. Cerny, P. Anderer, X. Long, and R. M. Aarts, "A deep transfer learning approach for wearable sleep stage classification with photoplethysmography," npj Digital Medicine, vol. 4, no. 1, p. 135, 12 2021.

[23] B. M. Wulterkens, P. Fonseca, L. W. A. Hermans, M. Ross, A. Cerny, P. Anderer, X. Long, J. P. v. Dijk, N. Vandenbussche, S. Pillen, M. M. v. Gilst, and S. Overeem, "It is All in the Wrist: Wearable Sleep Staging in a Clinical Population versus Reference Polysomnography," Nature and Science of Sleep, vol. 13, p. 885, 2021.

[24] H. Korkalainen, J. Aakko, B. Duce, S. Kainulainen, A. Leino, S. Nikkonen, I. O. Afara, S. Myllymaa, J. Töyräs, and T. Leppänen, "Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea," Sleep, no. May, pp. 1–10, 5 2020.

[25] R. Huttunen, T. Leppänen, B. Duce, A. Oksenberg, S. Myllymaa, J. Töyräs, and H. Korkalainen, "Assessment of obstructive sleep apnea-

related sleep fragmentation utilizing deep learning-based sleep staging from photoplethysmography," Sleep, no. June, pp. 1–10, 2021.

[26] S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, J. M. Samet, and P. W. Wahl, "The Sleep Heart Health Study: design, rationale, and methods." Sleep, vol. 20, no. 12, pp. 1077–85, 12 1997.

[27] G.-Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, and S. Redline, "The National Sleep Research Resource: towards a sleep data commons," Journal of the American Medical Informatics Association, vol. 25, no. 10, pp. 1351–1358, 10 2018.

[28] X. Chen, R. Wang, P. Zee, P. L. Lutsey, S. Javaheri, C. Alcántara, C. L. Jackson, M. A. Williams, and S. Redline, "Racial/Ethnic Differences in Sleep Disturbances: The Multi-Ethnic Study of Atherosclerosis (MESA)," SLEEP, 6 2015.

[29] S. Redline, P. V. Tishler, T. D. Tosteson, J. Williamson, K. Kump, I. Browner, V. Ferrette, and P. Krejci, "The Familial Aggregation of Obstructive Sleep Apnea," American Journal of Respiratory and Critical Care Medicine, vol. 151, no. 3_pt_1, pp. 682–687, 3 1995.

[30] R. B. Berry, R. Brooks, C. Gamaldo, S. M. Harding, R. M. Lloyd, S. F. Quan, M. T. Troester, and B. V. Vaughn, "AASM Scoring Manual Updates for 2017 (Version 2.4)," Journal of Clinical Sleep Medicine : JCSM : Official Publication of the American Academy of Sleep Medicine, vol. 13, no. 5, p. 665, 2017.

[31] S. Hara, T. Shimazaki, H. Okuhata, H. Nakamura, T. Kawabata, K. Cai, and T. Takubo, "Parameter optimization of motion artifact canceling PPG-based heart rate sensor by means of cross validation," in 2017 11th International Symposium on Medical Information and Communication Technology (ISMICT). IEEE, 2 2017, pp. 73–76.

[32] C. Park, H. Shin, and B. Lee, "Blockwise PPG Enhancement Based on Time-Variant Zero-Phase Harmonic Notch Filtering," Sensors, vol. 17, no. 4, p. 860, 4 2017.

[33] D. Pollreisz and N. TaheriNejad, "Detection and Removal of Motion Artifacts in PPG Signals," Mobile Networks and Applications, pp. 1–11, 8 2019.

[34] K. T. Tanweer, S. R. Hasan, and A. M. Kamboh, "Motion artifact reduction from PPG signals during intense exercise using filtered X-LMS," in 2017 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 5 2017, pp. 1–4.

[35] A. Temko, "Accurate Heart Rate Monitoring during Physical Exercises Using PPG," IEEE Transactions on Biomedical Engineering, vol. 64, no. 9, pp. 2016–2024, 9 2017.

