2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871295

Deep-Learning based Sleep Apnea Detection using SpO2 and Pulse Rate

Pragya Sharma*, Ali Jalali, Maulik Majmudar, Kuldeep Singh Rajput, and Nandakumar Selvaraj

Abstract— **This work presents automated apnea event detection using blood oxygen saturation (SpO2) and pulse rate (PR), conveniently recorded with a pulse oximeter. A large, diverse cohort of patients (n=8068, age$\geq$40 years) from the sleep heart health study dataset with annotated sleep events have been employed in this study. A deep-learning model is trained to detect apnea in successive 30 s epochs and performances are assessed on two independent sub-cohorts of test data. The proposed algorithm showcases the highest test performance of 90.4% area under the receiver operating characteristic curve and 58.9% area under the precision-recall curve for epoch-based apnea detection. Additionally, the model consistently performs well across various apnea subtypes, with the highest sensitivity of 93.4% for obstructive apnea detection followed by 90.5% for central apnea and 89.1% for desaturation associated hypopnea. Overall, the proposed algorithm provides a robust and sensitive approach for sleep apnea event detection using a noninvasive pulse oximeter sensor.**

Clinical relevance— **The study establishes high sensitivity for automated epoch-based apnea detection across a diverse study cohort with various comorbidities using simply a pulse oximeter. This highly cost-effective approach could also enable convenient sleep and health monitoring over long-term.**

I. INTRODUCTION

Sleep-disordered breathing (SDB) is a respiratory disorder of recurring partial or complete cessation of breathing during sleep. This progressive condition is commonly known as sleep apnea, and it is highly prevalent in 6% to 17% of the general adult population and as high as 49% in the advanced age groups with moderate to severe apnea levels [1]. The risk factors for developing SDB are complicated, and the American academy of sleep medicine (AASM) suggests screening for sleep apnea in all adult patients with heart failure (HF), elevated blood pressure (BP), atrial fibrillation, resistant hypertension, type-2 diabetes or stroke [2]. Still, it has been reported that about 85% of those with SDB are undiagnosed [3]. This is likely due to a lack of patient awareness and cumbersome testing methodology by overnight polysomnography (PSG). PSG is an often expensive gold-standard test conducted in a sleep center, with sensors including nasal airflow meter, pulse oximeter, respiratory effort chest belts, electrocardiogram (ECG), electroencephalogram, and others. The alternate home sleep apnea test (HSAT) allows the flexibility to test from the comfort of home, but requires the user to wear obtrusive nasal cannula, chest belts, ECG, etc. [4]. The acquired overnight data from either PSG or HSAT are then scored based on AASM practice standards [5] by trained technicians. Therefore, cost-effective comfortable rapid screening and continuous monitoring remain a challenging problem.

There has been tremendous progress particularly in at-home apnea monitoring, focusing on alternate screening approaches with one or more combinations of noninvasive sensors. ECG typically captures respiratory dynamics and the derived heart rate variability (HRV) contains information of elevated sympathetic nervous system (SNS) activity during arousals, that are often associated with hypopnea. ECG has been widely studied for sleep apnea screening with publicly available Physionet 2000 dataset [6]. However, the algorithms requiring robust ECG processing are not generalizable to other ECG datasets [7]. SpO2 and photoplethysmogram (PPG) waveform derived pulse rate variability (PRV) are shown to provide good performance for apnea detection [8]. However, pulse oximeters typically do not output PRV, as it requires additional computational and memory burden to output PRV, which is still considered as research-level feature. Hence, the public datasets involving pulse oximeters only provide clinically useful pulse rate (PR) and SpO2 outputs. SpO2 desaturation-associated apneic events alone [9] can severely underestimate the overall apnea event count and duration, as only about 78% apneas and 54% hypopneas are associated with significant desaturations [10].

