Journal of Medical and Biological Engineering (2021) 41:758–766
https://doi.org/10.1007/s40846-021-00646-8

ORIGINAL ARTICLE

End-to-End Sleep Apnea Detection Using Single-Lead ECG Signal and 1-D Residual Neural Networks

Roneel V. Sharan[1] · **Shlomo Berkovsky**[1] · **Hao Xiong**[1] · **Enrico Coiera**[1]

Received: 21 March 2021 / Accepted: 26 July 2021 / Published online: 29 July 2021

Abstract

Purpose Sleep apnea causes heart rate variability (HRV). HRV can be detected from the electrocardiography (ECG) signal and descriptors of HRV during sleep have been shown to be useful predictors of sleep apnea. In this work, we study the use of raw ECG signal and deep one-dimensional residual neural network (1-D ResNet) for end-to-end sleep apnea detection.

Methods Our method uses raw single-lead ECG signal as an input to a 1-D convolutional neural network (CNN) with residual connections, exploiting CNN's ability to learn distinguishing signal characteristics directly from the ECG signal and thereby forgoing the need for human engineered signal processing, feature extraction, and feature selection. In addition, we use weighted cross-entropy loss to account for the imbalance of apnea and non-apnea segments in our dataset, Bayesian optimization for fine-tuning the network hyperparameters, and data from current and adjacent epochs for predicting the label of the current epoch. The final ECG-based apnea detection network is evaluated on a dataset of 70 overnight ECG recordings.

Results The proposed method achieved an accuracy of 93.05% (AUC = 0.9819) in detecting sleep apnea segments when considering adjacent epochs, thus, outperforming several baseline techniques. Furthermore, the method achieved 100% accuracy in separating sleep apnea recordings from normal recordings.

Conclusion Our simple yet robust approach to ECG-based apnea detection demonstrates high accuracy. It has the potential to improve detection and diagnosis of sleep apnea and improve quality of life and health outcomes for millions of people worldwide.

Keywords Bayesian optimization · Electrocardiography · Heart rate variability · Obstructive sleep apnea · Residual neural network

1 Introduction

Inadequate sleep has been associated with a range of mental and physical health problems [1]. The 2011–2014 national health interview survey in the United States shows that 31.6% of adults get insufficient sleep [2] and a similar 2016 Australian survey shows that 33–45% of adults are affected by inadequate sleep and its daytime consequences [3]. Sleeping disorders are a common cause of inadequate sleep and one of the most common sleeping disorders is sleep apnea.

Sleep apnea is the involuntary cessation of breathing during sleep. Obstructive sleep apnea (OSA) is the most common type of sleep apnea, characterized by repeated episodes of partial (hypopnea) or complete obstruction (apnea) of the upper airways during sleep, limiting airflow to the lungs. The severity of sleep apnea is measured by the apnea–hypopnea index (AHI) which is defined as the number of apnea and hypopnea events per hour of sleep. The prevalence of OSA (AHI ≥ 5) in adults in the general population ranges from 9 to 38% [4].

Individuals with OSA experience a significant sleep disturbance leading to excessive daytime sleepiness and fatigue, which can potentially cause automobile accidents [5]. OSA is also associated with a number of medical conditions. In particular, there is an increased risk for coronary artery disease, congestive heart failure, and hypertension in individuals with severe sleep apnea [6].

Early diagnosis and treatment of OSA can help reverse symptoms, improve cognitive performance and quality of life [7], and reduce cardiovascular risks [8]. Polysomnography (PSG) is a widely used procedure in the diagnosis of

✉ Roneel V. Sharan
roneel.sharan@mq.edu.au

[1] Australian Institute of Health Innovation, Macquarie University, Sydney, NSW 2109, Australia

OSA. PSG is a type of sleep study where multiple physiological signals are recorded during sleep using various sensors and channels. The test is normally performed by sleep technicians in specialized sleep laboratories and physiological changes such as brain activity, eye movement, heart rhythm, muscle activity, respiratory effort, nasal pressure, and blood oxygen saturation levels are recorded overnight while the subject is asleep [9]. The recorded data is divided into small time windows (or *epochs*), the classifications of which is then used to determine the presence and severity of sleep apnea for the subject.

Setting up this multi-parametric test is time consuming. There can also be subjectivity in the analysis of the multiple hours of PSG data and the lack of recommendations about screening and the high costs of diagnostic PSG renders OSA underdiagnosed [10]. Recent research advancements in this area allow for the diagnosis to be achieved with fewer physiological signals, requiring fewer sensors and channels. Portable home sleep apnea tests are, therefore, being increasingly used to test medically uncomplicated subjects [11, 12].

