A flexible deep learning architecture for temporal sleep stage classification using accelerometry and photoplethysmography

Olsen, Mads; Zeitzer, Jamie M.; Richardson, Risa N. ; Davidenko, Polina; Jennum, Poul J.; Sorensen, Helge B.D.; Mignot, Emmanuel

Published in:
I E E E Transactions on Biomedical Engineering

Link to article, DOI:
10.1109/TBME.2022.3187945

Publication date:
2023

Document Version
Peer reviewed version

TBME-02270-2021.R2

A flexible deep learning architecture for temporal sleep stage classification using accelerometry and photoplethysmography

Mads Olsen IEEE member, Jamie M. Zeitzer, Risa N. Richardson, Polina Davidenko, Poul J. Jennum*, Helge B. D. Sørensen* IEEE Senior member, Emmanuel Mignot*

***Abstract*— Wrist-worn consumer sleep technologies (CST) that contain accelerometers (ACC) and photoplethysmography (PPG) are increasingly common and hold great potential to function as out-of-clinic (OOC) sleep monitoring systems. However, very few validation studies exist because raw data from CSTs are rarely made accessible for external use. We present a deep neural network (DNN) with a strong temporal core, inspired by U-Net, that can process multivariate time series inputs with different dimensionality to predict sleep stages (wake, light-, deep-, and REM sleep) using ACC and PPG signals from nocturnal recordings. The DNN was trained and tested on 3 internal datasets, comprising raw data both from clinical and wrist-worn devices from 301 recordings (PSG-PPG: 266, Wrist-worn PPG: 35). External validation was performed on a hold-out test dataset containing 35 recordings comprising only raw data from a wrist-worn CST. An accuracy=0.71±0.09, 0.76±0.07, 0.73±0.06, and κ=0.58±0.13, 0.64±0.09, 0.59±0.09 was achieved on the internal test sets. Our experiments show that spectral preprocessing yields superior performance when compared to surrogate-, feature-, raw data-based preparation. Combining both modalities produce the overall best performance, although PPG proved to be the most impactful and was the only modality capable of detecting REM sleep well. Including ACC improved model precision to wake and sleep metric estimation. Increasing input segment size improved performance consistently; the best performance was achieved using 1024 epochs (~8.5 hrs.). An accuracy=0.69±0.13 and κ=0.58±0.18 was achieved on the hold-out test dataset, proving the generalizability and robustness of our approach to raw data collected with a wrist-worn CST.**

***Index Terms*— mHealth, deep learning, wrist actigraphy, sleep stage classification, consumer sleep technologies.**

I. Introduction

Wrist-worn consumer sleep technologies (CST) that use accelerometers (ACC) and photo-plethysmography (PPG) to estimate sleep are increasingly common and hold great potential to function as inexpensive and convenient out-of-clinic (OOC) sleep monitoring systems [1]. Unfortunately, data from CSTs often rely on proprietary algorithms and raw data are rarely accessible for external use. Hence, there is an unmet need to validate raw sensor data from wrist-worn CSTs against gold-standard polysomnography (PSG) recordings [2].

Wrist-worn ACC, commonly known as actigraphy, measures physical activity and has been used to identify rest/activity in ambulatory settings as an alternative to PSG. Using actigraphy, sleep is simply defined as the absence of physical activity and is traditionally identified using threshold-based algorithms such as the Cole-Kripke algorithm [3]. Actigraphy based algorithms perform reasonably well to detect sleep but the fundamental limitation is that the coarse sampling rate of most actigraphy devices makes it difficult to capture motionless wakefulness; consequently such algorithms often overestimate sleep [4]. Recent studies have shown that wrist actigraphy sampled with high resolution captures motion with much finer detail and enables identification of breathing [5] and even heart rate [6]. Therefore, high resolution actigraphy has the potential to improve the detection of motionless wakefulness and may even enable classification of sleep stages. Sundararajan et al. recently presented a random forest classifier, utilizing 36 features extracted from nocturnal high resolution, triaxial ACC recordings, that significantly improved sleep-wake detection when compared to the Cole-Kripke algorithm [7]. However, they concluded that complete sleep stage classification was challenging due to the absence of discriminative features.

PPG measures the peripheral pulse wave, and through identification of inter-beat-intervals enables analysis of pulse rate variability (PRV), a comparable surrogate of heart rate variability that maps changes in the autonomic nervous system [8]. These autonomic changes have proven discriminative in

submission date: December 9th, 2021. Funding of STAGES was provided by the Klarman Family foundation. Funding of the TBI group was provided by a Patient-Centered Outcomes Research Institute (PCORI) Award (CER-1511-33 005). Amazfit, a Zepp Inc. brand, funded the devices for this study by a contract to Emmanuel Mignot registered at clinical.gov as NCT04429906.

Mads Olsen, Jamie M. Zeitzer, Polina Davidenko and Emmanuel J. Mignot are with the Department of Psychiatry and Behavioral Sciences, Stanford University, CA 94304 USA (e-mail: mads_olsen123@hotmail.com; jzeitzer@stanford.edu; mignot@stanford.edu).

Mads Olsen, Helge B. D. Sorensen is with the Biomedical Signal Processing & AI Research Group, Department of Health Technology, Technical University of Denmark, 2800 Kongens Lyngby, Denmark (e-mail: hbds@dtu.dk).

Poul Jennum is with the Danish Center for Sleep Medicine, Department of Clinical Neurophysiology, Rigshospitalet, 2600 Glostrup, Denmark (e-mail: poul.joergen.jennum@regionh.dk).

Risa N. Richardson is with the Division of Pulmonary and Sleep Medicine, Morsani College of Medicine, University of South Florida, Tampa, Florida USA (e-mail: risan@usf.edu).

*Shared last author

sleep research for the detection of sleep stages [9], [10], cortical arousals [11], [12], and sleep apnea [13]. Common for most classification algorithms that use pulse or heart rate signals is that they build on a sequence of preprocessing steps to extract the HRV/PRV signal, e.g., pulse detection, ectopic beat detection, pulse interpolation, and in turn require preparation and extraction of hand-crafted features [14]–[21].

In general, feature-based approaches carry the risk of cumulating errors and are constricted to engineered features that are not guaranteed to be optimal for the classification task. Contrarily, deep learning (DL)-based algorithms are data-driven systems that can learn feature representations directly from raw data and thus do have these restrictions. Korkalainen et al. recently reported promising sleep stage classification performance using a deep neural network (DNN) trained on raw PSG-based PPG signals [22]. However, DL based systems usually require large amounts of diverse data to generalize well [23]. This constitutes a problem since big datasets with diverse data from wrist-worn consumer devices do not exist, yet. Furthermore, data from wrist-worn devices are more prone to noise and data loss, and signal quality varies considerably between devices when compared to data from clinical equipment collected in a controlled environment [2]. Hence, until large datasets with diverse data exist, there is an unmet need to evaluate the tradeoffs of different preprocessing frameworks on signals from wrist-worn devices.

Sleep is segmented into 30 s *epochs* and is scored into 5 distinct classes, namely: wake (W), N1, N2, N3, and rapid eye movement (REM) sleep, according to scoring guidelines provided by standardization organizations, e.g., American Academy of Sleep Medicine (AASM)[24]. N1, N2, and N3 constitute non-REM (NREM) sleep, and can be subdivided into light sleep (N1 and N2) and deep sleep (N3). Sleep is a dynamic process with a cyclic pattern that cycles through NREM and REM sleep with a period of approximately 90-110 minutes. The most recent and best performing sleep stage classification algorithms are temporal models that are either based on recurrent frameworks, e.g., long-short term memory (LSTM) [9], [19], [25] and gated recurrent units (GRU) [22], [26] or convolutional neural network (CNN) architectures, e.g., dilated convolutions [27] and the residual U-Net architecture [28]. While there is consensus that including contextual information from neighboring epochs increase performance, the segment size that these temporal models are trained on vary considerably between studies; from minutes [26], [28], to hours[22], and to the entire recording length [27]. An optimal input segment size for sleep stage classification remains to be established.

In this paper, we present a flexible DNN with a strong temporal core, inspired by U-Net [28]–[30], to capture long-term dependencies. The model was trained to classify sleep stages (wake, light sleep, deep sleep, and REM sleep) using PPG and ACC recordings. Using this model, we investigated the impact of preprocessing across different datasets. We compare performance of different modality combinations to assess the modality importance, both in terms of overall performance, and with respect to sleep metrics, and related the performance to State-of-the-art (SOTA) works. Finally, we evaluated the importance of input segment size. A preliminary version of this work has been reported [31].

