Downloaded from orbit.dtu.dk on: Apr 17, 2024

DTU Library

A deep transfer learning approach for sleep stage classification and sleep apnea detection using wrist-worn consumer sleep technologies

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

Published in:
IEEE Transactions on Biomedical Engineering

Link to article, DOI:
10.1109/TBME.2024.3378480

Publication date:
2024

Document Version
Peer reviewed version

Link back to DTU Orbit

Citation (APA):
Olsen, M., Zeitzer, J. M., Richardson, R. N., Musgrave, V. H., Sorensen, H. B. D., Mignot, E., & Jennum, P. J. (in press). A deep transfer learning approach for sleep stage classification and sleep apnea detection using wrist-worn consumer sleep technologies. *IEEE Transactions on Biomedical Engineering*. https://doi.org/10.1109/TBME.2024.3378480

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2024.3378480

A deep transfer learning approach for sleep stage classification and sleep apnea detection using wrist-worn consumer sleep technologies

Mads Olsen IEEE member, Jamie M. Zeitzer, Risa N. Richardson, Valerie H Musgrave, Helge B. D. Sørensen†* IEEE Senior member, Emmanuel Mignot*, Poul J. Jennum*

Abstract*— Obstructive sleep apnea (OSA) is a common, underdiagnosed sleep-related breathing disorder with serious health implications Objective - We propose a deep transfer learning approach for sleep stage classification and sleep apnea (SA) detection using wrist-worn consumer sleep technologies (CST). Methods – Our model is based on a deep convolutional neural network (DNN) utilizing accelerometers and photo-plethysmography signals from nocturnal recordings. The DNN was trained and tested on internal datasets that include raw data from clinical and wrist-worn devices; external validation was performed on a hold-out test dataset containing raw data from a wrist-worn CST. Results - Training on clinical data improves performance significantly, and feature enrichment through a sleep stage stream gives only minor improvements. Raw data input outperforms feature-based input in CST datasets. The system generalizes well but performs slightly worse on wearable device data compared to clinical data. However, it excels in detecting events during REM sleep and is associated with arousal and oxygen desaturation. We found; cases that were significantly underestimated were characterized by fewer of such event associations. Conclusion - This study showcases the potential of using CSTs as alternate screening solution for undiagnosed cases of OSA. Significance - This work is significant for its development of a deep transfer learning approach using wrist-worn consumer sleep technologies, offering comprehensive validation for data utilization, and learning techniques, ultimately improving sleep apnea detection across diverse devices.

***Index Terms*— deep learning, sleep stage classification, consumer sleep technologies, sleep disordered breathing.**

I. Introduction

Obstructive sleep apnea (OSA) is a highly prevalent sleep-related breathing disorder (SDB) associated with neurocognitive impairment, cardiovascular disease (CVD), and increased mortality [1]. It is estimated to affect almost one billion of individuals worldwide [2] and is increasingly prevalent due to its strong association with obesity [3], [4] and cardiovascular diseases [5]. OSA remains underdiagnosed, and up to 80% of cases remain untreated, contributing to its status as a hidden health crisis [6], [7]. The underdiagnosis of OSA can be attributed to patients frequently ignoring symptoms until the condition has advanced, a lack of awareness among individuals and professionals, and the complexity of diagnostic procedures and management [7]. The American Academy of Sleep Medicine (AASM) supports *"the adoption of more aggressive and comprehensive OSA diagnosis and treatment programs"* to mitigate the implications of undiagnosed OSA [8].

Wrist-worn consumer sleep technologies (CST) are now commonplace and can measure physical activity and vital signs with increasingly high resolution and quality [9]. Studies have demonstrated promise in detecting sleep stages using CSTs [10]–[18]. Leveraging CSTs as out-of-clinic (OOC) sleep monitoring systems, particularly for OSA screening, offers the opportunity for large-scale studies in the public. This also enables identification and potential treatment of OSA. Validation studies of wrist-worn CSTs against the gold standard are limited [19]–[21], despite promising results using clinical

The reports of Mads Olsen were supported in part by Augustinus Fonden, in part by Knud Højgaards Fond, in part by Marie and M.B. Richters Fond, in part by Danmark Amerika Fonden, in part by Reinholdt W. Jorck og Hustrus Fond, in part by Otto Mønsted Fonden, in part by Copenhagen Center for Health Technology, and in part by the Technical University of Denmark.

The work of STAGES was supported by Klarman Family Foundation. The work of TBI Group was supported by Patient-Centered Outcomes Research Institute (PCORI) Award under Grant CER-1511-33 005. Amazfit, a Zepp Inc. brand, funded the devices for this work through a contract to Emmanuel Mignot registered at clinical.gov under Grant NCT04429906.

*(Helge B. D. Sørensen, Emmanuel Mignot, and Poul J. Jennum contributed equally to this work.) (Corresponding author: Mads Olsen.)

Mads Olsen is with the Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94304 USA, and with the Biomedical Signal Processing & AI Research Group, Department of Health Technology, Technical University of Denmark, 2800 Kongens Lyngby, Denmark (e-mail: madsol@dtu.dk, somnio.ai@gmail.com).

Jamie M. Zeitzer, Valerie H Musgrave, and Emmanuel Mignot are with the Department of Psychiatry and Behavioral Sciences, Stanford University, USA.

†Helge B. D. Sørensen was with the Biomedical Signal Processing & AI Research Group, Department of Health Technology, Technical University of Denmark, Denmark. Helge B. D. Sørensen passed away Febuary 15th, 2023. This paper acknowledges his significant contributions, and the authors would like to honor his memory.

Risa Nakase-Richardson is with James A. Haley Veterans Hospital (Research) and the Division of Pulmonary and Sleep Medicine, Morsani College of Medicine, University of South Florida, USA.

Poul J. Jennum is with the Danish Center for Sleep Medicine, Department of Clinical Neurophysiology, Rigshospitalet, Denmark.

This article has supplementary material.

Authorized licensed use limited to: Technical University of Denmark DTU Library. Downloaded on April 04,2024 at 08:46:44 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2024.3378480

data from related modalities, e.g. electro cardiogram (ECG) [22], [23]. This may be because most CSTs rely on proprietary algorithms and lack external access to raw data. Hence, there is an unmet need to validate raw sensor data from wrist-worn CSTs against gold standard polysomnography (PSG) recordings [24].

Detecting OSA using CSTs is a challenging task for multiple reasons. Firstly, datasets from CSTs are scarce and typically small. Secondly, data from wrist-worn devices is prone to noise and data loss and lacks standardization, resulting in variable technical specifications and quality compared to data obtained using clinical equipment in controlled environment. Given the large amount of overnight data that can be collected with CSTs, however, the presence of low signal quality is not necessarily an impenetrable barrier. For example, Papini et al. (2020) showed that segments with low signal quality could be removed to yield a substantial increase in sleep apnea detection performance on data from wrist-worn CST [21]. Finally, various OSA phenotypes exist [25], and datasets may contain patients with varying degrees of OSA severity, resulting in different expressions of OSA events in the observed signals. Indeed, OSA can manifest both as complete (apnea) and partial (hypopnea) cessations of airflow. It is also just one pathophysiological subtype of SDB abnormalities that can occur; other less frequent subtypes such as central sleep apnea (CSA) and hypoventilation have different pathophysiologies [26]. Although colloquially, OSA is often equated with high apnea-hypopnea index (AHI) and is likely the dominant phenotype in almost all cases, the exact nature of events detected with these devices cannot be distinguished. In this study, the class "SA" encompasses all sleep apnea and hypopnea subtypes, including central, obstructive, and mixed events.

Studies targeting SA detection using data from CSTs can be categorized into three groups: scalar-, segmentation-, and event-based approaches. Scalar-based approaches predict an overall AHI using scalar values extracted from a CST and demographic descriptors [19], [27]. Segmentation-based approaches divide the recording into subsegments and predict the presence of SA events for each subsegment, usually using hundreds of handcrafted features within each subsegment [20], [21]. Event-based systems predict individual apneic events and rely on extracting surrogate signals from the input modalities [22], [28]. While segmentation-based approaches capture a more nuanced representation of the condition when compared to scalar-based approaches, operating with a fixed segment size does not comport well with the intrinsic nature of apnea events that have different durations [26]. Furthermore, the existing approaches rely on extensive signal preprocessing, which may accumulate errors and risk removing important discriminative information from the underlying signal.