[36] M. Aboy, J. McNames, T. Thong, D. Tsunami, M. Ellenby, and B. Goldstein, "An Automatic Beat Detection Algorithm for Pressure Signals," IEEE Transactions on Biomedical Engineering, vol. 52, no. 10, pp. 1662–1670, 10 2005.

[37] 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," American Journal of Physiology-Heart and Circulatory Physiology, vol. 317, no. 5, pp. H1062–H1085, 11 2019.

[38] K. Kotzen, P. H. Charlton, A. Landesberg, and J. A. Behar, "Benchmarking Photoplethysmography Peak Detection Algorithms Using the Electrocardiogram Signal as a Reference," 2021 Computing in Cardiology (CinC), pp. 1–4, 9 2021.

[39] P. Hamilton, "Open source ECG analysis," in Computers in Cardiology. IEEE, 9 2002, pp. 101–104.

[40] A. Chocron, R. Efraim, F. Mandel, M. Rueschman, N. Palmius, T. Penzel, M. Elbaz, and J. A. Behar, "Digital biomarkers and artificial intelligence for mass diagnosis of atrial fibrillation in a population sample at risk of sleep disordered breathing," Physiological Measurement, vol. 41, no. 10, 7 2020.

[41] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," 9 2016.

[42] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-training for Speech Recognition," 4 2019.

[43] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding Transfer Learning for Medical Imaging," Advances in Neural Information Processing Systems, vol. 32, 2 2019.

[44] H. Phan and K. Mikkelsen, "Automatic Sleep Staging: Recent Development, Challenges, and Future Directions," pp. 1–28, 2021.

[45] Y. J. Lee, J. Y. Lee, J. H. Cho, and J. H. Choi, "Inter-rater reliability of sleep stage scoring: a meta-analysis," Journal of Clinical Sleep Medicine, 7 2021.

[46] H. Danker-Hopfe, P. Anderer, J. Zeitlhofer, M. Boeck, H. Dorn, G. Gruber, E. Heller, E. Loretz, D. Moser, S. Parapatics, B. Saletu, A. Schmidt, and G. Dorffner, "Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard." Journal of sleep research, vol. 18, no. 1, pp. 74–84, 3 2009.

[47] M. Elgendi, "On the analysis of fingertip photoplethysmogram signals," Current cardiology reviews, vol. 8, no. 1, pp. 14–25, 2012.

[48] K. Li, W. Pan, Q. Jiang, and G.-Z. Liu, "A Method to Detect Sleep Apnea based on Deep Neural Network and Hidden Markov Model using Single-Lead ECG signal," Neurocomputing, vol. 294, 3 2018.

[49] T. L. Lee-Chiong, Sleep: A Comprehensive Handbook, L.-C. Teofilo L, Ed. John Wiley & Sons, 2005.

[50] M. A. Carskadon, E. D. Brown, and W. C. Dement, "Sleep fragmentation in the elderly: Relationship to daytime sleep tendency," Neurobiology of Aging, vol. 3, pp. 321–327, 12 1982.

[51] "Automatic sleep staging of eeg signals: recent development, challenges, and future directions," Physiological measurement, vol. 43, 4 2022. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/35320788/

[52] Z. Jia, Y. Lin, J. Wang, X. Wang, P. Xie, and Y. Zhang, "Salientsleepnet: Multimodal salient wave detection network for sleep staging," CoRR, vol. abs/2105.13864, 2021. [Online]. Available: https://arxiv.org/abs/2105.13864

[53] 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 Transactions on Biomedical Engineering, 5 2021. [Online]. Available: http://arxiv.org/abs/2105.11043 http://dx.doi.org/10.1109/TBME.2022.3147187

[54] Y. Gal and Z. A. Uk, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning Zoubin Ghahramani," 2016.