II. Material and Methods

A conceptual visualization of the proposed DNN is presented in Fig. 1. It consists of three key modules. Firstly, a conformation module serves to prepare the input segments to conform with the subsequent temporal module. Secondly, the prepared segments are processed into feature maps in a temporal module inspired by U-Net [29], modified to operate on time series inputs instead of images, similarly to DeepSleep [30] and U-Sleep [28], which serves to enhance the propagation of temporal information across the entire input segment. Lastly, a segment classifier module, inspired by U-Sleep [28], serves to segment and classify the feature maps into sleep stage vectors that constitute the output predictions of the model.

A. Deep neural network architecture

The complete architecture of the proposed DNN will be presented in the following. Please refer to the supplementary material for a layer-based version of the model.

Let $\mathbf{X}^{(s)} \in \mathbb{R}^{T_s \times F_s \times C_s}$, $s = \{1, \dots, S\}$ denote the s-th time series segment where T_s, F_s, and C_s are the temporal, spatial, and channel dimensions, respectively. The conformation module, $\varphi_Z: \mathbf{X}^{(s)} \to \widetilde{\mathbf{X}} \in \mathbb{R}^{\tilde{T}_s \times \tilde{F}_s \times C}$, where $C = \sum_s C_s$ and $\tilde{T}_s = 2^{log_2(\lceil T_s \rceil)}$ and $\tilde{F}_s = 2^{log_2(\lceil F_s \rceil)}$, performs concatenation, reshaping, and zero-padding of the input. $\lceil \cdot \rceil$ denotes the ceil operator. Input segments are concatenated along the channel axis, thus their spatial-temporal dimensions must match. Then they are reshaped into a 3D vector and finally zero-padded to ensure that the spatial-temporal dimensions are a power of 2, such that output dimensions remain integers during up- and down-sampling throughout the temporal module.

The temporal module $\varphi_U: \widetilde{\mathbf{X}} \to \mathbb{R}^{\tilde{T}_s \times \tilde{F}_s \times C_U}$ consists of an encoder φ_E and a decoder φ_D, each of which consists of M blocks. This module can adapt to both one- and two-dimensional inputs by changing kernel size and stride of the convolutional layers, as denoted in the following by (2D)/(1D).

The encoder, $\varphi_E: \widetilde{\mathbf{X}} \to \mathbb{R}^{\tilde{T}_s/2^M, \lceil \tilde{F}_s/2^M \rceil, 2^{M/3} C_U}$ serves to learn feature representations from the segment at different scales by reducing the spatial-temporal resolution and increasing the feature dimension incrementally. Each block m in the encoder φ_E, has the same, simple composition that consists of a 2D convolutional layer with $2^{(m-1)/3} C_U$ filters of size $(K, 3)/(K, 1)$, a Gaussian Error Linear Unit (GELU) activation function [32], a batch-normalization normalization layer [33], and a second 2D convolutional layer with $2^{m/3} C_U$ filters of shape $(2, 2)/(2,1)$ and with stride: $(2, 2)/(2,1)$. Thus, the number of feature maps increase with a factor of $\sqrt[3]{2}$, and the spatial-temporal resolution is reduced by a factor of 2 with each block m. C_U and K were found by experimentation. Finally, the output from φ_E is processed by a similar block without spatial-temporal reduction before it is processed by the decoder.

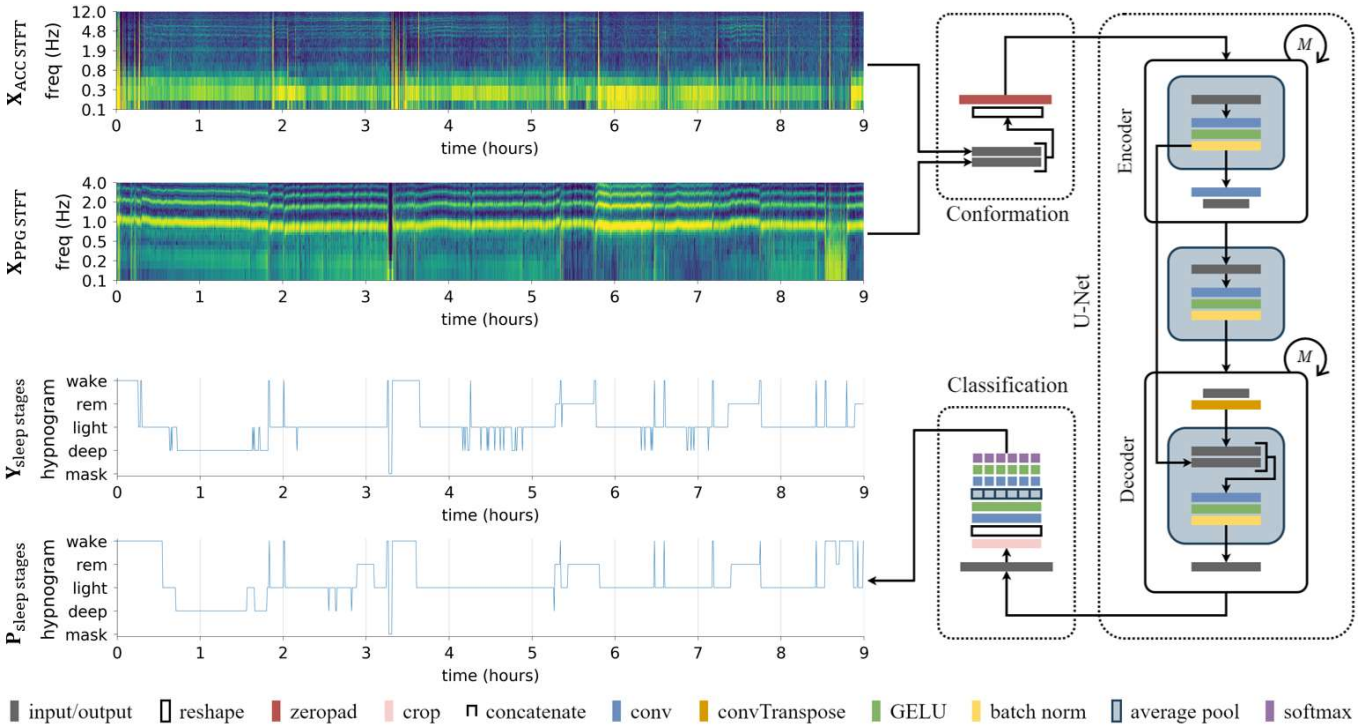

Fig. 1. Conceptual representation of the proposed deep neural network (DNN) in an example recording. Two time-aligned spectrograms: $\boldsymbol{X}_{ACC\ STFT} \in \mathbb{R}^{N\times 64}$ and $\boldsymbol{X}_{PPG\ STFT} \in \mathbb{R}^{N\times 64}$, are firstly concatenated, reshaped, and zero-padded to conform to the subsequent temporal module. Then the segments are processed in the deep convolutional neural network, inspired by U-Net [28]–[30] that consists of M encoder and decoder blocks. Finally, the output is segmented into sleep epochs of 30 s duration and classified into 4 classes: wake, light sleep, deep sleep. The classification module is inspired by the segment classifier from U-Sleep [28]. The argmax of the model predictions, $\boldsymbol{P}_{\text{sleep stages}}$, $argmax$: $\mathbb{R}^{T\times 4} \rightarrow \mathbb{R}^{T\times 1}$, is presented along with the ground truth, $\boldsymbol{Y}_{\text{sleep stages}}$, hypnogram for comparison. Periods with data loss are labeled with *mask*. M: Number of encoder and decoder blocks in U-net; T: number of sleep epochs; N: duration in seconds of the recording; GELU: Gaussian Error Linear Unit activation function [32]; conv: convolution; convTranspose: transposed convolutional; batch norm: batch normalization [33]; STFT: Short Time Fourier Transform; ACC: Accelerometry; PPG: Photoplethysmography.