Deep learning (DL) systems are universal approximators that can learn optimal features for a given task directly from raw data [29]. However, DL-based systems usually require extensive amounts of data to ensure high precision [30], which constitutes a problem for these small CST datasets. Studies have addressed this limitation by employing a transfer learning paradigm, where classifiers are pretrained on abundant, clinical datasets using signal modalities that are related to those of the target dataset. Kotzen et al. (2022) and Radha et al. (2021) successfully enhanced performance on a PPG-based sleep stage prediction task by pretraining their systems using ECG from clinical studies [17], [31]. Defining auxiliary learning tasks in a dataset has been another successful strategy to improve the performance of the main objective. Studies have demonstrated that including discriminative information through auxiliary prediction tasks, such as sleep stage prediction, can enhance SA detection performance [21].

In this study we present a DL model-framework for the detection of both SA events and sleep stages in nocturnal recordings. The system learns from minimal processed accelerometer (ACC) and photoplethysmography (PPG) signal modalities. Our model is pretrained on a large clinical cohort and was applied to both clinical- and CSTs datasets. In the experimental section we evaluated the impact of transfer learning and feature enrichment through the auxiliary sleep stage learning task. Furthermore, we compare the performance of our system to state-of-the-art (SOTA) works, both, including feature-based systems and established systems for alternative SA detection, hereunder the WatchPAT and the Belun sleep platform [27], [32]. Finally, we validate model performance on an external test set from a wrist-worn CST to assess model generalizability.

II. Material and Methods

Our proposed model is presented in Fig. 1. The model has two equivalent streams that work on two different objectives, sleep stage classification and SA detection, respectively. Both streams contain a deep convolutional neural network architecture, inspired by U-Net [33], DeepSleep [34] and U-Sleep [35], that was presented in our previous work [18]. This architecture constitutes the temporal feature extractor that serves to learn feature maps at different scales across the entire input segment. Sleep stage information is incorporated into the SA stream by concatenating feature maps from the sleep stage stream and the SA stream as shown in Fig. 1. Finally, the feature maps from each model stream are segmented and classified into sleep stage and SA vectors, respectively, which constitute the output predictions from the DNN.

A. Deep neural network architecture

A complete presentation of the proposed DNN architecture is presented in the supplementary material. The feature maps from each model stream are processed by two similar segmentation classifiers. Both segmentation classifiers apply cropping and reshaping operations to remove zeros that were padded and to reshape the feature maps into a 2D vector. Then, these vectors are segmented into the desired output resolution by a temporal average pool operator. A non-linear GELU activation function is applied before and after the average pool operator using a $(1,1)$ convolutional layer. Finally, a softmax

Authorized licensed use limited to: Technical University of Denmark DTU Library. Downloaded on April 04,2024 at 08:46:44 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2024.3378480

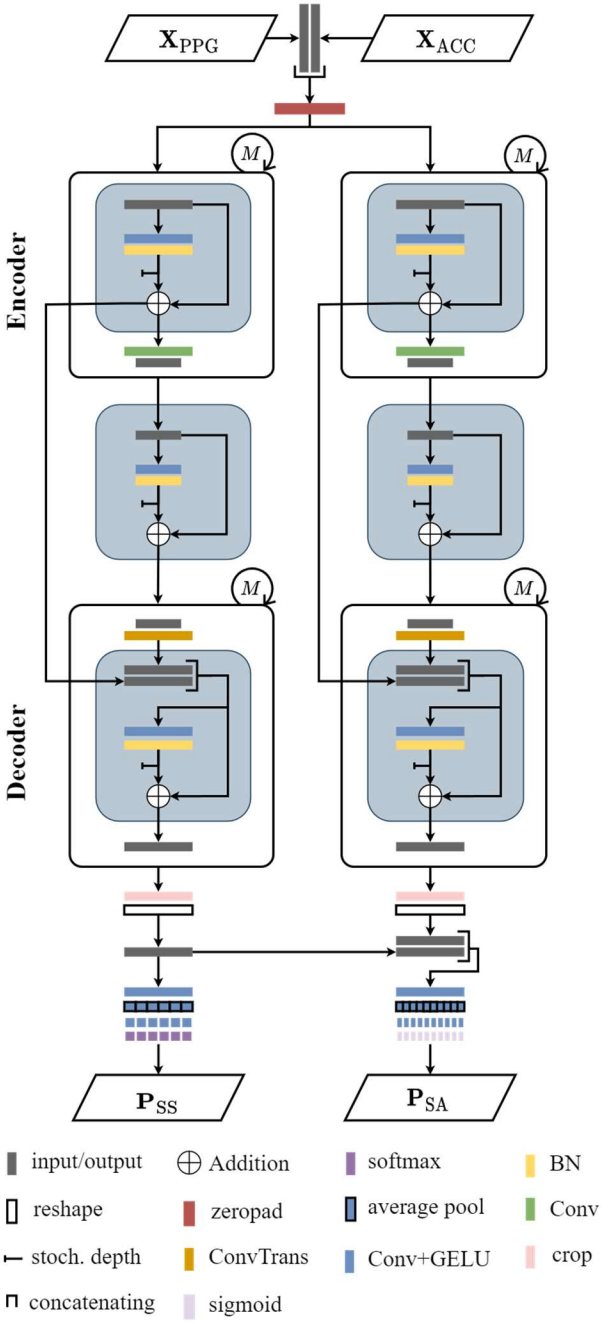

Fig. 1. Conceptual visualization of the proposed deep neural network (DNN). Two time-aligned time series: $\boldsymbol{X}_{PPG} \in \mathbb{R}^{32N\times1\times1}$ and $\boldsymbol{X}_{ACC} \in \mathbb{R}^{32N\times1\times3}$ are processed in two equivalent streams that work on two different objectives, namely sleep stage classification and apnea classification, respectively. Both streams are comprised of a deep convolutional neural network architecture with a residual convolutional block structure, inspired by U-Net [33], DeepSleep [34] and U-Sleep [35], that serves to learn feature maps at different scales across the entire input segment. Sleep stage information is incorporated into the SA stream by concatenating feature maps from the sleep stage stream into the SA steam. Finally, the feature maps from each model stream are segmented and classified into sleep stage: $\boldsymbol{P}_{SS} \in \mathbb{R}^{N/30\times4}$ and apnea: $\boldsymbol{P}_{SA} \in \mathbb{R}^{N\times1}$ vectors, respectively, which constitute the output predictions from the DNN. Each model stream was trained separately. T_{SS}: number of output timesteps for the sleep stage stream; N: duration in seconds of the recording; M: number of encoder and decoder blocks; stoch depth: Stochastic depth [48]; GELU: Gaussian Error Linear Unit activation function [49], conv: convolution, convTranspose: transposed convolutional; BN: batch normalization [50]. ACC: Accelerometry; PPG: Photoplethysmography; SA: Sleep apnea

and a sigmoid layer classifies each timestep in the final feature vectors into output vectors comprising sleep stage and SA predictions, respectively.

The sleep stage segmentation classifier, $\varphi_{\mathrm{SS}}: \mathbb{R}^{\tilde{T}_S\times\tilde{F}_S\times C_U} \rightarrow \mathbb{R}^{T_{SS}\times4}$, where T_{SS} is the number of output timesteps for the sleep stage stream, processes the feature maps into a vector, $\mathbf{P}_{SS} \in [0,1]^{T_{SS}\times4}$, that for each sleep epoch assigns a probability for each of the four sleep stages: wake, light sleep, deep sleep, and REM sleep. We classify sleep into 4 distinct categories, aligning with our previous research [18]. These categories include wakefulness (W), light sleep (L/N1 and N2 combined), deep sleep (D/N3), and rapid eye movement (R/REM) sleep. Our classification differs from the conventional AASM scoring system as we employ proxy modalities that indirectly capture the gold standard sleep signal, necessitating this adjustment in sleep staging categories. The SA segmentation classifier: $\varphi_{\mathrm{SA}}: \mathbb{R}^{\tilde{T}_S\times\tilde{F}_S\times C_U} \rightarrow \mathbb{R}^{T_{\mathrm{SA}}\times1}$, where T_{SA} is the number of output timesteps for the SA stream, processes the feature maps into a vector, $\mathbf{P}_{\mathrm{SA}} \in [0,1]^{T_{\mathrm{SA}}\times1}$, that for each second assigns a probability for the binary classification problem: no-SA or SA. The SA class contains all SA subtypes (i.e., central, obstructive, hypopnea, mixed events).

B. Data

Data used in our experiments come from the Stanford Technological Analytics and Genomic in Sleep (STAGES) study, the Traumatic Brain Injury (TBI) study, the Amazfit Health (Health) study, and the Multi-Ethnic Study of Atherosclerosis (MESA) sleep study. An overview these cohorts, including demographic information, sleep-related metrics, and data origin source, is presented in Table I. Information about recruitment, inclusion and exclusion criteria, technical details of the data collected, and data scoring guidelines is descripted in supplementary material and in previous publications [18], [36].