This work presents a fully automated epoch-based algorithm for sleep apnea monitoring using the ubiquitous pulse oximeter derived SpO2 and PR signals. While PSG traditionally uses a finger-tip pulse oximeter, additional device form factors including in-ear, wristbands, etc. can be used [11], that provide additional flexibility and comfort for the patients. This study showcases the use of SpO2 and PR from this cost-effective unobtrusive device alone for apnea detection while improving compliance due to comfortable sensing. A convolutional neural network (CNN) based deep-learning (DL) model is designed to perform binary apnea detection for each 30 s epoch of pulse oximeter vitals. This provides two-fold measurements of 1) apnea duration, and 2) apnea episode count. This real-time implementable DL algorithm for sleep apnea event detection is evaluated on a diverse cohort of patients with sleep disorders and cardiovascular comorbid conditions.

II. METHODS

A. Dataset

Sleep heart health study (SHHS) dataset [12], a prospective cohort study of the cardiovascular and other consequences of SDB, has been used in this work. All participants were 40 years or older with no history of sleep apnea treatment or ongoing home oxygen therapy at the baseline visit. The dataset was collected over a decade with two

Authors are with Biofourmis Inc, Boston, MA 02110, USA.
(*correspondence e-mail: pragya.sharma@biofourmis.com)

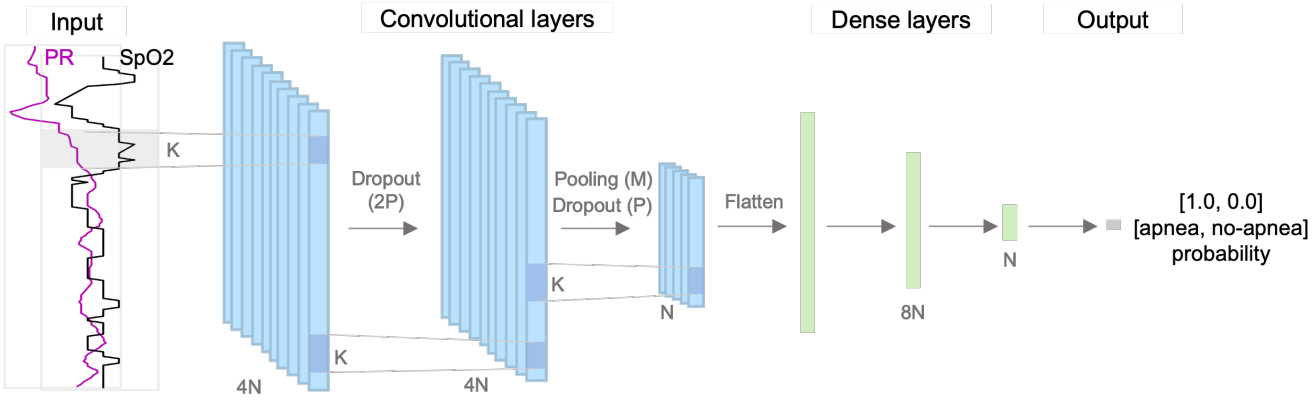

Fig. 1. DL architecture for epoch-based apnea detection with K=5, N=8, P=0.2, M=3. Input shows 150 s PR and SpO2 segments with apnea, indicated by ground truth [apnea, no-apnea] probability of [1, 0]. Input is fed to three 1D convolutional layers with kernel sizes, K, filter sizes (4N,4N,N), and strides (2,1,1). Dropout and maxpool layers are placed with dropout probability, P and pool size, M. Two fully connected dense layers with sizes (8N,N) are placed at the end. The last softmax layer generates apnea probability.

polysomnogram visits: SHHS1 (1995 – 1998), and SHHS2 (2001 – 2003), producing 5763 and 2651 patient records, respectively. The current analysis includes 5424 patients from SHHS1 and 2644 patients from SHHS2 that were free of any missing signals.