The decoder, $\varphi_{\mathrm{D}}: \mathbb{R}^{\tilde{T}_S/2^M,[\tilde{F}_S/2^M],2^{M/3}C_U} \rightarrow \mathbb{R}^{\tilde{T}_S\times\tilde{F}_S\times C_U}$ serves to merge and process the feature representations from the encoder at each resolution to enhance tractability throughout the network and to reshape the segment to its original temporal resolution. Each block m in the decoder φ_D consists of a transposed convolutional layer with kernel- and stride of size $(2,2)/(2,1)$ that performs spatial-temporal up-sampling (nearest neighbor) and a convolutional operation on the input. The up-sampled feature map is concatenated with the feature map that has the corresponding temporal resolution originating from the encoder. Finally, the concatenated feature maps are processed by a similar block to that of the encoder (See Fig. 1).

The classification module $\varphi_C: \mathbb{R}^{\tilde{T}_S\times\tilde{F}_S\times C_U} \rightarrow \mathbb{R}^{T\times 4}$, where T is the number of sleep epochs, serves to segment and classify the feature maps into sleep stage vectors. It is inspired by the segment classifier proposed by Perslev, et al. [28]. The feature maps are cropped to remove the zeros that were padded in the conformation module and reshaped into a 2D vector. The 2D vector is segmented by a temporal average pool operator that reduces the temporal axis to match the desired temporal output resolution of 30 s, i.e., 1 sleep epoch. A (1,1) convolution with GELU activation is applied before and after the average pool operator to increase flexibility. Finally, the softmax function, which treats the sleep stage classes as mutually exclusive, calculates the probability function over all classes for each timestep. These predictions can be further processed into a single representative class label for each timestep to obtain a hypnogram: argmax: $\mathbb{R}^{T\times 4} \rightarrow \mathbb{R}^{T\times 1}$ (See Fig 1.).

B. *Loss function*

Let $\mathbf{X}^{(s)}$, $s = \{1, 2\}$ denote two time series segments of ACC and PPG, respectively. Then, let $f: \mathbf{X}^{(s)} \rightarrow \mathbf{P} \in \mathbb{R}^{T\times 4}$ be the proposed DNN that takes $\mathbf{X}^{(s)}$ as input and outputs 4 class predictions for each output timestep: $t = \{1, \ldots, T\}$, such that the probability of sleep stage k, at timestep t, is given by $P_{tk} = \exp(Z_k) \,/\, \sum_{i=1}^{4} \exp(Z_i)\,, k \in \{1, \ldots, 4\}$, where $\mathbf{Z}$ is the output from the layer before the softmax layer. Let $\mathbf{Y} \in \{0,1\}^{T\times 4}$ be the corresponding one-hot encoded target vector. The objective is to estimate the parameters of f, found through optimization, that minimizes the loss function, given by the balanced categorical cross-entropy:

$$\mathcal{L}(\mathbf{P}, \mathbf{Y}) = -\frac{1}{BT}\sum_{b=1}^{B}\sum_{t=1}^{T}\sum_{k=1}^{4}\frac{1-\beta}{1-\beta^{n_k}} Y_{btk} \log\left(P_{btk}\right) \qquad (1)$$

Where $\mathcal{L}$ is the mean loss for the given batch, T is the output segment length, i.e., number of 30 s sleep epochs, n_k is the number of samples of class k in batch b, and β=0.999. $\mathbf{Y}_{btk}\log\left(\mathbf{P}_{btk}\right)$ is the categorical cross entropy that induce exponential penalty to the loss function the further away the prediction $\mathbf{P}$ is from the target $\mathbf{Y}$. $\frac{1-\beta}{1-\beta^{n_k}}$ is a balancing factor included to account for the class imbalance that naturally exists for sleep stage classification. It is based on the idea that as the

TABLE I
DATA COHORT OVERVIEW WITH DEMOGRAPHIC AND SLEEP RELATED INFORMATION.

Parameter	TBI	STAGES PSG	STAGES Arc	Health	p-value
ACC source	GT3X actigraph (100 Hz)	Amazfit Arc (25 Hz)		Amazfit Health (25 Hz)	
PPG source	PSG (100 Hz)	PSG (128 Hz)	Amazfit Arc (25 Hz)	Amazfit Health (50 Hz)	
Participants (Train/test)	231 (185/46)	35 (18/17)		35 (0/35)	
Gender, % male	81.4	45.7		40.0	<0.001
Age, $\frac{\mu\pm\sigma}{\text{years}}$	38.4±20.6	38.3±13.6		36.2±13.6	0.800
BMI, $\frac{\mu\pm\sigma}{\text{kg/m}^2}$	26.2±5.2	29.3±8.5		28.6±7.8	0.005
Wake, %	24.3±17.2	29.5±12.5		16.9±6.3	0.089
REM, %	14.5±8.6	13.5±6.1		10.8±4.8	0.034
Light, %	41.5±15.3	44.5±9.9		42.5±12.4	0.589
Deep, %	19.4±12.0	12.3±7.6		14.8±10.4	<0.001
AHI, $\boldsymbol{\mu \pm \sigma}$ (none, mild, moderate, severe)	17.6±20.2 (72, 78, 39, 41)	13.1±10.6 (11,11, 8, 5)		16.0±24.0 (9, 17, 4, 5)	0.552
ArI, $\boldsymbol{\mu \pm \sigma}$ (none, mild, moderate, severe)	21.1±15.2 (3, 98, 86, 44)	16.2±10.9 (4, 14, 14, 3)		12.2±10.1 (7, 21, 5, 2)	0.002

All sleep values come from the PSG recording associated with the participants. Statistical comparison of the gender fraction was made with chi-square test; all other statistical comparisons were made with one-way analysis of variance. None: AHI<5; mild: 5≤AHI<15; moderate: 15≤AHI<30, severe: 30≤AHI; μ: mean; σ: standard deviation; BMI: Body mass index; PSG: Polysomnography; AHI: Apnea-Hypopnea Index; ArI: Arousal Index; STAGES: Stanford Technological Analytics and Genomics in Sleep study; TBI: Traumatic Brain Injury study; Health: Amazfit Health study; ACC: Accelerometry; PPG: Photoplethysmography

number of samples for a given class increase, the additional benefit of additional data points will diminish [34].

C. Data

Data used in our experiments come from the Stanford Technological Analytics and Genomic in Sleep (STAGES) study, the Traumatic Brain Injury (TBI) study, and from the Amazfit Health (Health) study. Demographic and sleep related information are presented in Table I. Inclusion criteria is shown in supplementary material.

1) The Stanford Technological Analytics and Genomics in Sleep (STAGES) study

Participants for this study were patients referred to the Stanford Sleep Center in Redwood City for PSG examination whom upon request agreed to participate in the study. All procedures were pre-approved by the Stanford University Institutional Review Board (IRB #36071). Participants were equipped with a low-cost wrist-worn device, the Amazfit Arc (Huami, Inc.), for collection of ACC and PPG while undergoing in lab nocturnal PSG recording. A customized application was developed in collaboration with Huami, Inc. for the collection of raw sensor data from the Amazfit Arc device. It records tri-axial acceleration data at 25 Hz with 12-bit resolution at a dynamic range of $\pm 8g$ and optic data also at 25 Hz with its PPG sensor. Data were transferred in real time from the device to a smartphone via Bluetooth throughout the recording and finally uploaded to a cloud storage when the recording ended. A timestamp was saved along with the raw sensor data every second throughout the recording to ensure that data from the device and PSG could be synchronized. The associated sleep stage and sleep related event scorings were scored according to the 2007 AASM criteria [24]. Of the 323 participants, data from 201 participants were lost due to a practical problem that was mitigated in a subsequent update of the application. Of the remaining 122 recordings, 42 were excluded as they were shorter than 4 hours and a further 45 were excluded because the PPG was so noisy that there was no distinct heart rate throughout the recording. Two datasets were defined with the data from the remaining 35 complete recordings: STAGES Arc, which contains ACC and PPG from the Arc device and STAGES PSG, which contains ACC from the Arc device and PPG from the overlapping PSG recording.

2) Traumatic Brain Injury (TBI) study

Nocturnal PSG and overlapping ACC recordings from 271 participants involved in a study of sleep disordered breathing in patients undergoing rehabilitation from a traumatic brain injury [35] were also included. The study extent and scope is described elsewhere [35]. Of these, 12 were excluded because they were missing either ACC or PPG signals, and further 28 were excluded because they had unstable sample frequency. Tri-axial acceleration data were recorded with a GT3X actigraph (Actigraph Corp, Pensacola FL) at 100 Hz with 12-bit resolution at a dynamic range of $\pm 6g$. Data and timestamps were extracted from the device after the recording and was synchronized to the PSG. The finger-probe PPG signal was extracted from the overlapping PSG recording. The associated sleep stage and sleep related event scorings were scored according to the 2007 AASM criteria [24].