Acknowledgment

The Multi-Ethnic Study of Atherosclerosis (MESA) Sleep Ancillary study was funded by NIH-NHLBI Association of Sleep Disorders with Cardiovascular Health Across Ethnic Groups (RO1 HL098433). MESA is supported by NHLBI funded contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168 and N01-HC-95169 from the National Heart, Lung, and Blood Institute, and by cooperative agreements UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420 funded by NCATS. The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

The Cleveland Family Study (CFS) database is only available for non-commercial use. The Cleveland Family Study (CFS) was supported by grants from the National Institutes of Health (HL46380, M01 RR00080-39, T32-HL07567, RO1-46380). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve University). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

Appendix I
List of patients in MESA-test

0001, 0021, 0033, 0052, 0077, 0081, 0101, 0111, 0225, 0310, 0314, 0402, 0416, 0445, 0465, 0483, 0505, 0554, 0572, 0587, 0601, 0620, 0648, 0702, 0764, 0771, 0792, 0797, 0800, 0807, 0860, 0892, 0902, 0904, 0921, 1033, 1080, 1121, 1140, 1148, 1161, 1164, 1219, 1224, 1271, 1324, 1356, 1391, 1463, 1483, 1497, 1528, 1531, 1539, 1672, 1693, 1704, 1874, 1876, 1900, 1914, 2039, 2049, 2096, 2100, 2109, 2169, 2172, 2183, 2208, 2239, 2243, 2260, 2269, 2317, 2362, 2388, 2470, 2472, 2488, 2527, 2556, 2602, 2608, 2613, 2677, 2680, 2685, 2727, 2729, 2802, 2811, 2828, 2877, 2881, 2932, 2934, 2993, 2999, 3044, 3066, 3068, 3111, 3121, 3153, 3275, 3298, 3324, 3369, 3492, 3543, 3554, 3557, 3561, 3684, 3689, 3777, 3793, 3801, 3815, 3839, 3886, 3997, 4110, 4137, 4171, 4227, 4285, 4332, 4406, 4460, 4462, 4497, 4501, 4552, 4577, 4649, 4650, 4667, 4732, 4794, 4888, 4892, 4895, 4912, 4918, 4998, 5006, 5075, 5077, 5148, 5169, 5203, 5232, 5243, 5287, 5316, 5357, 5366, 5395, 5397, 5457, 5472, 5479, 5496, 5532, 5568, 5580, 5659, 5692, 5706, 5737, 5754, 5805, 5838, 5847, 5890, 5909, 5957, 5983, 6015, 6039, 6047, 6123, 6224, 6263, 6266, 6281, 6291, 6482, 6491, 6502, 6516, 6566, 6567, 6583, 6619, 6629, 6646, 6680, 6722, 6730, 6741, 6788

Appendix II
FE Model Diagrams

TABLE IV: Description of the network used to perform Sleep Staging using Feature Engineering.

Shape	Layer	Description
$1200 \times n_x$	Input	n_x is the number of features
1200×512	Dense	
1200×256	Dense	
1200×128	Dense	
1200×16	Dense	
$1200 \times n_e$	Dense	n_e is the embedding size
$1200 \times 2n_h$	LSTM	n_h is the number of hidden units
$1200 \times 2n_h$	LSTM	n_h is the number of hidden units
	Dropout	
1200×256	Dense	
	Dropout	
1200×128	Dense	
	Dropout	
1200×64	Dense	
1200×4	Dense	Softmax Activation

Appendix III
DTS Model Diagrams

TABLE V: Description of the network used to perform Sleep Staging using IPR time series.

Shape	Layer			Description
$1200 \times 256 \times 1$	Input			10 hours 2Hz
$1200 \times 256 \times F$	1D-Conv	$\| - \times 2$	$\|$	$i = [1, 2, 3]$
$1200 \times 128/i \times F$	MaxPool		$\|$ - X 3	F = [16, 32, 64]
$1200 \times 128/i \times F$	Residual		$\|$	
1200×2048	Flatten			
1200×128	Dense			
$1 \times 1200 \times 128$	Reshape			
$1 \times 1200 \times 128$	Dilated 1D-Conv	$\|$ -X 6	$\|$	kernel = 7
$1 \times 1200 \times 128$	Residual		$\|$ - X 2	dilation = [1,2,4,8,16,32]
$1 \times 1200 \times 128$	Dropout		$\|$	
$1 \times 1200 \times 4$	1D-Conv			Softmax Activation