C. Preprocessing

Data from all datasets were processed using the following initial preprocessing steps. Both ACC and PPG were interpolated to a uniform time series with a sampling rate of 32 Hz. Signals 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 7.4%, 1.3%, 5.3%, and 2.1% of the recording time for the STAGES, TBI, Health, MESA dataset, respectively.

Minimal denoising and normalization steps were applied to both input modalities to account for the inter- and intra-variations that exist between datasets. SA manifestation is presented for signal segment samples for each modality in the supplementary material.

For the ACC signal, $\mathbf{X}_{\mathrm{ACC}} \in \mathbb{R}^{32N\times1\times3}$, each directional vector was processed by a total variation filter (TV) followed by a differential operator to denoise the signal and to remove baseline wander, following the procedure proposed by Chen et al. (2021) [28]. Then, each directional vector was normalized to

Authorized licensed use limited to: Technical University of Denmark DTU Library. Downloaded on April 04,2024 at 08:46:44 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2024.3378480

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

Cohort	N	AHI, $\mu \pm \sigma$	Age, $\frac{\mu \pm \sigma}{\text{years}}$	BMI, $\frac{\mu \pm \sigma}{\text{kg/m}^2}$	Gender, % Male	Modality source		Sleep apnea detector			
						ACC (Sample rate)	PPG (Sample rate)	Warmup	Train	Test	External Test
MESA	1355	34.5±21.3	69.1±8.9	28.7±5.2	45.5	-	PSG (256 Hz)	1068		271	
TBI	231	17.6±20.2	38.4±20.6	26.2±5.2	81.4	GT3X (100 Hz triaxial)	PSG (100 Hz)		185	46	
STAGES PSG	35	13.1±10.6	38.3±13.6	29.3±8.5	45.7	Amazfit Arc (triaxial 25 Hz)	PSG (128 Hz)		18	17	
STAGES ARC	35	13.1±10.6	38.3±13.6	29.3±8.5	45.7	Amazfit Arc (triaxial 25 Hz)	Amazfit Arc (25 Hz)		18	17	
Health PSG	35	16.0±24.0	36.2±13.6	28.6±7.8	40.0	Amazfit Health (triaxial 25 Hz)	PSG (128 Hz)				35
Health	35	16.0±24.0	36.2±13.6	28.6±7.8	40.0	Amazfit Health (triaxial 25 Hz)	Amazfit Health (50 Hz)				35
Total	1726							1068	221	351	70

μ: mean; σ: standard deviation; PSG: Polysomnography; AHI: Apnea-Hypopnea Index; ArI: Arousal Index; STAGES: Stanford Technological Analytics and Genomics in Sleep study; TBI: Traumatic Brain Injury study. MESA: Multi-Ethnic Study of Atherosclerosis; Health: Amazfit Health Study.

have a median of zero and an inter-quartile range (IQR) between -1 and 1.

The PPG signal, $\mathbf{X}_{\text{PPG}} \in \mathbb{R}^{32N\times1\times1}$, was initially bandpass filtered with a passband frequency range of $[0.1, 8]$ Hz. Then, an adaptive version of the IQR normalization method was implemented where quartiles are calculated by sliding window size of 300 s, to account for the non-ergodic behavior of the PPG modality, as presented in our earlier work. Outliers outside 20 times the IQR-range was clipped.

The presented approach, that learns from raw, normalized data, was benchmarked against a PPG surrogate-based approach, that require a sequence of preprocessing steps similar to our previously presented approach [37]. The PPG surrogate signals, $\mathbf{X}_{\text{PPG surrogate}} \in \mathbb{R}^{4N\times1\times2}$, were extracted from the PPG pulse peaks, which were found by adaptive pulse segmentation [38]. For each pulse peak, the amplitude modulation (AM, i.e., the peak amplitude), the frequency modulation (FM, i.e., duration interval between consecutive beats) were extracted. These surrogate signals were interpolated using PCHIP and resampled to 4 Hz. Please refer to supplementary Fig. 1 for signal examples during apnea.

D. Training of the model streams

Each dataset was partitioned into a training set and a test set as reported in Table I. The training set was further partitioned into an evaluation set and a training set consisting of 30 % and 70 % of the recordings, respectively. 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 selected from the class set {W, L, D, R} or {non-SA, SA}, depending on which model stream was trained. Then, an input segment was randomly sampled with size T_S and with starting point $\{1, \dots, T_r - T_S\}$, where T_S is the input segment size and T_r is the duration of recording r. The segment was repeatedly sampled until it contained at least 1 representation of the selected class. Segments from the evaluation- and test sets were sampled in an ordered manner, such that each sleep epoch from each recording was evaluated only once.

Each model stream was trained separately. First, the sleep stage model stream was trained to allow sleep stage information (i.e., feature maps from the sleep stage stream) to be fed as input to the SA model stream. Furthermore, both model streams were trained using a transfer learning approach. Here, each model was initially trained on the MESA dataset and in turn finetuned by retraining the model on the remaining internal datasets, i.e., STAGES and TBI (see Table I). The MESA dataset did not include ACC at high resolution, therefore, a decoy vector containing random samples from a normal distribution with zero mean and unit variance, i.e., $X \sim \mathcal{N}(0, 1)$, was inputted as replacement for the ACC channel to allow ACC to be included during fine-tuning. Finally, input-channel dropout with a probability of 10% was added during finetuning to ensure the model did not consider the impact of ACC redundant.

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

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

Where $\mathcal{L}$ is the average loss for the given batch, n_k is the number of samples of class k in batch b, β=0.999, and T is the output segment length, which was T_{SS} and T_{SA} for the sleep stage and SA model streams, respectively. The output resolution of the sleep stage model stream was 30 s, such that one prediction was assigned to each sleep epoch of 30 s duration. The output resolution of the SA model stream was 1 s. $Y_{btk}\log(P_{btk})$ is the categorical cross entropy that induce exponential penalty to the loss function the further away the prediction $\mathbf{P}$ is from the target $\mathbf{Y}$. Binary cross entropy was used for the SA model stream. $(1-\beta)/(1-\beta^{n_k})$ is a balancing factor that accounts for class imbalance. It is based on the idea

Authorized licensed use limited to: Technical University of Denmark DTU Library. Downloaded on April 04,2024 at 08:46:44 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2024.3378480

that as the number of samples for a given class increases, the benefit of additional data points diminishes [39].

For both model streams the loss presented in (1) was computed for each batch and was minimized using the ADAM [40] 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). The learning rate was initiated to 10^{-4} during finetuning. Epochs with more than 50% missing data, e.g., during Bluetooth fallouts, 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 [41]. 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. Please refer to the supplementary material for model learning curves.

E. Event formation

The proposed CNN is capable of processing signal segments of variable duration. This made it possible to process entire recordings without having to subsegment them into the input segment size the model was trained with. Model predictions were computed for each recording and model stream:

$$f_{\mathrm{SS}}: \mathbf{X}^{(s)} \in \mathbb{R}^{32N\times 1\times 4} \rightarrow \mathbf{P}_{\mathrm{SS}} \in [0,1]^{\mathrm{N}/30\times 4}$$
$$f_{\mathrm{SA}}: \mathbf{X}^{(s)} \in \mathbb{R}^{32\ \times 1\times 4} \rightarrow \mathbf{P}_{\mathrm{SA}} \in [0,1]^{N\times 1}$$

where N is the recording duration in seconds for the sleep stages and SA model streams, respectively. For each recording, the output sequence from the SA model stream, $\boldsymbol{P}_{\mathrm{SA}} \in \mathbb{R}^{N\times 1}$, was processed into a set of SA event predictions by applying morphological opening and closing operations, given by:

$$\text{Opening: } (\mathbf{P}_{\mathrm{SA}} > \theta) \circ A = ((\mathbf{P}_{\mathrm{SA}} > \theta) \ominus A) \oplus A$$
$$\text{Closing: } (\mathbf{P}_{\mathrm{SA}} > \theta) \bullet A = ((\mathbf{P}_{\mathrm{SA}} > \theta) \oplus A) \ominus A$$

Where θ is the SA classification threshold, $\ominus$ and $\oplus$ are morphological erosion and dilation, respectively, and A is a 1D morphological structuring element that controls the dilated and eroded area. The threshold, θ, and the shape of A were found by identifying the set of parameters, found through grid search, that optimized F1-score performance on the evaluation set. In practice, this meant that the output predictions were processed into a binary vector, $\mathbf{P}_{\mathrm{SA}} \in [0,1]^{N\times 1} \rightarrow \{0,1\}^{N\times 1}$ by applying the threshold, θ. Here, consecutive 1's conceptualizes SA events. Finally, events were removed if they were shorter than A s, and events separated by less than A s were merged. This process is illustrated in the supplementary material.