3) Amazfit Health (Health) study

Participants for this study were recruited as outlined for the STAGES study. All procedures for this study were approved by Stanford University IRB #55476. Participants were equipped with a wrist-worn CST, the Amazfit Health (Huami, Inc.), that similarly to Amazfit Arc records tri-axial acceleration data at 25 Hz with 12-bit resolution at a dynamic range of $\pm 8g$ but collects optical data at 50 Hz with its PPG sensor. The practical setup for the Health study and the further data collection follows the procedure described above for the STAGES study. Of the 54 participants that were recruited for this study, 14 were excluded due to a Bluetooth instability problem, and further 5 were excluded because their recordings were shorter than 4 hours. A total of 35 complete recordings were included and used as a hold-out test set.

D. Initial preprocessing

Data from all datasets were processed using the following initial preprocessing steps. Both ACC and PPG were resampled to a uniform time series with a sampling rate of 32 Hz. Signals

TABLE II
MODEL SETTINGS FOR PREPROCESSING FRAMEWORKS EXPERIMENT

Input	(T_s, F_s, C_s)	K	L	C_U	M	B	T
ACC low-res	$\left(\frac{N}{30}, 1, 1\right)$	(16,1)	(2,1)	16	5	2	1024
ACC	$(32N, 1, 3)$	(16,1)	(2,1)	6	15		
ACC STFT	$(N, 64, 1)$	(16,3)	(2,2)	16	10		
ACC features [7]	$\left(\frac{N}{30}, 1, 36\right)$	(16,1)	(2,1)	16	5		
PPG[22]	$(32N, 1, 1)$	(16,1)	(2,1)	6	15	2	1024
PPG STFT	$(N, 64, 1)$	(16,3)	(2,2)	16	10		
PPG surrogate [11]	$(4N, 1, 3)$	(16,1)	(2,1)	16	12		
PPG features [14]–[21]	$\left(\frac{N}{30}, 1, 294\right)$	(16,1)	(2,1)	16	5		

Please refer to supplementary material for additional experiments. T_s, F_s, and C_s for the input s, is the temporal, spatial, and channel dimensions, respectively. K: kernel size; L: stride; C_U: Filter width of temporal module; M: Number of encoder and decoder blocks. B: Batch size; T: segment size in epochs.

with a sampling rate higher than 32 Hz were lowpass filtered before down-sampling to guard against aliasing using a Chebyshev filter with a cutoff frequency of 12 Hz and a passband ripple of 0.05 dB. Signals with a sampling rate lower than 32 Hz were interpolated using piecewise cubic Hermite Interpolation polynomial (PCHIP). Periods with data loss were labeled as *mask*. Data loss affected a total of 1.3%, 7.4%, 9.7%, and 5.3% of the recording time for the TBI, STAGES PSG, STAGES Arc, and Health datasets, respectively. Further preprocessing is presented in the experimental section.

E. Evaluation

Each dataset was partitioned into a training and a test set as reported in Table I. The training set was further portioned into a training (70 %) and a evaluation set (30 %). Experiments that influence parameter choice are reported on the evaluation dataset, e.g., model optimization and hyperparameter tuning. The remaining experiments are reported on the test set.

The proposed model was trained with an online learning procedure, where, for the training set, segments were prepared in pseudo-class-balanced batches governed by the following sampling procedure. Firstly, a class was uniformly sampled from the class set {W, L, D, R}. Then, a segment with size T and starting point $\max(0, D_r - T)$, where D_r is the duration in epochs of recording r, was randomly sampled iteratively until it satisfied the condition of containing at least 1 epoch with the selected class. Segments from the evaluation- and test set were sampled in an ordered manner, such that each sleep epoch from each recording was evaluated only once.

Input segment sizes ranging from 1 to 1024 epochs (30 s to ~8.5 hrs.) were tested in the experimental section. Recordings that were shorter than the required input segment size were zero-padded. The extended part was in turn masked in the loss function and did not influence model parameter learning. The output predictions from an entire recording were formed by concatenating its corresponding predicted subsegments.

The loss presented in (1) was calculated for each batch and was minimized using ADAM [36] optimizer with a learning rate of 10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, which in turn was divided by a factor of $\sqrt{10}$ every time the performance of the evaluation set did not improve for more than 10 training epochs (i.e., a complete iteration through the training set). Epochs with more than 50% missing data were masked in the loss function and did not influence model parameter learning. The learning procedure was stopped when the evaluation performance did not improve over the course of 25 epochs; the model with the highest performance on the evaluation set was saved. All weights and biases of the network were initialized using Kaiming normal initialization [37]. The proposed model was built with Python 3.6.8, and the DNN was implemented in Keras 2.6.0 and Tensorflow 2.6.2.

F. Performance metrics

The ACC and PPG modalities were collected with different devices with different technical specifications for each dataset (See Table I). Therefore, performance was computed separately for each dataset. Sleep stage prediction constitutes a class imbalanced classification problem; thus, optimizing after overall accuracy will bias the result toward the most common sleep stage. To account for this, the performance of the experiments was evaluated with respect to multiple performance metrics, hereunder $F1$-score, accuracy, and Cohen's κ. $F1$-score was computed for each class separately, and accuracy and Cohen's κ was computed across all classes.

For each participant, the following sleep metrics were considered [38]: total sleep time (TST), sleep onset latency (SOL), wake after sleep onset (WASO), and sleep efficiency (SE). Performance with respect to each sleep metric is reported as root mean squared error, $RMSE = (\sum_{r=1}^{R}(\mathrm{y}_r - \hat{\mathrm{y}}_r)^2/R)^{1/2}$, where y_r and $\hat{\mathrm{y}}_r$ refer to the target and the predicted sleep metric, respectively, $r \in [1, .., R]$ where R is the number of recordings.

III. EXPERIMENTS

A. Deep neural network parameter tuning

Multiple experiments were performed to identify the set of parameters for the proposed DNN that produced the best overall performance on the evaluation set. Specifically, the following parameters were found through grid search: kernel size $K = 16$ and filter width $C_U = 16$. Please refer to the supplementary material for a complete overview of the conducted experiments.

B. Impact of preprocessing framework

The impact of preprocessing on each input modality was investigated by evaluating different preprocessing frameworks with an increasing number of processing steps. The parameters of the DNN were modified for each preprocessing framework to account for the discrepancies in input dimensionality. Specifically, the kernel size, K, the stride, L, and the filter width, C_U of the convolutional layers, and the number of encoder and decoder blocks, M, change. Table II presents the model settings for each preprocessing framework. Please refer to supplementary material for a detailed presentation of each preprocessing framework.

The following four frameworks were considered for the ACC modality: ACC low-res: $\mathbf{X}_{\mathrm{ACC\,low\,res}} \in \mathbb{R}^{N/30\times 1}$ is the Euclidian norm of the ACC signal, summarized within each sleep epoch. This corresponds to traditional actigraphy measurements such as those used as input to the Cole-Kripke

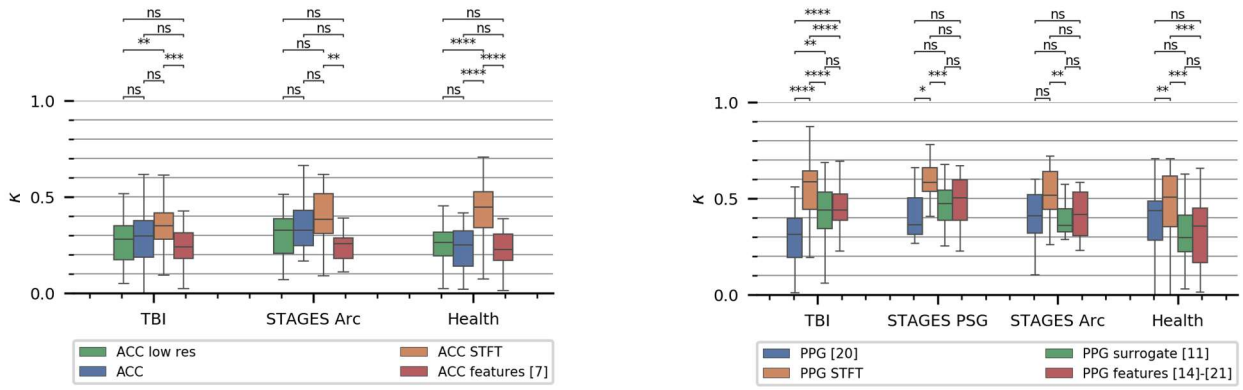

Fig. 2 – Impact of preprocessing for preprocessing of ACC (left) and PPG (right) by datasets. Please refer to supplementary material for further performance metrics as well as detailed information about how each preprocessing framework was implemented and processed by the proposed deep neural network. Pairwise t-tests for the performance metrics were conducted using a Bonferroni corrected significance level. ns: not significant; *: $p < 0.05/6$; **: $p < 0.01/6$; ***: $p < 0.001/6$; ****: $p < 0.0001/6$. ACC: Accelerometry; PPG: Photoplethysmography; STFT: Short time Fourier transform; STAGES: Stanford Technological Analytics and Genomics in Sleep study; TBI: Traumatic Brain Injury study. Health: Amazfit Health study.