While various approaches have been investigated [13–15], one line of research that has gained significant attention over the last two decades utilizes heart rate data. Apnea and hypopnea events lead to variations in heart rate, where the heart rate decreases during apnea and increases during recovery [16]. This characteristic behavior, known as heart rate variability (HRV), has been the subject of multiple studies investigating objective detection of apnea epochs using signal processing and machine learning techniques [17–21].

1.1 Related Work

Past techniques for HRV based apnea classification involve detecting the location of the R peaks in the ECG signal using the QRS detection algorithms followed by feature engineering. The R–R interval is computed as the difference between two consecutive R peaks and the resulting signal is used for feature extraction. Time and frequency domain features of the R–R interval signal are by far the most commonly used features [19–23].

For the computation of frequency domain features, the R–R interval signal needs further transformation into the frequency domain. The Lomb–Scargle periodogram [24] has been shown to be more suited for estimating the power spectral density of the unevenly sampled R–R interval signals then fast Fourier transform based methods [25]. Analysis of normalized spectral energy in four frequency bands is used: ultra-low frequency (0–0.003 Hz), very-low frequency (0.003–0.04 Hz), low frequency (0.04–0.15 Hz), and high frequency (0.15–0.4 Hz), together with the ratio of energy in the high and low frequency bands [26–28]. However, spectral analysis in finer frequency sub-bands is also useful [19]. Frequency-domain analysis of the R–R interval signals has generally produced better results than time-domain analysis [17].

Other feature extraction techniques such as the ECG-derived respiratory (EDR) signal [19, 21, 22, 29], cardiopulmonary coupling (CPC) [19], QRS morphology and subspace projections [22], and wavelet transform [30] have also been studied. While various classification methods have been experimented with, support vector machines (SVM) have been a popular choice of classifier due to their superior accuracy compared to conventional classification methods [18, 20–22, 30].

However, over the last decade, deep learning techniques have outperformed feature engineering-based techniques in apnea detection. Convolutional neural networks (CNNs), a popular deep learning method, originally used for image classification, have been shown to learn discriminative class characteristics directly from images [31]. CNNs have also been adopted in physiological signal classification applications by transforming the signal to an image-like representation [32, 33].

In detecting sleep apnea using ECG, the R–R interval signal has been used as an input to CNN [28, 34–36]. HRV is an unevenly sampled data. In [28, 34], cubic interpolation [37] is utilized to resize the data to a common dimension and padding is another technique to resize signals to equal dimensions [38, 39]. In [35], the R–R interval signal is adjusted to an image-like representation for classification using CNN. More recently, CNNs have been used to detect apnea directly from the ECG signal [40, 41].

1.2 Raw ECG and 1-D ResNet for Apnea Detection

In this study, we explore the use of raw single-lead ECG signal as a direct input to a one-dimensional CNN (1-D CNN) for detecting sleep apnea epochs. Our method offers various advantages and improvements compared to earlier works using a similar approach to this problem [40, 41]. With the conventional plain CNN, the convergence of the model starts to degrade with a deep architecture [42]. This problem can be addressed using residual learning, such as using identity and projection shortcuts [42]. While residual learning has been applied to the R–R interval signal for this purpose [36], in this work, we extend the residual learning framework to a 1-D CNN for raw ECG-based apnea detection. The resulting network is referred as a 1-D residual neural network (1-D ResNet) [43, 44] and evaluated against various conventional classification methods and a plain network used in [40, 41].

In addition, the duration of apnea episodes can last over a minute [45]. However, only a small portion of an episode may be present in an epoch, which typically have a duration of 30 or 60 s, and the episode can span multiple adjacent epochs [46]. With conventional classification methods, using data from epochs adjacent to the current epoch being

predicted, has shown to strengthen the ability of the classifier to distinguish between apnea and non-apnea epochs [19, 23, 34, 47]. In this work, we extend this approach to a deep learning framework. In particular, we evaluate two scenarios: using ECG data from the current epoch only to predict the current label and using ECG data from the current and adjacent, previous and following, epochs to predict the label of the current epoch.

Also, in medical applications, the number of disease samples is generally much lower than the number of normal samples which can lead to a biased model, particularly in deep learning applications. Conventionally, this problem has been addressed by balancing the dataset [40], such as using oversampling, undersampling, and data augmentation [35]. In this work, we explore the use of a weighted classification loss function [48] with the advantage of utilizing the full dataset without the need for data duplication, data removal, or generation of synthetic data.