F. Performance metrics

To quantify the overlap between two events, a commonly used metric is the intersection over union (IoU) [42]. In IoU, a predicted event is considered a true positive if it exhibited an IoU $> \delta$ with a true event, and otherwise it is considered a false positive, where δ is the overlapping criterion. Likewise, a true event was considered as a false negative if it did not exhibit an IoU $> \delta$ with a predicted event. When δ is small a predicted event which exhibits at least some small overlap with a true event might be considered a true positive, and vice versa. From these, recall: $\mathrm{Re} = \mathrm{N_{TP}}/(\mathrm{N_{TP}} + \mathrm{N_{FN}})$, precision: $\mathrm{Pr} = \mathrm{N_{TP}}/(\mathrm{N_{TP}} + \mathrm{N_{FP}})$, and F1-score: $\mathrm{F1} = 2(\mathrm{Re}\cdot\mathrm{Pr})/(\mathrm{Re} + \mathrm{Pr})$ were computed for the overlapping criterion $\delta = 0.1$, where $\mathrm{N_{TP}}, \mathrm{N_{FN}}$, and $\mathrm{N_{FP}}$ refer to number of true positives, false negatives, and false positives, respectively. Furthermore, the AHI was computed as the number of events divided by total sleep time (TST). Here, TST was computed as the aggregated time spent in non-Wake predicted sleep stages. Models that do not predict sleep stages rely on manually scored sleep stages. The correlation between the predicted and the true AHI was assessed both by Spearman's correlation, ρ. Finally, recordings were categorized into the following AHI severity groups: none, mild, moderate, and severe, which corresponds to AHI<5, 5≤AHI<15, 15≤AHI<30, 30≤AHI, respectively. Here, the linearly weighted Cohen's κ was used to measure the agreement between the true and predicted severity group, and the accuracy was reported for the binary classification tasks of categorizing recordings by these thresholds: AHI<5, AHI<15, and AHI<30.

G. Context analysis of apnea events

To investigate our model's performance within different contexts, SA events were categorized according to their subtype (i.e., OSA, CSA, and hypopnea) and dependent on what context they were appearing in, hereunder sleep stage context (i.e., wake, light, deep, and REM sleep), associated events (i.e., oxygen desaturations and arousal events), and by their duration. Criteria were defined sleep stages such that SA events were only assigned to one sub-context group. Specifically, an SA event was assigned to the sleep stage it originates from. Formally written, the difference in start time between the associated sleep stage, $t_{\mathrm{SS},0}$, and the SA event, $t_{\mathrm{SA},0}$, must be: $\Delta t_0 = t_{\mathrm{SS},0} - t_{\mathrm{SA},0} < 0$, and their $IoU(t_{\mathrm{SS}}, t_{\mathrm{SA}}) > 0$. Next, an SA event was assigned to an associated event if the associated event occurred because of/following the SA event. Formally, the difference in start time between the associated sleep stage, $t_{\mathrm{AE},0}$, and the SA event, $t_{\mathrm{SA},0}$, must be: $\Delta t_0 = t_{\mathrm{AE},0} - t_{\mathrm{SA},0} > 0$, and their $IoU(t_{\mathrm{AE}}, t_{\mathrm{SA}}) > 0$. In addition, the SA event duration was extended with 5 s to capture associated events that are annotated immediately after the SA event ends. Furthermore, SA events were categorized by their subtype, t_{sub}. Formally, $IoU(t_{\mathrm{sub}}, t_{\mathrm{SA}}) > 0$. Finally, SA events were partitioned into three groups by their duration, given by: $t_{\mathrm{SA}} < 15$, $15 \le t_{\mathrm{SA}} < 30$, $30 \le t_{\mathrm{SA}}$.

These groups were further categorized by adopting the under- and overestimation groups from Papini et al. (2020), which outlines the recordings that are misclassified to a severe degree [21]:

$$\text{Considerably underestimation: } \begin{cases} \mathrm{AHI_{pre}} < \frac{1}{2}\mathrm{AHI} - 2.5, & 5 \le \mathrm{AHI} < 15 \\ \mathrm{AHI_{pre}} < \frac{2}{3}\mathrm{AHI} - 5, & \mathrm{AHI} \ge 15 \end{cases}$$
$$\text{Considerably overestimation: } \begin{cases} \mathrm{AHI_{pre}} > 2\mathrm{AHI} + 5, & 0 \le \mathrm{AHI} < 5 \\ \mathrm{AHI_{pre}} > \frac{3}{2}\mathrm{AHI} + 7.5, & \mathrm{AHI} \ge 5 \end{cases}$$

Authorized licensed use limited to: Technical University of Denmark DTU Library. Downloaded on April 04,2024 at 08:46:44 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2024.3378480

TABLE II
OVERALL PERFORMANCE FOR THE PROPOSED APPROACH ON THE 6 TEST DATASETS AND PERFORMANCE OF RELATED STATE-OF-THE-ART WORKS

Dataset train/test	Model/System	SS. aux	w.u.	Input	Train/ Test	F1 score (by event)	AHI metrics					
							Cohen's κ (weighted)	ρ	p-value	Cohen's κ / AUC ROC AHI≥5	AHI≥15	AHI≥30
Warmup MESA	Res-U-Net	✓		PPG (PSG) raw	1084/ 242	0.63 ± 0.16	0.55	0.76	<0.001	0.00/ 0.50	0.57/ 0.79	0.48/ 0.74
Internal TBI	Res-U-Net	✓	✓	ACC (wrist) raw PPG (PSG) raw	185/ 46	0.44 ± 0.24	0.74	0.78	<0.001	0.61/ 0.80	0.87/ 0.97	0.73/ 0.81
Internal STAGES PSG	Res-U-Net	✓	✓	ACC (wrist) raw PPG (PSG) raw	18/17	0.37 ± 0.21	0.56	0.73	0.001	0.00/ 0.50	0.75/ 0.86	1.00/ 1.00
Internal STAGES Arc	Res-U-Net	✓	✓	ACC (wrist) raw PPG (wrist) raw	18/17	0.29 ± 0.17	0.42	0.20	0.343	0.20/ 0.59	0.20/ 0.59	0.00/ 0.50
External Health PSG	Res-U-Net	✓	✓	ACC (wrist) raw PPG (PSG) raw	0/35	0.42 ± 0.24	0.64	0.75	<0.001	0.68/ 0.84	0.66/ 0.86	0.64/ 0.75
External Health	Res-U-Net	✓	✓	ACC (wrist) raw PPG (wrist) raw	0/35	0.32 ± 0.21	0.57	0.62	<0.001	0.30/ 0.63	0.61/ 0.79	0.64/ 0.75
SOMNIA	Papini [21]	✓		ACC (wrist) features PPG (wrist) features	250/ 188	0.44*	0.51	0.67		0.39/ 0.80	0.51/ 0.82	0.49/ 0.84
	Belun [27]	✓	✓	ACC (ring) features PPG (ring) features	8417/ 79	-	0.42	-	-			
	WatchPAT [32]	✓	✓	ACC, PAT, pulse rate, SpO2, snoring	0/ 500		0.53	-				

* Epoch-based detection (the rest are event-based)
$\mu \pm \sigma$ (average and standard deviation); ρ: Spearman's correlation; AUC ROC: Area under the receiver operator curve; Res-U-Net: The presented residual U-Net model; SS. Aux: Added model output from auxiliary sleep stage model stream; w.u.: Warm-up – referring to model pretraining on a clinical dataset; ACC: Accelerometry, PPG: Photoplethysmography; PAT: Peripheral arterial tonometry; AHI: Apnea-hypopnea index; MESA: Multi-Ethnic Study of Atherosclerosis sleep; TBI: Traumatic Brain Injury study. Health: Amazfit Health study; STAGES: Stanford Technological Analytics and Genomics in Sleep study.