algorithm [3]. ACC: $\mathbf{X}_{\mathrm{ACC}} \in \mathbb{R}^{32N\times 3}$ is the raw, normalized, triaxial ACC modality. Each directional vector was normalized to have a median of zero and an inter-quartile range (IQR) between -1 and 1. ACC STFT: $\mathbf{X}_{\mathrm{ACC\ STFT}} \in \mathbb{R}^{\mathrm{N}\times 64}$ is the average of the Short Time Fourier Transformed (STFT) ACC signal of each directional vector using a Blackman window function with a duration of 10 s, with a 9 s overlap, and using 128 sampling points to calculate the Fourier transform. Frequency content outside the range $(0,16]$ Hz was removed to produce an output of size: $\mathbb{R}^{\mathrm{N}\times 64}$. ACC features: $\mathbf{X}_{\mathrm{ACC\ features}} \in \mathbb{R}^{N/30\times 36}$ are the ACC features; following the procedure of Sundararajan et. al. the ACC vectors were processed into 3 surrogate feature signals: $\mathbf{X}_{\mathrm{EN}}$, $\mathbf{X}_{\mathrm{z-angle}}$, and $\mathbf{X}_{\mathrm{LIDS}}$. 12 features were extracted from each of these 3 surrogate signals, yielding a total of 36 features per 30 s sleep epoch [7].

Four preprocessing frameworks were also evaluated for the PPG modality. The PPG signal, $\mathbf{X}_{\mathrm{PPG}} \in \mathbb{R}^{32\ \times 1}$, was initially bandpass filtered between with a passband frequency range of $[0.1, 8]$ Hz. Then, an adaptive version of the IQR normalization method was implemented, to account for the amplitude variation that typically exists for the PPG modality; the median and quartiles were calculated for a sliding window of size 300 s. Finally, outliers outside 20 times the IQR-range were clipped. PPG STFT: $\mathbf{X}_{\mathrm{PPG\ STFT}} \in \mathbb{R}^{N\times 64}$ is the STFT PPG signal using a Blackman window function with a duration of 10 s, with a 9 s overlap, and using 512 sampling points to calculate the Fourier transform. Frequency content outside the range $(0,4]$ Hz was removed to produce the output with size: $\mathbb{R}^{N\times 64}$. $\mathbf{X}_{\mathrm{PPG\ surrogate}} \in \mathbb{R}^{4N\times 3}$ are the PPG surrogate signals [11] extracted from the PPG pulse peaks, which were found using adaptive pulse segmentation [39]. For each pulse peak, the amplitude modulation (AM), i.e., the peak amplitude, the frequency modulation (FM), i.e., duration interval between consecutive beats, and the baseline wander (BW), i.e., the combined FM and AM, were extracted. The PPG surrogate signals were interpolated using PCHIP and resampled to 4 Hz. $\mathbf{X}_{\mathrm{PPG\ features}} \in \mathbb{R}^{N/30\times 294}$ are the PPG features. Many of the features from the feature-based approaches overlap [14]–[21], therefore, it was chosen to implement the union of these features as one common approach. 98 features were extracted using the following segment sizes, centered over a given sleep epoch: 30 s, 150 s, and 270 s, creating a feature pool of 294 features per 30 s sleep epoch (See supplementary material).

The overall κ test performance is presented for each preprocessing framework in Fig. 2. These results show that for both modalities spectral preprocessing outperforms all other frameworks across all datasets, and in most cases significantly. A significant performance gain is found for high resolution ACC (ACC STFT) when compared to traditional, low-resolution ACC for the TBI dataset and the hold-out test dataset.

C. Benchmark

The performance of the proposed approach is compared to SOTA works in Table III. Only those who report on 4 sleep stage classes were included. The performance for the proposed approach is presented for the three internal test sets and the hold-out test set for different modality combinations to allow comparison between works with different modality inputs.

Overall, Table III shows that in most cases, the combined approach has the overall best performance across datasets. For the PPG modality, we observe that the model performs better on STAGES PSG ($\kappa = 0.58 \pm 0.11$) when compared to STAGES Arc ($\kappa = 0.51 \pm 0.14$). Data from these datasets came from simultaneous recordings from the same participants but using different PPG sensors. Furthermore, it was observed that the overall performance is equivalent for both the internal and the hold-out test sets that contain wrist-worn data: STAGES Arc ($\kappa = 0.59 \pm 0.09$) and Health ($\kappa = 0.58 \pm 0.18$), though the latter had more variation.

All approaches that used both ACC and PPG were feature-based [14]–[20]. The best performing of these, i.e., Wulterkens et al., reported similar performance to that of the proposed approached: $\kappa = 0.62 \pm 0.12$ [19], while all other approaches had substantially lower performance. One feature-based approach reported a performance increase from $\kappa = 0.55$ to $\kappa = 0.65$ by pretraining their classifier on features extracted from ECG [21]. Korkalainen et al., who presented the only DL approach trained on raw PSG-based PPG signals, reported a performance of $\kappa = 0.54$ [22]. Sundararajan et al. [7] presented a feature-based approach solely trained on high resolution ACC recording. They reported significantly lower performance when compared to the presented approach.

TABLE III
OVERALL PERFORMANCE OF THE PROPOSED APPROACH ON INTERNAL- AND HOLD-OUT TEST SET, AND PERFORMANCE OF RELATED STATE-OF-THE-ART WORKS.

Dataset train/test	Input	Model	F1				Accuracy	Cohen's κ	RMSE			
			Wake	Light	Deep	REM			TST	SOL	WASO	SE
TBI 185/46	ACC STFT	CNN	0.64±0.17	0.54±0.11	0.49±0.25	0.27±0.23	0.55±0.10	0.35±0.14	41.8	28.7	37.2	10.0
	PPG STFT	CNN	0.71±0.14	0.63±0.16	0.70±0.24	**0.71±0.24**	0.71±0.11	**0.58±0.16**	49.5	23.7	37.1	10.5
	ACC+PPG STFT	CNN	**0.72±0.14**	**0.65±0.11**	**0.72±0.23**	0.65±0.22	**0.71±0.09**	0.56±0.13	**37.3**	**17.3**	**32.4**	**8.0**
STAGES PSG 18/17	ACC STFT	CNN	**0.81±0.10**	0.66±0.08	0.42±0.26	0.18±0.18	0.61±0.09	0.38±0.16	**31.7**	57.7	53.7	7.5
	PPG STFT	CNN	0.79±0.12	0.67±0.10	0.56±0.23	0.77±0.18	0.69±0.07	0.58±0.11	47.9	63.5	44.0	7.2
	ACC+PPG STFT	CNN	**0.81±0.09**	**0.76±0.07**	**0.68±0.26**	**0.79±0.19**	**0.76±0.07**	**0.64±0.09**	37.0	**41.2**	**28.8**	**5.7**
STAGES Arc 18/17	ACC STFT	CNN	**0.81±0.10**	0.66±0.08	0.42±0.26	0.18±0.18	0.61±0.09	0.38±0.16	**31.7**	57.7	53.7	7.5
	PPG STFT	CNN	0.76±0.14	0.74±0.09	0.45±0.29	**0.65±0.22**	0.68±0.10	0.51±0.14	70.7	52.1	66.6	13.8
	ACC+PPG STFT	CNN	0.79±0.09	**0.75±0.06**	**0.59±0.27**	0.64±0.23	**0.73±0.06**	**0.59±0.09**	43.7	**24.4**	**31.4**	**6.9**
Health 0/35	ACC STFT	CNN	0.80±0.14	0.66±0.12	0.50±0.24	0.26±0.20	0.64±0.11	0.45±0.15	67.7	59.5	42.3	9.1
	PPG STFT	CNN	0.76±0.15	**0.67±0.13**	0.54±0.26	0.51±0.26	0.68±0.13	0.50±0.19	61.1	68.2	61.8	12.7
	ACC+PPG STFT	CNN	**0.83±0.13**	0.66±0.13	**0.63±0.26**	**0.58±0.25**	**0.69±0.13**	**0.58±0.18**	**49.2**	**56.3**	**40.4**	**7.9**
134/24	ACC features [7]	RF	0.55	0.57	0.21	0.12						
135/80	ACC+PPG features [14]	BLD					0.59±0.09	0.42±0.12	34.3	10.2	25.3	7.4
60/60	ACC+PPG features [15]	LDC	0.70	0.71	0.62	0.67	0.69	0.52±0.14				
50/50	ACC+PPG features [17]	LDC					0.77	0.58				
543/292	ACC+PPG features [19]	LSTM	0.73±0.17	0.78±0.10	0.69±0.24	0.74±0.18	0.76±0.07	0.62±0.12	34.3	25.3	40.6	6.7
23/23	HR, sleep metric, demographic features [20]	SVM+ XGB					0.73±0.12	0.43±0.21				
584,60/60	PPG features [21] (pretrain w. ECG)	LSTM	0.71	0.75	0.73	0.80	0.76±0.08	0.65±0.11				
805/89	PPG (PSG) [22]	CNN+ GRU	0.74	0.67	0.54	0.71	0.69	0.54				