Both the SHHS datasets include overnight SpO2 and PR signals with 1 Hz sampling frequency (fs). An 'ox stat' signal is also available at the same fs to indicate the quality status. The annotated apnea and hypopnea events from sleep technicians required both events to be associated with $\geq$ 10 s change in respiratory signal. Similar to current AASM guidelines, apnea events only needed to be associated with significant airflow reduction. Hypopneas were identified with $\geq 30\%$ reduction in any respiratory signal for > 2 breaths or at least 2% desaturation for more subtle changes in breathing. To comply with the current guidelines, hypopnea annotations without either an arousal or at least a 3% desaturation event are ignored.

A train-test participant split is carried out in the SHHS2 cohort based on stratified apnea-hypopnea index (AHI) severity category (none, mild, moderate, and severe) with 30% (n=793) participants for testing and the remaining 70% (n=1851) for training and validation. The entire SHHS1 cohort (n=5424) is reserved as another test dataset.

B. Data Preprocessing

The algorithm starts with minimal PR and SpO2 processing. Poor samples are rejected based on 'ox stat' index. Next, PR samples outside the range of 40-150 beats per minute and SpO2 samples outside 70-100% are rejected. These outliers may be due to data outages or motion artifacts. The rejected timestamp data are replaced by linear interpolation. To reduce individual baseline effects, PR is standardized and SpO2 mean is subtracted for each participant.

A desaturation event is identified at a SpO2 sample point if a drop $\geq \text{SpO2}_{drop}$ from baseline ($\text{SpO2}_{baseline}$) occurs in the next 30 s, with a slope $\geq 0.1\%$ per second, resulting in a minimum value of SpO2_{nadir}. The $\text{SpO2}_{baseline}$ is the mean SpO2 in the previous 60 s window. After a desaturation event onset is detected, the event end is the minimum of time when signal reaches 1) $\text{SpO2}_{baseline} - 1$, or 2) $1.5 \times \text{SpO2}_{nadir}$, or 3) 120 s. The extracted desaturation events and given arousal annotations are used to keep valid hypopnea events based on current AASM guidelines. A SpO2_{drop} value of 2.9%, close to standard 3% is selected to minimize the error between estimated and given desaturation count.

C. Apnea Detection

A DL model is trained to perform binary classification of apnea or no-apnea for each 30 s epoch. The ground truth labels for epochs are generated from the corrected continuous annotations based on the following rules: 1) if an epoch contains an entire apnea event, it is labeled positive; 2) if an event spans across multiple epochs, the first epoch is positive if at least 50% of the epoch has apnea, the last epoch is positive if at least 50% has apnea or if the first epoch has $< 50\%$ apnea; 3) Any intermediate epochs, if present, are positive.

A LeNet-like architecture with 1D convolutional layers is used to extract features from input PR and SpO2 signal segments and detect apnea for each epoch. Each segment is 150 s long and composed of 2 previous and 2 future epochs, with the label corresponding to the center 30 s epoch. Fig. 1 shows the entire architecture. Model is implemented in Python using Keras library with PlaidML backend. The optimal hyperparameters are obtained using the NNI library by maximizing the validation data receiver operating characteristic area under the curve (ROCAUC). The training is performed with Adam optimizer to minimize categorical cross-entropy loss, with early stopping based on validation loss stability. The validation data is a 20% split out of the training data.

As the dataset is imbalanced with about 10% positive apnea epochs, output probability scores are converted to apnea or no-apnea labels with a threshold that optimizes