Furthermore, a deep learning network has a number of hyperparameters, variables which determine how the network is trained. These variables can be tuned using trial and error, as in [41]. However, training deep learning models in this fashion can be time consuming and the obtained hyperparameters may not be optimal. In this work, we use Bayesian optimization for fine-tuning the 1-D ResNet hyperparameters. The Bayesian optimization algorithm aims to minimize an objective function in a bounded domain. The generalization performance of the Bayesian optimization learning algorithm is modeled from a Gaussian process and it has been shown to be effective in various hyperparameter optimization tasks [49, 50]. To analyze the robustness of our proposed method in ECG-based apnea classification, we analytically compare our results against several earlier methods.

The rest of the paper is organized as follows. In Sect. 2, we overview the dataset used in this work and the proposed and baseline methods. Experimental results are provided in Sect. 3 followed by the discussion of the results and conclusions in Sect. 4.

2 Method

We first describe the dataset used in this work followed by the proposed and baseline methods.

2.1 Dataset

This work utilizes the Apnea-ECG dataset [51, 52]. The dataset contains 70 overnight ECG recordings divided equally into training and test subjects: 35 recordings (total recording time of 285.42 h) for training and 35 for testing (total recording time of 288.38 h). The ECG signals are sampled at 100 Hz and vary in duration from 401 to 578 min. Human experts annotated each 60 s window (or epoch) as apnea or non-apnea based on simultaneously recorded respiration and other related signals. The age of the subjects in the dataset ranges from 27 to 63 years with 30 male subjects in the training dataset and 27 male subjects in the test dataset.

The recordings of every subject are grouped into one of the following three classes: Class A (apnea)—at least one hour with an AHI $\geq$ 10 and at least 100 min with apnea, Class B (borderline)—at least one hour with AHI $\geq$ 5 and 5–99 min with apnea, and Class C (control)—fewer than 5 min with apnea. This dataset was originally released for two tasks: classifying each epoch as apnea or non-apnea and classifying each subject's recording as apnea (Class A) or normal (Class C). These two tasks also form the subject of investigation in this work.

2.2 Proposed Method

An overview of the proposed method in apnea and non-apnea epoch classification is given in Fig. 1. The *training* block contains the ResNet model, which is trained, and the hyperparameters optimized, on the *training data*. At the end of the training process, we obtain a *trained ResNet*. The performance of the *trained ResNet* is evaluated on the *test data*.

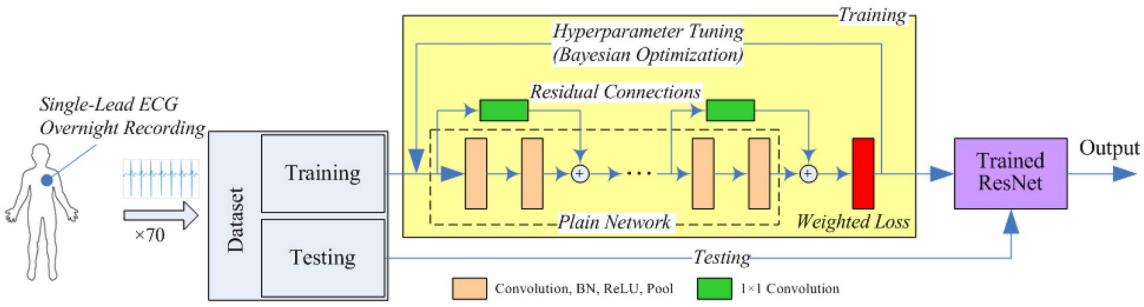

Fig. 1 An overview of the proposed method used for classifying raw ECG signal epoch into apnea and non-apnea

The dataset contains 17,010 epochs (6514 of them being apnea epochs) from the 35 training recordings and 17,268 epochs (6550 apnea epochs) from the 35 test recordings. We use raw ECG signals as a direct input to a 1-D ResNet and do not apply any preprocessing. We consider two strategies for the problem of predicting the label of each 60 s epoch. In the first approach, we use data from the *current* epoch $\mathbf{x}_i$ to predict the label y_i of the epoch. In the second approach, we use data from a window of the *current and adjacent* (both previous and following) epochs $\mathbf{x}_{i'} = [\mathbf{x}_{i-a}, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_{i+a}]$ to predict the label of the current epoch. The total number of combined epochs is 3 when $a = 1$ and 5 when $a = 2$.