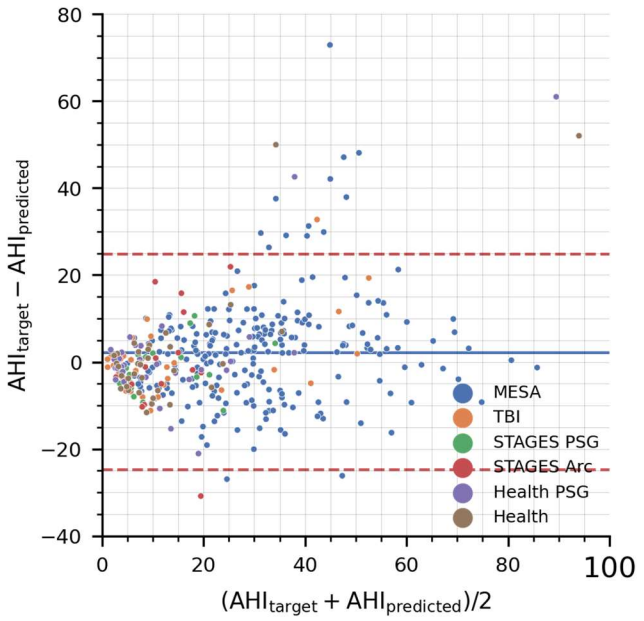

Fig. 2 – Bland Altman plot between target AHI and predicted AHI. Red dashed and solid blue lines represents the mean and the standard deviation, respectively. MESA: Multi-Ethnic Study of Atherosclerosis; TBI: Traumatic Brain Injury study. STAGES: Stanford Technological Analytics and Genomics in Sleep study; Health: Amazfit Health study. PSG: Polysomnography (PSG-derived PPG signal).

III. EXPERIMENTS

A. Deep neural network parameter selection

The following parameters were used for our proposed DNN: An input segment size of 256 sleep epochs, corresponding to 128 min, a batch size of $B = 8$, a kernel size, $K = 9$, a filter width, $C_U = 8$, and network depth, $M = 11$. Please refer to our previous study that presents a comprehensive experimental section to determine the optimal set of parameters for the proposed DNN [18]. Furthermore, determined through grid search, the best performing post processing parameters were determined to be $\theta = 0.42$ and $A = 6$, presented in the supplementary material.

B. Benchmark

Both event-based and recording-based performance of the proposed approach are presented in Table II for all test datasets used in this study, and the corresponding Bland Altman plot between target AHI and predicted AHI are presented in Fig. 2. Competing approaches are shown for comparison. Comparison between works on databases with wearable data indicate that the proposed approach has similar or better performance to that of the best performing SOTA works. The proposed approach achieved $\kappa = 0.57$ and $\rho = 0.62$ ($p < 0.001$) on the external, wearable dataset (Health) using raw input, whereas Papini et al. (2020) achieved $\kappa = 0.51$ and $\rho = 0.67$(p<0.01) [21] using processed features on the SOMNIA dataset. Similarly, the proposed approach show similar or better performance when compared to SOTA works that use established, wearable system, hereunder the Belun Sleep platform [27], $\kappa = 0.42$ and WatchPAT [32], $\kappa = 0.53$.

Authorized licensed use limited to: Technical University of Denmark DTU Library. Downloaded on April 04,2024 at 08:46:44 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2024.3378480

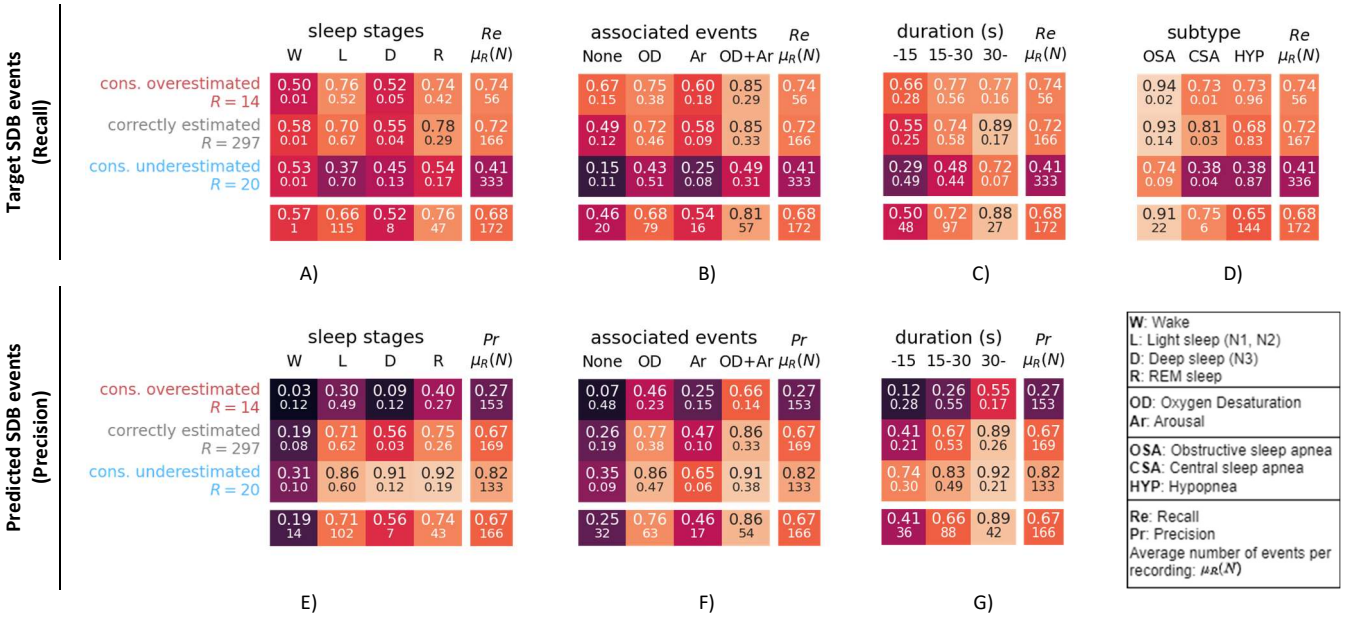

Fig. 4 – Context analysis of scored and predicted SA events grouped by AHI estimation groups adapted from [21], representing correctly (gray) estimated cases, and considerably over- (red) and underestimated (blue) cases. Upper: scored events grouped by A) sleep stages, B) associated events, C) duration, and D) SA subtype. Upper cell value: recall of events within context group. Lower cell value: Proportion of events within context group (each context sums to 1). Marginalized lower cell value: Subject-average number of scored events marginalized over all context groups (right column) or AHI estimation groups (lower row). Lower: predicted events grouped by E) sleep stages, F) associated events, and G) duration. Upper cell value: precision of events within context group. Lower cell value: Same as above. Only datasets that use PSG-derived PPG. These are MESA, TBI, STAGES PSG, and Health PSG.

The STAGES and the Health study both had PPG recordings from the wearable device and from the overlapping PSG recording. The model was tested on both PPG sources independently to assess the importance of the PPG source. The F1-score performance metric dropped from 0.42 ± 0.24 to 0.32 ± 0.21 for the Health study and from 0.37 ± 0.21 to 0.29 ± 0.17 for the STAGES study, when using the PPG signal recorded with the wearable devices when compared to using the PSG-derived PPG signal (Health PSG).

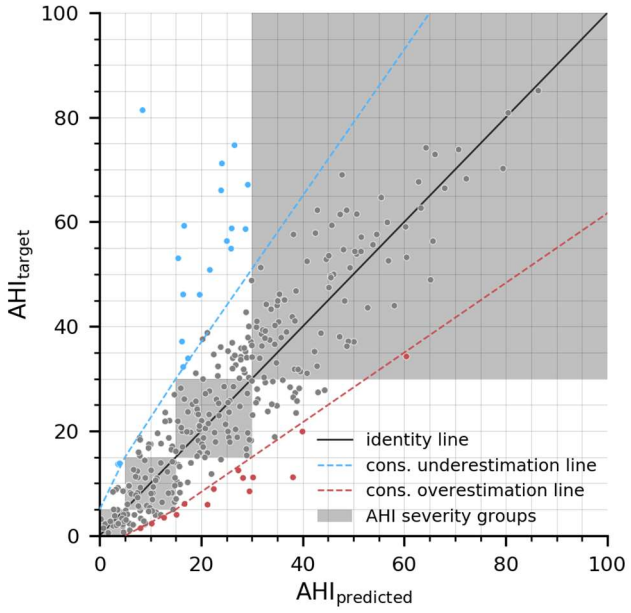

Fig. 32 – Correlation between target AHI and predicted AHI. The correlation between the predicted and the true AHI across test datasets used in the event context analysis. These are MESA, TBI, STAGES PSG, and Health PSG. The AHI severity groups (gray boxes) assign agreement between the true and predicted AHI severity groups, given by none, mild, moderate, and severe, which corresponds to AHI<5, 5≤AHI<15, 15≤AHI<30, 30≤AHI, , respectively. Red and blue dashed lines represents considerable over- and underestimation lines adapted from [21]. TBI: Traumatic Brain Injury study. STAGES: Stanford Technological Analytics and Genomics in Sleep study; Health: Amazfit Health study. PSG: Polysomnography (PSG-derived PPG signal).