Three different input configurations were considered for the internal test sets and the hold-out test set. Best performing inputs are highlighted in bold font for each dataset. Performance metrics are reported by $F1$ score for each sleep stage, overall accuracy, and overall Cohen's κ, and root mean squared error (RMSE) for the following sleep metrics all reported in minutes: Total sleep time (TST); Sleep onset latency (SOL); Wake after sleep onset (WASO); Sleep efficiency (SE). Please refer to supplementary material for performance on additional sleep metrics. $F1$ score, accuracy, and Cohen's κ performance is reported either as $\mu \pm \sigma$ (average and standard deviation) across recordings or as a single value across all sleep epochs. ACC: Accelerometry, PPG: Photoplethysmography; STFT: Short time Fourier transform; HR: Heart rate; CNN: Convolutional neural network; RF: Random Forest; GRU: Gated recurrent unit; LSTM: Long-short term memory; BLD: Bayesian linear discriminant; LDC: Linear discriminant classifier. SVM: Support vector machine, XGB: gradient boosting decision tree; STAGES: Stanford Technological Analytics and Genomics in Sleep study; TBI: Traumatic Brain Injury study. Health: Amazfit Health study.

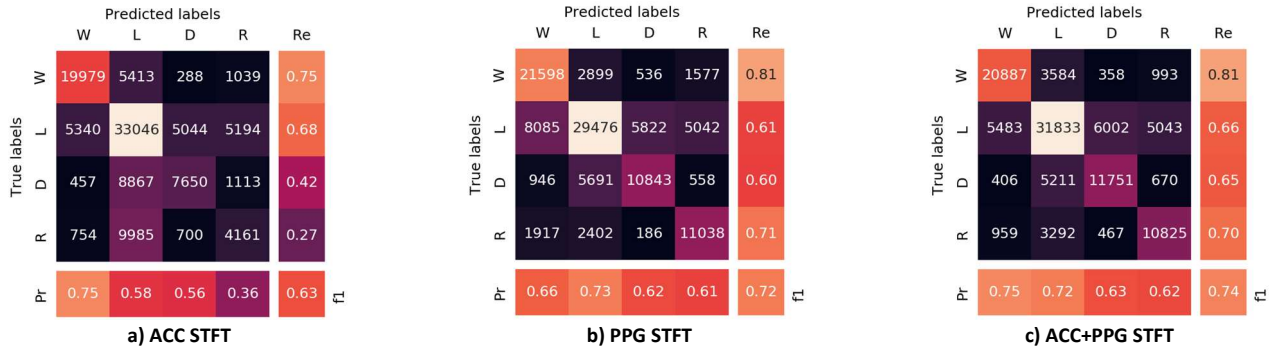

Fig. 3 – Confusion matrices showing epoch-by-epoch comparison of predicted and target classes shown for three different input modalities for all test datasets. Please refer to supplementary material for by-dataset confusion matrices. W: Wake; L: Light sleep (N1, N2); D: Deep sleep (N3); R: REM sleep (Rapid Eye Movement sleep); Pr: Precision; Re: Recall; f1: $F1$-score; ACC: Accelerometry; PPG: Photoplethysmography; STFT: Short time Fourier transform

D. Modality importance

Fig. 3 presents confusion matrices with respect to each sleep stage class for each modality combination, across all datasets. While the overall performance across all sleep stages is better for the PPG based approach, Fig. 3 reveals that the ACC based approach produce less false positives to wake, $\mathrm{Pr} = 0.75$, when compared to the PPG based approach, $\mathrm{Pr} = 0.66$. The ACC based approach struggles to distinguish between the remaining sleep stages, and most significantly to identify REM sleep, where it performs with $\mathrm{Pr} = 0.36$ and $\mathrm{Re} = 0.27$. The combined approach proves to incorporates the best from the two modalities; it has the high performance from the PPG modality and the improved precision to wake from the ACC modality. Investigation of Table III elucidates that adding ACC improves the RMSE for almost all sleep metrics and improves performance with 8 percentage points for both CST-based datasets, i.e., STAGES Arc and Health.

TABLE IV
MODEL SETTINGS FOR THE INPUT SEGMENT SIZE EXPERIMENT

Input	(T_s, F_s, C_s)	K	L	C_U	M	B	T
ACC + PPG STFT	$(N, 64, 2)$	(16,3)	(2,2)	16	10	2	1024
					8	8	256
					6	64	32
					5	256	8
					5	2048	1

T_s, F_s, and C_s for the input s, is the temporal, spatial, and channel dimensions, respectively. K: kernel size; L: stride; C_U: Filter width of temporal module; M: Number of encoder and decoder blocks. B: Batch size; T: segment size in epochs.

E. Impact of input segment size

In this experiment, the importance of temporal modelling was assessed by training the proposed DNN for different input segment sizes. Five input segment sizes were tested with durations of 30 s, 240 s, 960 s, 7680 s, and 30720 s, which corresponds to 1, 8, 32, 256, and 1024 sleep epochs. Model settings for this experiment is presented in the Table IV. By design, the batch size, B, and the input segment size, T, have a reciprocal relationship, such that the number of sleep epochs evaluated in a batch is constant, i.e., 2048.

Fig. 4 shows the κ test performance for the different input segment sizes, T, presented by sleep stage. This experiment shows that performance increases with input segment size across all sleep stages. The impact is mostly significant for smaller window sizes; though REM sleep drastically improve when input segment size is changed from 32 to 256 epochs; the latter corresponds to 128 minutes, i.e., more than a sleep cycle.

IV. DISCUSSION

The proposed DL-based approach achieved a participant averaged performance of $Acc = 0.71, 0.76, 0.73$ and $\kappa = 0.56, 0.64, 0.59$ on the internal test sets, i.e., TBI, STAGES PSG, and STAGES Arc, respectively, and with $Acc = 0.69$ and $\kappa = 0.58$ on the hold out test set, i.e., Health, when using both ACC and PPG as input, both processed with the STFT and with an input segment size of 1024 sleep epochs (~8.5 hrs.).

The impact of preprocessing was investigated by evaluating different preprocessing frameworks. It was shown that for both modalities STFT preprocessing achieved the highest performance, outperforming both low-resolution, raw [22], surrogate [11], and feature-based approaches [7], [14]–[21]. Spectral decomposition is likely to achieve great generalization because it enhances the quasiperiodic signals such as heart rate and breathing contained in the ACC and PPG modalities by dispersing noise across all frequencies. While raw signals are information-rich, they are also very noisy, especially from wrist-worn CSTs, and the morphology of the PPG pulse wave may vary based on sensor quality, sensor placement, skin thickness and skin tone. The datasets used in this study may not be diverse and large enough for the DNN model to work well on raw PPG and ACC signals. Data augmentation may partly address this limitation; however, this was not tested in this study. On the other hand, feature-based approaches are based on capturing the essence of the modalities and in turn to reduce the complexity of the signals. Here, information may unintentionally be removed during this process, as there is no guarantee that the chosen features are optimal for the classification task. Furthermore, feature-based approaches rely on a sequence of processing steps and risk culminating errors.