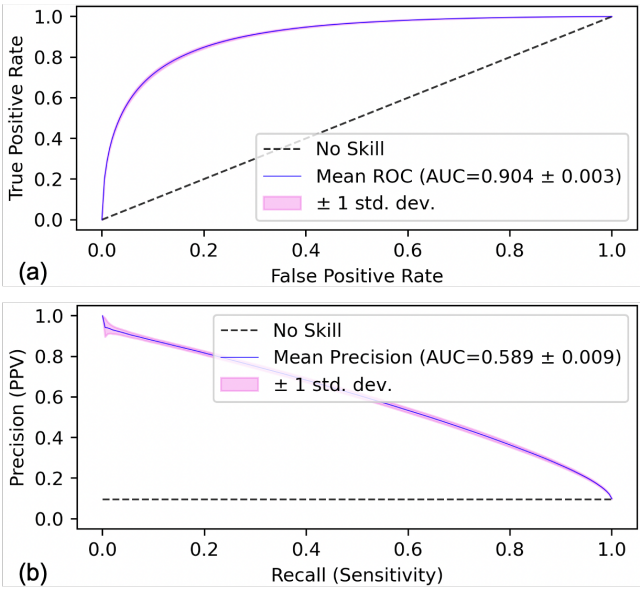

Fig. 2. Mean (a) ROC and (b) PR curves on the SHHS2 test dataset.

geometric mean (g-mean) of sensitivity and specificity,

$$G - mean = \sqrt{Sensitivity \times Specificity}. \quad (1)$$

The optimization is performed on the training data to derive the optimal threshold for each train-test split.

A robust testing design over 25 trials validates the performance and stability of the DL model architecture. In each trial, model is trained with randomly selected 70% SHHS2 participants and tested on the remaining cohort. Results section reports average performances over these trials.

D. Apnea Episodal Performance Analysis

In addition to epoch-based apnea detection, the analysis is extended to determine apnea episodal performance [13]. For this, consecutive positive epochs are combined as one episode. Episode sensitivity and positive predictive value (PPV) are defined as,

$$\text{Episode Sensitivity} = \frac{\#TP_{true}}{\#TP_{true} + \#FN}, \text{ and} \quad (2)$$

$$\text{Episode PPV} = \frac{\#TP_{pred}}{\#TP_{pred} + \#FP}. \quad (3)$$

The TP_{true} is defined as a true episode that is at least partially overlapped by one or more predicted episodes and TP_{pred} is a predicted episode that is at least partially overlapping one or more true episodes.

III. RESULTS

A. Apnea Epoch Detection Performance

Table I shows the CNN DL model performance for train and test datasets including metrics: accuracy, sensitivity, specificity, F1-score, PPV, precision-recall area under the curve (PRAUC) and ROCAUC. Balanced sensitivity and specificity scores are obtained after g-mean optimization.

TABLE I
AVERAGE APNEA EPOCH DETECTION PERFORMANCE

Epoch Performance (%)	SHHS2 Train	SHHS2 Test	SHHS1 Test
Accuracy	82.2	82.2	84.3
Sensitivity	82.8	82.9	68.9
Specificity	82.2	82.1	86.4
F1-score	47.8	47.9	51.1
PPV	33.6	33.6	40.6
PRAUC	58.9*	58.9	56.4
ROCAUC	90.4	90.4	86.2
Average performance (25 trials) with $\sigma \leq 1$, *: $1 < \sigma \leq 2$			

TABLE II
AVERAGE APNEA EPISODAL PERFORMANCE

Episodal Performance (%)	SHHS2 Test	SHHS1 Test
PPV	34.2*	41.5*
Sensitivity	79.4	66.3
Average performance (25 trials) with $\sigma \leq 1$, *: $1 < \sigma \leq 2$		

Figs. 2a and 2b show the ROC and PR curves for the SHHS2 test dataset, with the solid line indicating mean performance across 25 trials. The mean (m) and standard deviation (σ) of AUC across these trials are reported as $m \pm \sigma$.

This epoch-based analysis is also equivalent to the apnea durational performance, as described in [13], with 30 s resolution. With $> 80\%$ SHHS2 sensitivity, we have a high probability of apnea epoch detection and with precision or PPV of 0.41 for SHHS1, we beat the random prediction with apnea prevalence < 0.10. High AUC scores of 0.86 and 0.56 are observed for the ROC and PR curves even after a huge data imbalance. The standard deviations across performances in 25 trials are $< 1\%$ in almost all cases.