The architecture of the 1-D ResNet is shown in Fig. 2. The plain network architecture was informed by offline experiments with a number of parameters and is inspired by the architecture of [38]. In the input layer, the ECG signal is normalized using zero mean and unit standard deviation to remove the effect of subject variability on the signal. Considering the adjacent epochs, the plain network for classification of 5 combined epochs (5 min with 30,000 sample points), consists of twelve convolutional layers, each followed by a batch normalization (BN) layer [53], rectified linear unit (ReLU) [54], and a max pooling layer [55]. Each convolutional layer has 32 filters. The kernel size for the first nine convolutional layers is 10×1 while the filter sizes of the remaining three convolutional layers are 6×1, 5×1, and 4×1, all with a stride of 1×1. The inclusion of a batch normalization layer after each convolutional operation makes the learning process faster and more stable [53]. The first nine max pooling layers have a pool size of 6×1 and the remaining three are 4×1, 3×1, and 2×1. Each max pooling layer uses a stride of 2×1 which halves the input and we couple a pooling layer with each convolutional layer.

The residual connections are added to the network every two sets of layers [42]. Projection shortcuts, a couple of 1×1 convolutions with stride 2×1, are used to match the dimension reduction of the pooling layers, transforming the plain network to a residual network. The final layers include a fully connected layer and a softmax layer [56]. The training data has many more non-apnea than apnea epochs. A weighted classification layer is used in the final layer to account for this class imbalance. This is realized using a weighted cross-entropy loss between the prediction scores Y and training targets T computed as

$$L = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{K} c_i T_{ni} \log\left(Y_{ni}\right) \tag{1}$$

where N represents the number of observations, K—the number of classes, and c—the class weights. The final network, therefore, consists of 70 layers.

The ResNet model was trained from scratch using adaptive moment estimation (Adam) [57], which we found to be more robust than stochastic gradient descent with momentum [56]. The Adam optimization algorithm adapts the learning rate using the moving average of the first and second moments of the gradients. The estimators for the bias-corrected first and second moments, $\hat{m}$ and $\hat{v}$, respectively, for the current training iteration t are given as

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t} \text{ and } \hat{v}_t = \frac{v_t}{1-\beta_2^t}, \tag{2}$$

respectively, where β_1 and β_2 are the hyperparameters of the algorithm. The model weight w is then updated as

$$w_t = w_{t-1} - \eta\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} \tag{3}$$

where η is the step size and ε is a small scalar.

We conducted a grid search to determine the effective range of the *initial learn rate*, *learn rate drop factor*, *learn rate drop period*, and *L2 regularization*. These parameters were then fine-tuned within the limits of this range using Bayesian optimization [50] with the aim of finding a hyperparameter set that minimizes the cross-validation error on the training dataset as

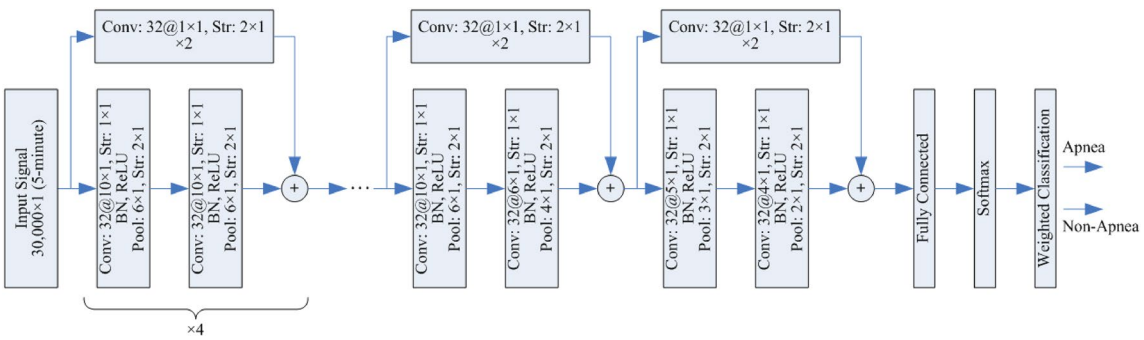

Fig. 2 Architecture of the 1-D ResNet used in this work

$$x^* = \arg\min_{x \in X} f(x) \tag{4}$$

where f is the objective function to minimize, the input x is part of the feasible set X, and x^* is the hyperparameter set yielding the lowest value. Bayesian optimization works by building a probabilistic model of the objective function. This is then searched using an acquisition function to determine the hyperparameters to evaluate on the true objective function. As an example, fine-tuning of the *initial learn rate* and *L2 regularization* using Bayesian optimization is illustrated as a surface plot in Fig. 3. The objective function is minimization of the cross-validation error. In practice, we could minimize the error by concurrently tuning more than two variables but for illustration we use only two variables.