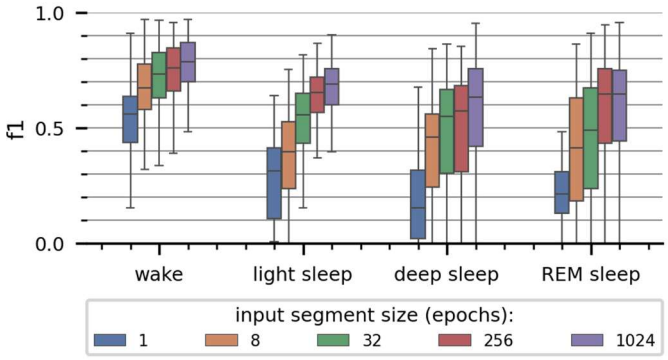

Fig. 4 – Input segment size importance. Performance as a function of input segment size for the combined approach (ACC+PPG STFT) with respect to sleep stages, across all datasets. 1 epoch is 30 s. Please refer to supplementary material for by-dataset performance. f1: $F1$-score ACC: Accelerometry; PPG: Photoplethysmography; STFT: Short time Fourier transform.

SOTA comparison analysis showed that the proposed approach had similar performance to the best performing feature-based algorithms [19] but using less processing steps. Performance remained high for presented approach when applied to the hold test set, proving its generalizability to data collected with wrist-worn CSTs. Radha et al. presented a transfer learning approach that utilize nocturnal ECG-recordings to improve sleep stage classification performance when applied to wrist-worn PPG recordings [21]. It is well established that performance increases with the amount of training data [26]; we are confident that adding more training data will improve performance of the presented approach.

For the ACC modality, the presented approach proved to outperform the feature based approach, presented by Sundararajan et al. [7], across all sleep stages (See Table III). Furthermore, a significant performance gain was found for high resolution ACC (ACC STFT) when compared to traditional, low-resolution ACC for the TBI dataset and the hold-out test dataset (See Fig. 2). These results indicate that the high resolution ACC does capture discriminative bio signals, thereby addressing the fundamental limitation of coarsely sampled actigraphy that is incapable of capturing motionless wake [4].

Investigation of signal modality importance showed that the combined approach had the overall best performance across datasets. The PPG modality had the most impact on performance and was the only modality capable of detecting REM sleep well. Our experiments showed that adding ACC made the model produce less false positives to wake. Since all sleep metrics included in Table III were based on sleep-wake behavior, this explains why adding ACC improved RMSE for almost all sleep metrics. ACC was found to have most impact on the overall performance for the datasets with wrist-worn CST data. The PPG sensor had more noise on these datasets, which could explain why ACC has more impact.

We found that the model performed better on the PPG modality originating from the PSG (STAGES PSG) when compared to the modality originating from the wrist-worn CSTs (STAGES Arc). This finding was consistent across sleep stages, which indicates that the PPG modality from the wrist-worn CST is more prone to noise when compared to PPG modality from

clinical grade equipment used in PSG recordings; a finding that was confirmed by visual inspection of the recordings.

U-Net has previously been used for sleep stage classification but only applied to relatively short sleep segments (20 mins) [28]; thus this implementation does not utilize the full potential of the U-Net architecture. Our experiments showed that performance increased with input segment size. This indicates that long-term dependencies that exist in sleep, e.g., cyclic behavior of the sleep profile, may be leveraged to achieve higher performance. The largest input segment size of 1024 epochs (~8.5 hrs.) did in many cases have global context, covering the entire recording. This fact introduces a bias, since the model is unintentionally informed about the beginning and end of the recording, where participants are always awake. Nevertheless, the finding that performance increase with input segment size still holds for the remaining segment sizes.

It is widely recognized that large batch sizes may have adverse effects on model performance [40]. Experimentation using lower batch sizes for models with smaller input segment sizes showed no significant change in performance. Please refer to supplementary materials for these results.

The autonomic and motoric changes found in the PPG and ACC modalities are quite simple in nature, but they have complex long-term dependencies. Therefore, the presented U-Net type architecture was designed with a simple block structure but with a strong temporal core. Other, more complex block architectures could potentially improve performance further, but such were not explored. CNNs have the inductive bias of being translational invariant. This makes them especially useful for image recognition tasks, where the position of the object of interest is irrelevant. The same logic does not follow for spectrograms, as the position in a spectrogram has important meaning. Despite this, the presented CNN-based model proved to work well, which indicates that it is the relative changes of the bio-signals, rather than the absolute position, that is discriminatively important.

The sampling rates for the different devices used in this study did not share a common denominator (See Table I). The re-sampling rate of 32 Hz was chosen as it was the next higher power of 2 with respect to the smallest sampling rate of 25 Hz. This was operationally convenient as it minimized the number of zero-paddings an input segment required in the conformation module. up-sampling approximates the sequence through interpolation of the exiting data points, whereas down-sampling may unintentionally remove important information in the signal. Both procedures may negatively affect performance. STFT was the only spectral transformation that was evaluated. Other spectral decompositions with improved time and frequency resolution would be of interest to explore.

The performance reported in this work is limited by the amount of data and the datasets used. For instance, the TBI cohort contains data from patients undergoing rehabilitation from traumatic brain injuries. Their brain status is likely to influence their sleep profile and motoric expression. Comparison between studies is difficult due to the difference in experimental design, device type and quality, patient health, etc. Ideally, the algorithms should be benchmarked on a large, standardized, dataset with raw wrist-worn CST data. However, such datasets do not exist, yet.

A significant portion of the data from the CSTs were lost during data collection. Bluetooth fallouts were likely to happen during long recordings, e.g., during a participant's bathroom visits, where the Bluetooth range limit got violated. Data loss caused by this could be avoided by choosing a CST with local storage that does not rely on a stable Bluetooth connection throughout the recording. Premature stopping was another likely cause of data loss, which could happen if the phone or device ran out of battery, was shut off, or if the App froze or was closed. Our study design required participants to start, stop, and upload data themselves after the recording, using an application on their phone. This design choice was necessary because the clinical coordinators were not present in the morning after the recording. We identified that the most significant cause of data loss was that data was never uploaded due to this problem, causing 62 % of the recordings to be excluded in the STAGES study. A procedural change mitigated this in the subsequent Health study, where participants were requested to assure that the data had been transferred in a follow-up phone interview. This successfully reduced the number of exclusions to 26 %. While most user errors can be mitigated by using simple and clear user guidelines, ultimately, limiting the number of requested user actions is preferable in the future. Finally, we found that the Amazfit Health device recorded with better signal quality compared to its predecessor, Amazfit Arc. 45 recordings were excluded due to low signals quality when using the latter compared to 0 for the former.

The validation presented here was only performed using nocturnal recordings. Additional validation studies of sleep stage classification using 24-hour sleep recordings could be useful, specially to assess patients with hypersomnia disorders. Importantly however, EEG based sleep is currently not easily measurable over the entire day while individuals are mobile, thus comparison with the PSG gold standard would only be possible under conditions of 24-hour bedrest.

V. Conclusion

In this study we introduce a flexible DNN with a strong temporal core that can process multivariate time series inputs with different dimensionality to predict sleep stages using ACC and PPG signals. The proposed approach exhibits strong classification performance compared with feature-based approaches and approaches that input raw data. The model was designed with a strong temporal core to capture long-term dependencies, and it proved to increase performance with input segment size. It was established that combining both ACC and PPG result in equally or better performance when compared to each modality separately. The PPG modality had the most impact on performance and was the only modality capable of detecting REM sleep well, whereas ACC improved wake precision, thus improving sleep metric estimation. Performance remained high for the presented DNN when it was applied to the hold test set, proving its generalizability to data collected with a wrist-worn CST. We believe that the presented work establishes wrist-worn CSTs that measure ACC and PPG as potential OOC sleep monitoring systems, given raw data is accessible and provided data stable data collection is ensured.

ACKNOWLEDGEMENTS

We would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results. Funding of STAGES was provided by the Klarman Family foundation. Funding of the TBI group was provided by a Patient-Centered Outcomes Research Institute (PCORI) Award (CER-1511-33 005). We thank Amazfit, a Zepp Inc. brand, for funding the devices for this study by a contract to Emmanuel Mignot registered at clinical.gov as NCT04429906.