B. Apnea Episodal Performance

Table II shows apnea episodal performance statistics, where specificity or true negative episodes are undefined. True apnea episodes are detected with a sensitivity of 79.4% and 66.3% for SHHS2 and SHHS1 test datasets respectively.

C. Apnea and Hypopnea Subtype Performance

Table III presents findings of SDB subtype detectability, including obstructive, central, mixed sleep apnea (OSA, CSA, MSA), and desaturation or arousal associated hypopnea (H-desat, H-arousal). The second column in Table III shows the average percentage of epochs in each subtype out of the total positive epochs, given in the last row. H-arousal constitutes the highest percentage of epochs followed by H-desat and OSA. The sensitivity column shows the average percentage of epochs correctly detected as positive. OSA, CSA, and H-desat, generally associated with desaturation, are detected with $> 89\%$ sensitivity for SHHS2. Further, a notable sensitivity of 69% is achieved for arousal-associated hypopnea using the presented hybrid SpO2-PR based algorithm.

IV. DISCUSSION

Sleep apnea diagnosis is a challenging problem due to a lack of comfortable, cost-effective, and accurate screening

TABLE III
AVERAGE APNEA AND HYPOPNEA SUBTYPE PERFORMANCE

	Epoch Percentage (%)		Sensitivity (%)	
	SHHS2 Test	SHHS1 Test	SHHS2 Test	SHHS1 Test
OSA	25.7	26.8	93.4	81.1
CSA	5.4	3.3	90.5^{*}	82.1
MSA	0.0	0.0	$80.3^{\#}$	65.1^{*}
H-desat	31.3	29.8	89.1	78.0
H-arousal	37.6	40.1	69.3^{*}	52.9^{*}
Total apnea epochs in SHHS2 Test: 80, 868, and SHHS1 Test: 611, 628. Average performance (25 trials) with $\sigma \leq 1$, *: $1 < \sigma \leq 2$, $^{\#}$:$2 < \sigma$				

solution. Several alternate apnea detection methods with noninvasive sensing either have inconsistent performance on large cohorts [7] or focus only on desaturation-associated events [9], thereby severely underestimating apnea episode count and duration. Further, most works directly estimate AHI, which does not account for apnea duration that is found to be strongly linked to mortality [14]. This work combines a convenient and off-the-shelf pulse oximeter sensor with a robust algorithm, that works well across various SDB subtypes. Unlike previous research studies that generally test on relatively small datasets with limited disease conditions and age groups, our test performance results are based on a heterogenous SHHS cohort with various cardiovascular and other comorbidities, mimicking a real-world application.

With 90.4% ROCAUC for apnea epoch detection across the entire SHHS2 cohort, the presented model outperforms earlier work with 86% ROCAUC for desaturation-only apneic event detection on the same cohort [9], without accounting for H-arousal. Combining H-desat and H-arousal hypopneas results in a sensitivity of 68.1% on the entire SHHS cohort. This is on par with 65.8% sensitivity with ECG sensor [15]. Further, the AASM inter-scorer reliability program reported only 65.4% agreement per epoch for hypopnea, with 16.4% scoring no event [16]. Around $53-69\%$ detection rate of H-arousal in the present work indicates the model's unique advantage to detect subtle arousal-related changes in PR [17]. This is one of the most challenging issues, with even HSAT devices being unable to detect H-arousals [4].

The study achieves good performances in both test cohorts, with relatively lower sensitivity for the SHHS1 dataset that could be attributed to the poorer oximeter and PSG data quality in the first visit. Only about 66% of the SHHS1 participants have above-average PSG quality compared to 87% for SHHS2. 11% of the SHHS1 participants have < 6 hr of usable oximetry signal, while 97% of the SHHS2 participants have $\geq 95\%$ good quality signal during sleep. SHHS1 also has a younger population with a mean age of 63 yrs compared to 68 yrs. The presented DL model is trained on a subset of the older age-group SHHS2 cohort and tested on a relatively more diverse and larger SHHS1 cohort. Retraining the DL model with a large cohort of young age group representation can also potentially address the difference in performance.