We use the same network as in Fig. 2 for combinations of 3 epochs with input dimension of 18,000. When only the current epoch is used to predict the current label, the input dimension is 6000. The first two sets of convolution, batch normalization, ReLU, and pooling layers and the accompanying residual connection are removed to account for the smaller input size.

2.3 Baseline Methods

The dataset provides the location of the R-peaks, computed using the QRS complex detection algorithm [58]. We use the R–R interval signal for implementing two baseline methods.

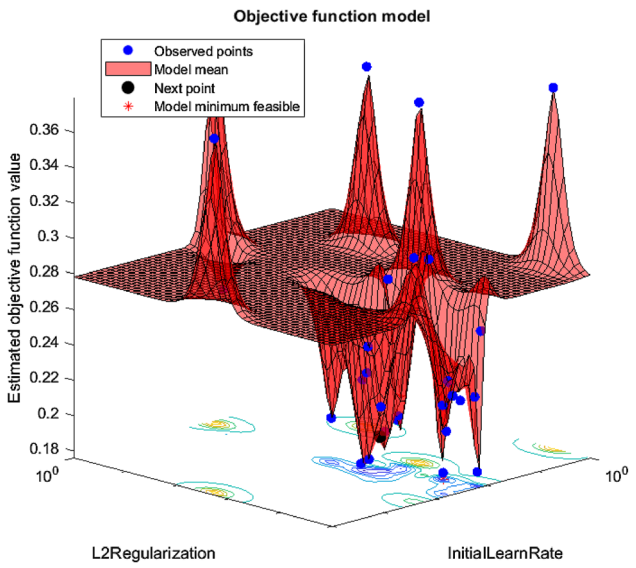

Fig. 3 Illustration of Bayesian optimization of the *initial learn rate* and *L2 regularization*

2.3.1 R–R Interval Signal Features

The first baseline method involves time and frequency domain features extracted from the R–R interval signal. We use 10 *time-domain* features [26, 27]. These include the mean, standard deviation, skewness, and kurtosis of the R–R intervals; root mean square and standard deviation of successive R–R interval differences; number of pairs of successive R–R intervals that differ by more than 20 ms and 50 ms; and fraction of the R–R intervals that differ by more than 20 ms and 50 ms.

For *frequency-domain* feature extraction, the spectral energy of the R–R interval data is computed using Lomb–Scargle periodogram. We analyzed the spectral energy in 32 equally spaced frequency bands up to the frequency of 0.4 Hz.

Apart from the individual time and frequency domain feature sets, we also consider their combined feature set. As such, this baseline method has 10 time-domain R–R interval features, 32 frequency-domain R–R interval features, and 42 features in the combined feature set. The classifiers for these feature sets are logistic regression (LR) [59] and SVM [60] with a radial basis function (RBF) kernel.

2.3.2 R–R Interval Signal and 1-D CNN

In the second baseline method, the R–R interval signal is fed directly into a 1-D CNN. The length of the R–R interval signal varies, depending on the heart rate. Since CNN requires a fixed size input, we use zero-padding and cropping to get to a common data size.

The 1-D CNN network architecture for classifying R–R interval signal is similar to the plain network shown in Fig. 2. The data input length is set to 400, 250, and 100 for 5 epochs, 3 epochs, and 1 epoch, respectively. There are 6 sets of convolutional, batch normalization, ReLU, and pooling layers when data from 5 epochs is combined, 5 sets when data from 3 epochs is combined, and 4 sets when the current epoch only is utilized.

2.4 Evaluation Metrics

The performance of the proposed and baseline methods is evaluated on the test dataset using accuracy, sensitivity, specificity, and the area under the curve (AUC). Accuracy is defined as the percentage of apnea and non-apnea epochs that were predicted correctly, sensitivity/specificity is the percentage of apnea/non-apnea epochs that were predicted correctly, and AUC corresponds to the area under the receiver operating characteristic (ROC) curve. For the class A versus class C subject classification task, we report the percentage of correctly classified subjects based on the epoch classifications and the definition of classes from

Sect. 2.1. For all the metrics, values closer to 1 or 100% indicate a strong performance of an algorithm in distinguishing between apnea and non-apnea epochs or subjects.

3 Results

Results using the proposed and baseline algorithms described in Sects. 2.2 and 2.3, respectively, are presented in this section. Results using the current epoch data only are presented first, followed by results using the current and adjacent epochs, and concluded by the results of recording-level predictions. We discuss the obtained results in Sect. 4.