Finally, the model's performance appears to remain stable when applied to the external cohort, as observed when comparing it with the warm-up and internal cohorts. In the warm-up and internal cohorts, Spearman's correlation, ρ, was 0.76($p<0.001$), 0.78 ($p<0.001$), and 0.73 ($p<0.001$), respectively, while in the external cohort, it was 0.75 ($p<0.001$) for the non-CST datasets.

C. *Model strengths and weaknesses*

To identify the strengths and weaknesses of our model, we investigate how well the model predicts SA events in different contexts. Fig. 3 presents the correlation between the target AHI and the predicted AHI categorized into AHI estimation groups adapted from [21]. Only datasets that use PSG-derived PPG and that have ACC were included in this analysis. Please refer to the supplementary material for a by-dataset presentation. These are TBI, STAGES PSG, and Health PSG. Fig. 4 (upper) presents proportion of scored events and the performance recall within each of the following contexts: A) sleep stages, B) associated events, C) event duration, and D) SA subtypes for each AHI estimation group. Likewise, Fig. 4 (lower) shows the proportion of predicted events and the performance precision within contexts: E) sleep stages, F) associated events, and G) event duration for each AHI estimation group.

Investigation of the marginal performance across all AHI estimation groups, Fig. 4, shows that our model performs better during REM sleep, $Re = 0.76$ and $Pr = 0.74$, when compared to the other sleep stages. The performance is substantially lower during deep sleep, $Re = 0.52$ and $Pr = 0.56$. Furthermore, the model is more sensitive to SA events if they are associated with

Authorized licensed use limited to: Technical University of Denmark DTU Library. Downloaded on April 04,2024 at 08:46:44 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2024.3378480

TABLE III
ABLATION STUDY - PERFORMANCE FOR DIFFERENT CONFIGURATIONS OF THE PROPOSED APPROACH ON THE 3 TEST DATASET GROUPS

Model	SS. aux*	Warm-up	Input		Warmup test MESA		Internal test TBI, STAGES PSG		External test Health	
			ACC	PPG	F1 Event-based	Cohen's κ (weighted) AHI-based	F1 Event-based	Cohen's κ (weighted) AHI-based	F1 Event-based	Cohen's κ (weighted) AHI-based
Res-U-Net	✓	✓	Raw	Raw	**0.67**	0.55*	**0.56**	0.66*	**0.42**	**0.57***
Res-U-Net		✓	Raw	Raw	**0.66**	0.50	0.55	0.69	0.41	0.50
Res-U-Net	✓		Raw	Raw			0.50	0.68*	0.36	0.42*
Res-U-Net			Raw	Raw			0.52	**0.77**	0.33	0.41
U-Net		✓	Raw	Raw	**0.65**	0.48	0.53	0.65	0.37	0.42
U-Net			Raw	Raw			0.48	0.61	0.26	0.37
Res-U-Net		✓		Raw			0.54	0.75	0.31	0.49
Res-U-Net				Raw			0.46	0.61	0.28	0.35
Res-U-Net			Raw				0.16	0.2	0.20	0.34
Res-U-Net		✓		Surrogate	**0.69**	0.57	0.51	0.61	0.34	0.42
Res-U-Net				Surrogate			0.44	0.58	0.27	0.36
GRU [22]		✓		Surrogate	**0.68**	**0.58**	0.50	0.69	0.36	0.38
GRU [22]				Surrogate			0.46	0.56	0.30	0.30
Papini [21]	✓	✓	Features	Features	**0.60**	0.48	0.51	0.51	0.35	0.38

* TST was computed automatically from the sleep stage model predictions. Best performing inputs are highlighted in bold font for each dataset. Please refer to the supplementary material for significance analysis. Res-U-Net: The presented residual U-Net model; U-Net: The presented U-Net model without residual connections; GRU: Gated recurrent unit [22]; SS. Aux: Added model output from auxiliary sleep stage model stream; Warm-up – referring to model pretraining on the MESA dataset; ACC: Accelerometry, PPG: Photoplethysmography; AHI: Apnea-Hypopnea index; MESA: Multi-Ethnic Study of Atherosclerosis sleep. N.B. MESA does not have an ACC modality; STAGES: Stanford Technological Analytics and Genomics in Sleep study; TBI: Traumatic Brain Injury study. Health: Amazfit Health study.

an oxygen desaturation, $Re = 0.68$, or an arousal event, $Re = 0.54$, or both, $Re = 0.81$, compared to events with no such association, $Re = 0.46$. A similar tendency is found for the model precision. The model is more sensitive and more precise to longer SA events, as it scores $Re = 0.88$ and $Pr = 0.89$ for events longer than 30 s compared to $Re = 0.50$ and $Pr = 0.41$ for events shorter than 15 s. Finally, our model was more sensitive to OSA, $Re = 0.91$, and CSA, $Re = 0.75$, events when compared to hypopnea events, $Re = 0.65$. This has a large impact on the overall performance, as most breathing events in the datasets are hypopneas. For comparison the interrater aggrement for each apnea subgroup is: OSA: 0.77, hypopneas: 0.65, and CSA: 0.52 [51].

The AHI group that is considerably underestimated is characterized by having many false negatives and consequently a low recall. The 20 recordings that were considerably underestimated had many short hypopnea events, 336 events on average, when compared to the correctly estimated group, 167 events on average. The underestimated group had a higher proportion of events during deep sleep, 13 %, and a small proportion of events during REM sleep, 17 %, compared to the other groups, 4-5 %, and 29-42 %, respectively.

Contrarily, the 14 recordings that were considerably overestimated were characterized by having many false positives and consequently a low precision. While this group had a low precision of $Pr = 0.27$, 52 % of predicted events were associated with oxygen desaturation or arousal, but only a fraction of these events was associated with a scored event. Troublingly, 48 % of the events were predicted without any associated event context. This is considerably higher than that of the other groups.

D. Ablation study

In this section, we conduct an ablation study to assess the impact of various modifications on the model's performance. Each line in Table III corresponds to a separate experiment, with each model trained from scratch. Detailed results are available in the supplementary material. The following experiments were conducted:

Architecture importance: We tested the significance of architecture by comparing the model with and without residual connections. Adding the residual connection improved the F1-score from 0.65, 0.53, and 0.37 to 0.66, 0.55, and 0.41 on the warm-up, internal, and external wearable test sets, respectively.

Auxiliary features: We added auxiliary features from a complementary network trained to classify sleep stages, which slightly improved the F1-score from 0.66, 0.53, and 0.41 to 0.67, 0.55, and 0.42 on the warm-up, internal, and external wearable test sets, respectively.

Pretraining: We assessed the importance of pretraining the model on a large clinical dataset, the MESA database containing only PPG data. This considerably improved performance from 0.50 to 0.55 on the internal wearable test set and from 0.36 to 0.42 on the external wearable test set. Similar improvements were observed for sleep stage prediction.

Modality importance: We examined the importance of individual modalities by inputting them separately. Using both ACC and PPG yielded the best performance. While adding ACC only slightly improved performance on the internal dataset, it notably improved the F1-score from 0.31 to 0.42 on the external wearable test set.

Feature-based vs. raw Data: Comparing feature-based approaches with training on raw data, we found that preprocessing input modalities into surrogate signals produced the best F1-score on the warm-up and external wearable test sets. Contrarily, the Res-U-Net model trained with raw signals performed best on the internal test set.

Comparison with related work: Our experiments indicate that the presented approach outperforms the implemented version of the approach by Papini et al. (2020) [21].

Authorized licensed use limited to: Technical University of Denmark DTU Library. Downloaded on April 04,2024 at 08:46:44 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2024.3378480

IV. Discussion

A. Model validation on non-CST datasets

It was chosen to use AHI-based metrics rather than event-based metrics when comparing across datasets, since the F1-score is very sensitive to class imbalance, as referenced in [43]. Notably, the MESA database's significantly higher performance is associated with its double average AHI compared to other datasets. Though, direct comparison between databases is only indicative, due to natural variability between databases. Our overall findings indicate that the predicted AHI aligns well with the target AHI, albeit with some moderate variance. In most cases, this variance only results in the model misclassifying the AHI by one severity group, as shown in supplementary Fig. 5. The proposed residual U-Net type DNN model's performance in the external cohort demonstrates stability when compared to its performance in both the warm-up and internal cohorts. In the warm-up and internal cohorts, Spearman's correlation, ρ, was observed to be 0.76 ($p<0.001$), 0.78 ($p<0.001$), and 0.73 ($p<0.001$), respectively. When applied to the external cohort, the model achieved a Spearman's correlation of 0.75 ($p<0.001$) for the non-CST datasets. This comparison suggests that the model's performance remains consistent when extended to the external dataset.