CODE AVAILABILITY:

https://github.com/MADSOLSEN/SleepStagePrediction

REFERENCES

[1] A. Henriksen *et al.*, "Using fitness trackers and smartwatches to measure physical activity in research: Analysis of consumer wrist-worn wearables," *J. Med. Internet Res.*, vol. 20, no. 3, 2018.

[2] I. Perez-Pozuelo *et al.*, "The future of sleep health: a data-driven revolution in sleep science and medicine," *npj Digit. Med.*, vol. 3, no. 1, pp. 1–15, 2020.

[3] R. J. Cole, D. F. Kripke, W. Gruen, D. J. Mullaney, and J. C. Gillin, "Automatic sleep/wake identification from wrist activity," *Sleep*, vol. 15, no. 5, pp. 461–469, 1992.

[4] A. Sadeh, "The role and validity of actigraphy in sleep medicine: An update," *Sleep Med. Rev.*, vol. 15, no. 4, pp. 259–267, 2011.

[5] M. Zinkhan and J. W. Kantelhardt, "Sleep Assessment in Large Cohort Studies with High-Resolution Accelerometers," *Sleep Med. Clin.*, vol. 11, no. 4, pp. 469–488, 2016.

[6] J. Zschocke *et al.*, "Detection and analysis of pulse waves during sleep via wrist-worn actigraphy," *PLoS One*, vol. 14, no. 12, pp. 1–18, 2019.

[7] K. Sundararajan *et al.*, "Sleep classification from wrist-worn accelerometer data using random forests," *Sci. Rep.*, vol. 11, no. 1, pp. 1–10, 2021.

[8] J. Sen Wong, W. A. Lu, K. T. Wu, M. Liu, G. Y. Chen, and C. D. Kuo, "A comparative study of pulse rate variability and heart rate variability in healthy subjects," *J. Clin. Monit. Comput.*, vol. 26, no. 2, pp. 107–114, 2012.

[9] M. Radha *et al.*, "Sleep stage classification from heart-rate variability using long short-term memory neural networks," *Sci. Rep.*, vol. 9, no. 1, pp. 1–11, 2019.

[10] M. Mendez and M. Matteucci, "Sleep staging from Heart Rate Variability: time-varying spectral features and Hidden Markov Models," *Int. J. Biomed. Eng. Technol.*, vol. 3, pp. 246–263, 2010.

[11] M. Olsen *et al.*, "Automatic, electrocardiographic-based detection of autonomic arousals and their association with cortical arousals, leg movements, and respiratory events in sleep," *Sleep*, vol. 41, no. 3, pp. 1–10, 2018.

[12] M. Basner, B. Griefahn, U. Müller, G. Plath, and A. Samel, "An ECG-based algorithm for the automatic identification of autonomic activations associated with cortical arousal," *Sleep*, vol. 30, no. 10, pp. 1349–1361, 2007.

[13] M. Olsen, E. Mignot, P. J. Jennum, and H. B. D. Sorensen, "Robust, ECG-based detection of Sleep-disordered breathing in large population-based cohorts," *Sleep*, vol. 43, no. 5, 2020.

[14] P. Fonseca *et al.*, "Validation of photoplethysmography-based sleep staging compared with polysomnography in healthy middle-aged adults," *Sleep*, vol. 40, no. 7, 2017.

[15] Z. Beattie *et al.*, "Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals," *Physiol. Meas.*, vol. 38, no. 11, pp. 1968–1979, 2017.

[16] O. Walch, Y. Huang, D. Forger, and C. Goldstein, "Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device," *Sleep*, vol. 42, no. 12, pp. 1–19, 2019.

[17] I. Fedorin, K. Slyusarenko, W. Lee, and N. Sakhnenko, "Sleep stages classification in a healthy people based on optical plethysmography and accelerometer signals via wearable devices," *2019 IEEE 2nd Ukr. Conf. Electr. Comput. Eng. UKRCON 2019 - Proc.*, pp. 1201–1204, 2019.

[18] D. M. Roberts, M. M. Schade, G. M. Mathew, D. Gartenberg, and O. M. Buxton, "Detecting sleep using heart rate and motion data from multisensor consumer-grade wearables, relative to wrist actigraphy and polysomnography," *Sleep*, no. March, pp. 1–19, 2020.

[19] B. M. Wulterkens *et al.*, "It is All in the Wrist : Wearable Sleep Staging in a Clinical Population versus Reference Polysomnography," no. June, 2021.

[20] Z. Liang and M. A. Chapa-Martell, "A Multi-Level Classification Approach for Sleep Stage Prediction With Processed Data Derived From Consumer Wearable Activity Trackers," *Front. Digit. Heal.*, vol. 3, no. May, pp. 1–16, 2021.

[21] M. Radha *et al.*, "A deep transfer learning approach for wearable sleep stage classification with photoplethysmography," *npj Digit. Med.*, vol. 4, no. 1, pp. 1–11, 2021.

[22] H. Korkalainen *et al.*, "Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea," *Sleep*, no. May, pp. 1–10, 2020.

[23] F. Emmert-Streib, Z. Yang, H. Feng, S. Tripathi, and M. Dehmer, "An Introductory Review of Deep Learning for Prediction Models With Big Data," *Front. Artif. Intell.*, vol. 3, no. February, pp. 1–23, 2020.

[24] C. Iber, S. Ancoli-Israel, A. L. Chesson Jr., and S. F. Quan, "The AASM Manual for the Scoring of Sleep and Associated Events: Rules Terminology and Technical Specifications 1st ed." p. 59, 2007.

[25] P. Fonseca *et al.*, "Automatic sleep staging using heart rate variability , body movements , and recurrent neural networks in a sleep disordered population," no. April, pp. 1–10, 2020.

[26] A. N. Olesen, P. Jørgen Jennum, E. Mignot, and H. B. D. Sorensen, "Automatic sleep stage classification with deep residual networks in a mixed-cohort setting," *Sleep*, vol. 44, no. 1, pp. 1–12, 2021.

[27] N. Sridhar *et al.*, "Deep learning for automated sleep staging using instantaneous heart rate," *npj Digit. Med.*, vol. 3, no. 1, 2020.

[28] M. Perslev, S. Darkner, L. Kempfner, M. Nikolic, P. J. Jennum, and C. Igel, "U-Sleep: resilient high-frequency sleep staging," *npj Digit. Med.*, vol. 4, no. 1, pp. 1–12, 2021.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9351, pp. 234–241, 2015.

[30] H. Li and Y. Guan, "DeepSleep convolutional neural network allows accurate and fast detection of sleep arousal," *Commun. Biol.*, vol. 4, no. 1, pp. 1–11, 2021.

[31] M. Olsen, H. B. D. Sørensen, P. J. Jennum, and E. Mignot, "Sleep stage prediction and slkeep disordered breathing detection using raw actigraphy and photoplethysmography from wearable consumer device," *Sleep*, vol. 43, no. Abstract Supplement, 2020.

[32] D. Hendrycks and K. Gimpel, "Gaussian Error Linear Units (GELUs)," pp. 1–9, 2016, [Online]. Available: http://arxiv.org/abs/1606.08415.

[33] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *Proc. 32nd Int. Conf. Mach. Learn. PLMR*, vol. 37, pp. 448–456, 2015.

[34] Y. Cui, M. Jia, T. Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 9260–9269, 2019.

[35] R. Nakase-Richardson *et al.*, "Comparison of Diagnostic Sleep Studies in Hospitalized Neurorehabilitation Patients With Moderate to Severe Traumatic Brain Injury," *Chest*, vol. 158, no. 4, pp. 1689–1700, 2020.

[36] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–15, 2015.

[37] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification Kaiming," *Biochem. Biophys. Res. Commun.*, vol. 498, no. 1, pp. 254–261, 2018.

[38] M. Ohayon *et al.*, "National Sleep Foundation's sleep quality recommendations: first report," *Sleep Heal.*, vol. 3, no. 1, pp. 6–19, 2017.

[39] P. H. Charlton *et al.*, "Extraction of respiratory signals from the electrocardiogram and photoplethysmogram: Technical and physiological determinants," *Physiol. Meas.*, vol. 38, no. 5, pp. 669–690, 2017.

[40] N. S. Keskar, J. Nocedal, P. T. P. Tang, D. Mudigere, and M. Smelyanskiy, "On large-batch training for deep learning: Generalization gap and sharp minima," *5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track Proc.*, pp. 1–16, 2017.