3.1 Current Epoch to Predict Current Label

Results for apnea and non-apnea epoch classification using data from the current epoch only are given in Table 1. In the analysis we focus on the predictive accuracy and AUC (as a single-number combination of sensitivity and specificity) metrics. With the R–R interval signal features, the frequency-domain features yield a better classification performance than time-domain features, with SVM consistently outperforming LR. There is only a minor improvement in the classification performance when the two feature sets are combined. The best classification accuracy of 79.61% (AUC = 0.8483) is achieved using the combined feature set and SVM classifier.

With a classification accuracy of 82.60% (AUC = 0.8975), the classification performance is further improved using the R–R interval signal with 1-D CNN. The use of raw ECG signal with 1-D CNN (plain network) yields an accuracy of 87.10% (AUC = 0.9398). However, the best classification accuracy of 89.30% (AUC = 0.9520) is achieved using the proposed raw ECG signal and a 1-D ResNet.

Table 1 Results of current epoch data only

Input	Classifier	Acc (%)	Sens (%)	Spec (%)	AUC
Time-domain features	LR	73.82	58.32	83.32	0.7800
	SVM	77.57	72.89	80.43	0.8298
Frequency-domain features	LR	78.55	67.64	85.23	0.8149
	SVM	79.12	73.00	82.87	0.8458
Combined features (time + frequency)	LR	76.48	62.42	85.10	0.8085
	SVM	79.61	70.51	85.18	0.8483
Raw R–R Intervals	CNN	82.60	76.34	86.44	0.8975
Raw ECG Signal	CNN	87.10	81.65	90.43	0.9398
Raw ECG Signal	ResNet	**89.30**	**85.62**	**91.55**	**0.9520**

Bold indicates best result (highest value) in each column

3.2 Current and Adjacent Epochs to Predict Current Label

In Table 2, we present results for apnea and non-apnea epoch classification using a combination of the current and adjacent epochs. As far as the time and frequency domain features of the R–R interval signal are concerned, the latter outperform the former, while the best classification accuracy is achieved again using the combined feature set and SVM classifier. The classification accuracy using 3 epochs is 82.65% (AUC = 0.8765), which is slightly inferior to 83.27% (AUC = 0.8772) achieved with 5 epochs. These increase to 86.83% (AUC = 0.9377) and 88.60% (AUC = 0.9467) using the R–R interval signal and 1-D CNN. The results are further improved with the use of raw ECG signal and 1-D CNN; accuracy of 91.08% (AUC = 0.9695) with 3 epochs and 91.72% (AUC = 0.9730) with 5 epochs. However, the best results are again achieved using the raw ECG and 1-D ResNet; accuracy of 91.94% (AUC = 0.9760) with 3 epochs and 93.05% (AUC = 0.9819) with 5 epochs. Overall, the accuracy and AUC scores produced by the 5 epochs data are superior to those produced using 3 epochs.

Table 2 Results of current and adjacent epochs data

Input	Classifier	Current Epoch ± 1 Epoch				Current Epoch ± 2 Epochs			
		Acc (%)	Sens (%)	Spec (%)	AUC	Acc (%)	Sens (%)	Spec (%)	AUC
Time-domain features	LR	71.21	62.50	76.57	0.7599	70.18	58.74	77.24	0.7518
	SVM	78.32	73.07	81.55	0.8398	78.39	66.32	85.85	0.8389
Frequency-domain features	LR	81.66	74.86	85.85	0.8473	81.93	71.73	88.22	0.8498
	SVM	82.01	76.20	85.58	0.8682	82.34	73.78	87.63	0.8695
Combined features (time + frequency)	LR	69.72	27.15	95.95	0.6155	78.83	63.30	88.42	0.8286
	SVM	82.65	77.39	85.89	0.8765	83.27	74.02	88.98	0.8772
Raw R–R intervals	CNN	86.83	81.89	89.88	0.9377	88.60	81.44	93.03	0.9467
Raw ECG signal	CNN	91.08	88.93	92.40	0.9695	91.72	87.88	94.09	0.9730
Raw ECG signal	ResNet	**91.94**	**89.33**	**93.55**	**0.9760**	**93.05**	**90.16**	**94.84**	**0.9819**

Bold indicates best result (highest value) in each column

3.3 Recording Prediction

Finally, we use the epoch prediction results to predict the label of the recordings in the test dataset as per the criteria described in Sect. 2.1. We achieve an accuracy of 100% in classifying the recordings of the subjects as apnea (Class A) and normal (Class C). The perfect accuracy is obtained by all the evaluated 1-D ResNet classification models using the raw ECG data.