We developed and tested a data-driven algorithm that achieves high epoch-level durational and episodal performances for apnea detection across various subtypes. The model performance is robust across multiple trials and achieves state-of-the-art results on large independent test cohorts with wide-ranging comorbidities. With featureless training and minimal time series processing, it is easier to evaluate the algorithm on a new dataset or deploy on real-world medical devices without requiring manual feature engineering or time-consuming model optimizations. Due to the comfort and ease of pulse oximeter placement by users themselves, clinicians can easily pre-screen at-risk patients from their homes, or monitor the effect of any treatment over its course by taking into account both apnea episode count and duration.

REFERENCES

[1] C. V. Senaratna *et al.*, "Prevalence of obstructive sleep apnea in the general population: A systematic review," *Sleep Medicine Reviews*, vol. 34, pp. 70–81, 2017.

[2] R. N. Aurora and S. F. Quan, "Quality measure for screening for adult obstructive sleep apnea by primary care physicians," *Journal of Clinical Sleep Medicine*, vol. 12, no. 08, pp. 1185–1187, 2016.

[3] K. K. Motamedi, A. C. McClary, and R. G. Amedee, "Obstructive sleep apnea: a growing problem," *Ochsner Journal*, vol. 9, no. 3, pp. 149–153, 2009.

[4] I. M. Rosen *et al.*, "Clinical use of a home sleep apnea test: an american academy of sleep medicine position statement," *Journal of Clinical Sleep Medicine*, vol. 13, no. 10, pp. 1205–1207, 2017.

[5] R. B. Berry *et al.*, "Rules for scoring respiratory events in sleep: update of the 2007 aasm manual for the scoring of sleep and associated events," *Journal of clinical sleep medicine*, vol. 8, no. 5, pp. 597–619, 2012.

[6] A. L. Goldberger *et al.*, "Physiobank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[7] G. B. Papini *et al.*, "On the generalizability of ECG-based obstructive sleep apnea monitoring: merits and limitations of the Apnea-ECG database," in *IEEE Engineering in Medicine and Biology Conference (EMBC)*. IEEE, 2018, pp. 6022–6025.

[8] R. Lazazzera *et al.*, "Detection and classification of sleep apnea and hypopnea using PPG and SpO2 signals," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 5, pp. 1496–1506, 2020.

[9] M. Deviaene *et al.*, "Automatic screening of sleep apnea patients based on the SpO2 signal," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 607–617, 2018.

[10] I. Ayappa *et al.*, "Immediate consequences of respiratory events in sleep disordered breathing," *Sleep medicine*, vol. 6, no. 2, pp. 123–130, 2005.

[11] T. Tamura *et al.*, "Wearable photoplethysmographic sensors—past and present," *Electronics*, vol. 3, no. 2, pp. 282–302, 2014.

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

[13] P. Sanders *et al.*, "Performance of a new atrial fibrillation detection algorithm in a miniaturized insertable cardiac monitor: Results from the reveal LINQ usability study," *Heart Rhythm*, vol. 13, no. 7, pp. 1425–1430, 2016.

[14] M. P. Butler *et al.*, "Apnea–hypopnea event duration predicts mortality in men and women in the sleep heart health study," *American journal of respiratory and critical care medicine*, vol. 199, no. 7, pp. 903–912, 2019.

[15] M. Olsen *et al.*, "Robust, ECG-based detection of sleep-disordered breathing in large population-based cohorts," *Sleep*, vol. 43, no. 5, p. zsz276, 2020.

[16] R. S. Rosenberg and 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. 447–454, 2014.

[17] C. Karmakar *et al.*, "Detection of respiratory arousals using photoplethysmography (PPG) signal in sleep apnea patients," *IEEE journal of biomedical and health informatics*, vol. 18, no. 3, pp. 1065–1073, 2013.