B. Addressing the performance drop on wearable CST data

The F1-score performance metric dropped substantially when using the PPG signal recorded with wearable devices when compared to using the PSG-derived PPG signal. This drop that we hypothesize is attributed to the lower signal quality of the PPG signal from the wrist-worn CST. Papini et al. (2020) showed a substantial increase in SA detection performance on data collected with a wrist-worn CST when segments with low signal quality were removed [21]. Implementing a quality-filtering mechanism for segments with suboptimal signal quality in the datasets used in this study is likely to enhance overall performance. The STAGES Arc dataset has $\kappa = 0.42$ and $\rho = 0.203$ ($p = 0.343$), which is substantially lower than the correlation metrics of the other datasets. From our investigation, it became evident that five severely misclassified recordings account for a significant portion of the variability in this small dataset (See supplementary Fig. 5 and 6). However, due to the limited sample size, it is challenging to formulate concrete hypotheses to explain this phenomenon.

C. Benchmark against related work

The performance of our proposed approach was benchmarked qualitatively against both directly comparable SOTA works that used data from wrist-worn CSTs [21] and against established, proprietary SA screening systems that utilize related signal modalities [27], [32]. Comparison between works on databases with wearable data indicate that the proposed approach has similar or better performance compared to these systems. It's important to note that a direct comparison between databases can only offer indicative insights due to the natural variability of AHI severity and underlying conditions between the cohorts. Nonetheless, from the internal benchmark our experiments show that the model trained using raw signals produced comparable or better results when compared to an approach that use surrogate signal inputs [22]. Direct comparison between Res-U-Net and the recurrent GRU model was only possible using the surrogate input, since the GRU model cannot process raw input. These models perform similarly across databases. An effort was made to implement the work of Papini et al (2020), though comparison is still challenging due to varying supervision resolutions and implementation biases. Our approach detects apnea events of varying duration, while Papini et al.'s method focuses on 30-second sleep epochs. Additionally, their approach relies on 212 "hand-crafted" features, but since they didn't release their code, implementing these features carries some implementation bias risk. Please refer to supplementary material for details on how this work was adopted.

D. Transfer learning

E. We investigated using existing, clinical datasets to improve generalizability of the system. Our experiments showed that pretraining the model on the MESA database that only contained PPG data improved performance considerably from a F1-score of 0.50 to 0.55 on the internal test set and from 0.36 to 0.42 external, wearable test sets. And weighted κ was non-significantly decreased from 0.68 to 0.66 and significantly increased from 0.42 to 0.57, respectively. Performance in terms of sleep stage prediction was also improved when pretraining on the MESA database. Please refer to the supplementary material for these results as well as the significance analysis. During pretraining on the MESA dataset, our model only learns from the PPG modality. To ensure the model did not consider the impact of ACC redundant, an input-channel dropout with a probability of 10% was added during finetuning on the target, internal dataset. Our experiments show that ACC proved to be substantially impactful – improving F1 test performance from 0.54 to 0.55 and from 0.31 to 0.41 on the internal and external validation sets, respectively. These experiments illustrate the efficacy of transfer learning when data is insufficient in size. The rational for integrating acceleration data for the detection of apnea events stems from the fact that Apnea events are often terminated by an arousal or a brief awakening, which can manifest as subtle or significant muscle movements. These movements, detectable through acceleration data, can serve as valuable markers to delineate the cessation of an apnea event. An intriguing direction to explore involves using movement signal surrogates, like ECG-based activity counts [52], instead of employing a decoy vector to compensate for the missing acceleration signal in the MESA study. While these surrogates may not precisely match raw acceleration data, they present a potential opportunity for enhancement in our approach. The dual-stream, temporal Res-U-Net model

The U-Net architecture has a simple block structure and a strong temporal core. It learns at multiple temporal scales which enables it to learn both short-term and long-term relationships. Tractability throughout the network is enhanced by inclusion of the residual connections, which proved to increase performance. This confirms findings from the semantic image segmentation research field [44]. The Res-U-Net model did show similar performance to the GRU model. Capturing the underlying physiological signals through abstraction is the central aim of this research, and while this can certainly be

Authorized licensed use limited to: Technical University of Denmark DTU Library. Downloaded on April 04,2024 at 08:46:44 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2024.3378480

achieved through signal processing, our study also delves into the application of deep learning to discover these abstractions. Our findings suggest that the particular feature-based approach we implemented may not be device-agnostic, as it demonstrated limitations when dealing with out-of-distribution data. It is important to note that the observed variations in performance across different datasets could potentially be attributed to the distinct characteristics of these datasets rather than the approach's inherent inability to generalize across devices. To robustly establish device agnosticism, it would be essential to conduct experiments with identical recordings across various devices and meticulously analyze the statistical differences in the outcomes.

Enriching the feature pool with features learned from the auxiliary training task of predicting sleep stages only gave minimal improvements to the performance. However, predicting sleep stages allowed us to infer TST from these predictions, and this in turn, helped estimate the AHI automatically.

The model streams were trained separately. Training these in parallel could enhance performance, as the model streams could potentially cooperatively complement each other in an iterative fashion. The results demonstrate that our model exhibits higher sensitivity to SA events that are linked to oxygen desaturations and/or arousals. Therefore, the inclusion of a third stream that predicts these auxiliary events may enhance performance further by leveraging contextual information related to sleep events that are correlated with or caused by SA.

F. Model strengths and weaknesses

Adopting the under- and overestimation criteria introduced by Papini et al. (2020) [21], resulted in 20 recordings (6 %) that considerably underestimated and 14 (4%) that considerably overestimated the AHI of 331 rec. It was chosen to only include recordings from the data sets with PSG-based PPG to remove the impact of estimation errors due to poor signal quality, as elaborated in section B of the Discussion. The underestimated group was characterized by having more hypopneas with shorter duration, and by having fewer SA events that were occurring during REM sleep and more during deep sleep compared to the correctly- and the overestimated group. In agreement with the findings in this study, existing heart rate-based SA detection systems show lower performance to hypopnea events when compared to actual SA events as well [21], [22]. Hypopneas has less impact on respiration and are typically shorter compared to OSA and CSA, where airflow completely stops [26]. Naturally, it follows that hypopneas have a smaller impact on associated signals as well, which may explain why they are more difficult to identify. The model's performance in relation to OSA and hypopneas aligns with the trends observed when comparing it to the inter-scorer agreement. However, when it comes to CSA, comparing the model's performance to inter-scorer agreement becomes more challenging. CSA is characterized by its infrequent occurrence, which can make it less prevalent in the dataset.

It was found that 52% of predicted events in the overestimated group were associated with oxygen desaturation or arousal, partly justifying the high false positive rate, as these events can mimic SA events. However, the cause of the remaining 48% of false predictions could not be identified, as they were not associated with any analyzed context. Papini et al. (2020) showed that their SA classification system was very dependent on the diagnoses of the patients under analysis [21]. SA detection is potentially sensitive to cardiac arrhythmias or may be more difficult in patients with autonomic nervous system dysfunctions, and for patients with pacemakers. Categorizing the recordings used in this study by underlying condition may add further insight into the fundamental strengths and limitations of our system.

The proposed model performed better to detect events in the context of REM sleep. SA events in the context of REM sleep are, in most patients, longer, more frequent, and associated with more pronounced hypoxemia when compared with events during NREM sleep [45]. This is due to muscle atonia that typically occurs during REM sleep. It may be reasonable to assume that the relative immobility of subjects during normal REM sleep may make detection of movements associated with arousals secondary to SA easier to detect. However, it's also important to consider that there might be more events in REM sleep, although further information on the number of events per sleep stage is needed to draw definitive conclusions. The proposed model performed better to detect SA events that were associated with an arousal– or/and a desaturation event. Recent studies have shown that nocturnal hypoxemia rather than the frequency of SA events, is the main driver of cardiovascular risk [46], [47]. While further studies in this area is needed, this fact would enhance the screening potential of the system. Lastly, the relatively small size of the testing data sample may have implications for the generalizability of these findings.

G. Future perspectives:

The validation presented here was only conducted using a single nocturnal recording for each participant. CSTs have the potential to monitor patients over multiple nights. It would be of great interest to explore the potential of improving AHI performance from an ensemble of multiple nights. Our study suggests that simple wearable devices have the potential to serve for identification of high-risk individuals who should be evaluated further for disease management. While our study compared raw signal modalities to other approaches, e.g., feature-based, and surrogate signals, we acknowledge the importance of testing optimal quantization and sampling rates of the raw signals in future research. These avenues hold potential for streamlining our models while maintaining or improving performance, making them more practical and less resource demanding.