4 Discussion and Conclusions

The proposed raw ECG signal and 1-D ResNet approach performed well on the apnea and non-apnea epoch classification task. The distinguishing characteristics of the ECG signal are typically derived from analysis of the time-series data and handcrafted features are then used to represent the ECG signal [19]. This inevitably leads to information loss. On the contrary, in this work, deep residual neural network is shown to accurately learn such temporal characteristics directly from the raw signal, forgoing the need for manual signal processing and feature engineering.

The performance of our method improved when the data of the epoch being predicted was complemented with data from adjacent epochs. The best results were achieved using 5-min windows, although, only marginally better than 3-min windows. Notably, 5-min window length is also recommended for analysis of short-term recordings by the Task Force of the European Society of Cardiology and the North American Society of Pacing Electrophysiology [61].

When compared to [40, 41], in our work the ECG signals do not require any preprocessing, the use of residual connections provides improvement over a plain network, and the imbalance in our dataset is catered for using a weighted cross-entropy loss, without the need for data deletion or duplication required in undersampling and oversampling, respectively. Our method is further strengthened by the use of ECG data from adjacent epochs and fine-tuning of network hyperparameters using Bayesian optimization, achieving the highest accuracy of 93.05% (AUC = 0.9819) when data from adjacent epochs is considered.

We propose a simple yet robust method for detecting sleep apnea. While sleep apnea affects millions of people worldwide, sleep studies for diagnosing sleep apnea are expensive, not readily available, involve a suite of expensive sensors requiring time to set up, and the multitude of signals captured during a sleep study requires significant time to analyze. The proposed method, on the other hand, relies only on single-lead ECG data which can be measured by commercial-grade wearable devices, such as the Apple Watch [62]. As such, the proposed method could be integrated with wearable devices for screening apnea subjects at the comfort of their home. The method is shown to achieve strong classification performance, at the epoch and subject level alike, having the potential to improve apnea detection and diagnosis. Treatment of sleep apnea can follow the diagnosis and clinical treatments, such as using positive airway pressure, have been associated with improvement in quality of life [63] and reduced mortality [64].

Our work, however, has some limitations. First, we evaluated only a small number of neural network implementations and architectures. While we obtained promising results, it should be highlighted that recently there has been a tremendous interest in the deep learning research and a multitude of new methods are regularly published. It is possible that some of these perform well with ECG data and sleep apnea detection task, and eventually outperform our results. We leave a thorough evaluation of other deep learning methods for future studies. Second, the use of single-lead ECG signal only may be seen as a limitation. Despite being relatively cheap and easy to administer, the obtained results could potentially be improved by fusing other sensing technologies. While not aiming to achieve the full instrumentation offered by the PSG setting, we posit that adding a pulse oximeter sensor may improve the accuracy of sleep apnea detection, as pulse oximetry captures complementary oxygen saturation information [65]. Finally, our training dataset is limited to only 35 subjects of which only 5 are females and only 10 are normal. As such, it is not clear if a model developed on this small and imbalanced dataset will generalize to larger and more diverse populations. In addition, the dataset does not include any records of comorbidities, which would be present in practical situations. The proposed method would need to be evaluated on a larger and more balanced dataset with a more diverse population and a range of comorbidities, so that it could be eventually used as a robust sleep apnea diagnostic tool.

Funding The authors received no funding from an external source.

Declarations

Conflict of interest The authors have no conflict of interest to declare.

References

1. Stein, M. B., Belik, S.-L., Jacobi, F., & Sareen, J. (2008). Impairment associated with sleep problems in the community: Relationship to physical and mental health comorbidity. *Psychosomatic Medicine, 70*(8), 913–919.
2. Black, L. I., Nugent, C. N., & Adams, P. F. (2016). Tables of adult health behaviors, sleep: National Health Interview Survey, 2011–2014. Available from: http://www.cdc.gov/nchs/nhis/SHS/tables.htm.
3. Adams, R. J., Appleton, S. L., Taylor, A. W., Gill, T. K., Lang, C., McEvoy, R. D., & Antic, N. A. (2017). Sleep health of