V. Conclusion

We present a flexible, deep learning architecture for the detection of sleep stages and SA events in nocturnal ACC and PPG recordings. The model learns from minimally preprocessed recordings to detect SA events and provides an exact onset and duration. It improves on our previously reported work, which achieved SOTA performance on a sleep stage classification task [18], by enhancing tractability throughout the network. The model was trained using two advanced training schemes to enrich the feature pool and to improve generalizability. This was achieved by feature enrichment through an auxiliary sleep stage prediction task, and through

Authorized licensed use limited to: Technical University of Denmark DTU Library. Downloaded on April 04,2024 at 08:46:44 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2024.3378480

transfer learning by utilizing existing clinical data. The latter proved to be most impactful.

The proposed approach has similar or better performance compared to feature-based systems that used data from wrist-worn CSTs [21] and against established, proprietary SA screening systems that utilize related signal modalities [27], [32]. Furthermore, the model performance remains stable when applied to an external dataset. A performance drop was observed when the model was applied to datasets with PPG signals recorded with wearable devices. Implementing a quality-filtering mechanism for segments with low signal quality could improve performance. Addressing these limitations will open the possibility for wrist-worn CSTs to become alternative screening systems to target undiagnosed cases of OSA or for the use as part of OSA management.

ACKNOWLEDGEMENTS

We would like to thank the National Sleep Research Resource team for their efforts in collecting, organizing, and making available PSG data used in this study. Some of the computing for this project was performed on the Sherlock cluster. 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. We would like to thank Huami Inc. for providing the Arc devices used in the collection of the STAGES dataset. Research reported in this article was funded through a Patient-Centered Outcomes Research Institute® (PCORI®) Award (CER-1511-33005). This work does not represent the views or opinion of the Department of Veterans Affairs or the U.S. Government. The statements presented in this publication are solely the responsibility of the author(s) and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute® (PCORI®), its Board of Governors or Methodology Committee, or official policy or position of the Veterans Health Administration (VHA), U.S. Government. Lasty, this material is the result of work supported with resources and the use of facilities at the James A. Haley Veterans' Hospital.

CODE AVAILABILITY:

https://github.com/MADSOLSEN/SleepDisorderedBreathingDetection/tree/master.

REFERENCES

[1] *S. Javaheri et al., "Sleep Apnea: Types, Mechanisms, and Clinical Cardiovascular Consequences," J. Am. Coll. Cardiol., vol. 69, no. 7, pp. 841–858, 2017.*

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

[3] *N. Sharma, J. Lee, I. Youssef, M. O. Salifu, and S. I. McFarlane, "Obesity, Cardiovascular Disease and Sleep Disorders: Insights into the Rising Epidemic," J. Sleep Disord. Ther., vol. 06, no. 01, pp. 1–7, 2017.*

[4] *P. E. Peppard, T. Young, J. H. Barnet, M. Palta, E. W. Hagen, and K. M. Hla, "Increased prevalence of sleep-disordered breathing in adults," Am. J. Epidemiol., vol. 177, no. 9, pp. 1006–1014, 2013.*

[5] *K. Kadhim et al., "Prevalence and Assessment of Sleep-Disordered Breathing in Patients With Atrial Fibrillation: A Systematic Review and Meta-analysis," Can. J. Cardiol., vol. 37, no. 11, pp. 1846–1856, 2021.*

[6] *S. Krishnan et al., "Comorbidities and quality of life in Australian men and women with diagnosed and undiagnosed high-risk obstructive sleep apnea," J. Clin. Sleep Med., vol. 18, no. 7, 2022.*

[7] *Frost & Sullivan, "Hidden Health Crisis Costing America Billions," Am. Acad. Sleep Med., pp. 1–25, 2016.*

[8] *N. F. Watson, "Health care savings: The economic value of diagnostic and therapeutic care for obstructive sleep apnea," J. Clin. Sleep Med., vol. 12, no. 8, pp. 1075–1077, 2016.*

[9] *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.*

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

[11] *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.*

[12] *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.*

[13] *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.*

[14] *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.*

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

[16] *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.*

[17] *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.*

[18] *M. Olsen et al., "A flexible deep learning architecture for temporal sleep stage classification using accelerometry and photoplethysmography," no. c, 2022.*

[19] *D. Benedetti et al., "Obstructive Sleep Apnoea Syndrome Screening Through Wrist-Worn Smartbands: A Machine-Learning Approach," Nat. Sci. Sleep, vol. Volume 14, no. May, pp. 941–956, 2022.*

[20] *S. Wu, M. Chen, K. Wei, and G. Liu, "Sleep apnea screening based on Photoplethysmography data from wearable bracelets using an information-based similarity approach," Comput. Methods Programs Biomed., vol. 211, p. 106442, 2021.*

[21] *G. B. Papini, P. Fonseca, M. M. van Gilst, J. W. M. Bergmans, R. Vullings, and S. Overeem, "Wearable monitoring of sleep-disordered breathing: estimation of the apnea–hypopnea index using wrist-worn reflective photoplethysmography," Sci. Rep., vol. 10, no. 1, pp. 1–15, 2020.*

[22] *M. Olsen, E. Mignot, P. J. Jennum, and H. B. Dissing Sorensen, "Robust, ECG-based detection of sleep-disordered breathing in large population-based cohorts," Sleep, vol. 43, no. 5, pp. 1–11, 2020.*

[23] *T. Penzel, G. B. M. Rg, M. A. L. Goldberges, and H. Peter, "The Apnea-ECG Database," pp. 255–258, 2000.*

[24] *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.*

[25] *A. Zinchuk and H. K. Yaggi, "Phenotypic Subtypes of OSA: A Challenge and Opportunity for Precision Medicine," Chest, vol. 157, no. 2, pp. 403–420, 2020.*

[26] *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.*

[27] *E. Yeh et al., "Detection of obstructive sleep apnea using Belun Sleep Platform wearable with neural network-based algorithm and its combined use with STOP-Bang questionnaire," PLoS One, vol. 16, no. 10 October, pp. 1–15, 2021.*

[28] *X. Chen, Y. Xiao, Y. Tang, J. Fernandez-Mendoza, and G. Cao,*

Authorized licensed use limited to: Technical University of Denmark DTU Library. Downloaded on April 04,2024 at 08:46:44 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2024.3378480

"ApneaDetector: Detecting Sleep Apnea with Smartwatches," Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol., vol. 5, no. 2, 2021.

[29] D. X. Zhou, "Universality of deep convolutional neural networks," Appl. Comput. Harmon. Anal., vol. 48, no. 2, pp. 787–794, 2020.

[30] 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.

[31] K. Kotzen, P. H. Charlton, S. Salabi, L. Amar, A. Landesberg, and J. A. Behar, "SleepPPG-Net: a deep learning algorithm for robust sleep staging from continuous photoplethysmography," pp. 1–11, 2022.

[32] O. C. Ioachimescu et al., "Performance of peripheral arterial tonometry–based testing for the diagnosis of obstructive sleep apnea in a large sleep clinic cohort," J. Clin. Sleep Med., vol. 6, no. 10, pp. 1663–1674, 2020.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, pp. 779–788, 2016.

[43] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," Pattern Recognit., vol. 91, pp. 216–231, 2019.

[44] Z. Zhang, Q. Liu, Y. Wang, and S. Member, "Road Extraction by Deep Residual U-Net," vol. 15, no. 5, pp. 749–753, 2018.

[45] K. P. Grace, S. W. Hughes, and R. L. Horner, "Identification of the Mechanism Mediating Genioglossus Muscle Suppression in REM Sleep," no. M.

[46] J. D. Gottlieb et al., "Hypoxia , Not the Frequency of Sleep Apnea , Induces Acute Hemodynamic Stress in Patients With Chronic Heart Failure," vol. 54, no. 18, 2009.

[47] O. Oldenburg et al., "Nocturnal hypoxaemia is associated with increased mortality in stable heart failure patients," pp. 1695–1703, 2016.

[48] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "arXiv : 1603 . 09382v3 [cs . LG] 28 Jul 2016 Deep Networks with Stochastic Depth."

[49] D. Hendrycks and K. Gimpel, "Gaussian Error Linear Units (GELUs)," pp. 1–9, 2016.

[50] 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.

[51] R.S. Rosenberg, S. Van Hout. "The American Academy of Sleep Medicine inter-scorer reliability program: respiratory events." Journal of clinical sleep medicine, vol 10, no 4, pp. 447-54.S, 2014.

[52] P, Fonseca, R.M. Aarts, X. Long, J. Rolink, S. Leonhardt. "Estimating actigraphy from motion artifacts in ECG and respiratory effort signals." Physiological Measurement.vol. 37, no. 1, 2015.

Authorized licensed use limited to: Technical University of Denmark DTU Library. Downloaded on April 04,2024 at 08:46:44 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.