Australian adults in 2016: Results of the 2016 Sleep Health Foundation national survey. *Sleep Health, 3*(1), 35–42.
4. Senaratna, C. V., Perret, J. L., Lodge, C. J., Lowe, A. J., Campbell, B. E., Matheson, M. C., Hamilton, G. S., & Dharmage, S. C. (2017). Prevalence of obstructive sleep apnea in the general population: A systematic review. *Sleep Medicine Reviews, 34*, 70–81.
5. Findley, L. J., Weiss, J. W., & Jabour, E. R. (1991). Drivers with untreated sleep apnea. A cause of death and serious injury. *Archives of Internal Medicine, 151*(7), 1451–1452.
6. Jean-Louis, G., Zizi, F., Clark, L. T., Brown, C. D., & McFarlane, S. I. (2008). Obstructive sleep apnea and cardiovascular disease: Role of the metabolic syndrome and its components. *Journal of Clinical Sleep Medicine, 4*(3), 261–272.
7. Engleman, H. M., & Douglas, N. J. (2004). Sleep · 4: Sleepiness, cognitive function, and quality of life in obstructive sleep apnoea/hypopnoea syndrome. *Thorax, 59*(7), 618–622.
8. Buchner, N. J., Sanner, B. M., Borgel, J., & Rump, L. C. (2007). Continuous positive airway pressure treatment of mild to moderate obstructive sleep apnea reduces cardiovascular risk. *American Journal of Respiratory and Critical Care Medicine, 176*(12), 1274–1280.
9. Chesson, A. L., Jr., Ferber, R. A., Fry, J. M., Grigg-Damberger, M., Hartse, K. M., Hurwitz, T. D., Johnson, S., Kader, G. A., Littner, M., Rosen, G., Sangal, R. B., Schmidt-Nowara, W., & Sher, A. (1997). The indications for polysomnography and related procedures. *Sleep, 20*(6), 423–487.
10. Senaratna, C. V., Perret, J. L., Lowe, A., Bowatte, G., Abramson, M. J., Thompson, B., Lodge, C., Russell, M., Hamilton, G. S., & Dharmage, S. C. (2019). Detecting sleep apnoea syndrome in primary care with screening questionnaires and the Epworth sleepiness scale. *Medical Journal of Australia, 211*(2), 65–70.
11. Colaco, B., Herold, D., Johnson, M., Roellinger, D., Naessens, J. M., & Morgenthaler, T. I. (2018). Analyses of the complexity of patients undergoing attended polysomnography in the era of home sleep apnea tests. *Journal of Clinical Sleep Medicine, 14*(4), 631–639.
12. Collop, N. A., Anderson, W. M., Boehlecke, B., Claman, D., Goldberg, R., Gottlieb, D. J., Hudgel, D., Sateia, M., & Schwab, R. (2007). Clinical guidelines for the use of unattended portable monitors in the diagnosis of obstructive sleep apnea in adult patients. *Journal of Clinical Sleep Medicine, 3*(7), 737–747.
13. Wu, C.-H., Lee, J.-H., Kuo, T. B. J., Lai, C.-T., Li, L. P. H., & Yang, C. C. H. (2020). Improving the diagnostic ability of the sleep apnea screening system based on oximetry by using physical activity data. *Journal of Medical and Biological Engineering, 40*(6), 858–867.
14. Mendonça, F., Mostafa, S. S., Ravelo-García, A. G., Morgado-Dias, F., & Penzel, T. (2019). A review of obstructive sleep apnea detection approaches. *IEEE Journal of Biomedical and Health Informatics, 23*(2), 825–837.
15. Dell'Aquila, C. R., Cañadas, G. E., & Laciar, E. (2020). A new algorithm to score apnea/hypopnea events based on respiratory effort signal and oximeter sensors. *Journal of Medical and Biological Engineering, 40*(4), 555–563.
16. Guilleminault, C., Winkle, R., Connolly, S., Melvin, K., & Tilkian, A. (1984). Cyclical variation of the heart rate in sleep apnoea syndrome: Mechanisms, and usefulness of 24 h electrocardiography as a screening technique. *The Lancet, 323*(8369), 126–131.
17. Penzel, T., McNames, J., de Chazal, P., Raymond, B., Murray, A., & Moody, G. (2002). Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings. *Medical and Biological Engineering and Computing, 40*(4), 402–407.
18. Sharma, H., & Sharma, K. K. (2016). An algorithm for sleep apnea detection from single-lead ECG using Hermite basis functions. *Computers in Biology and Medicine, 77*, 116–124.
19. de Chazal, P., & Sadr, N. (2016). Sleep apnoea classification using heart rate variability, ECG derived respiration and cardiopulmonary coupling parameters. In *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL*, (pp. 3203–3206).
20. Almazaydeh, L., Elleithy, K., & Faezipour, M. (2012). Obstructive sleep apnea detection using SVM-based classification of ECG signal features. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Diego, CA* (pp. 4938–4941).
21. Song, C., Liu, K., Zhang, X., Chen, L., & Xian, X. (2016). An obstructive sleep apnea detection approach using a discriminative Hidden Markov Model from ECG signals. *IEEE Transactions on Biomedical Engineering, 63*(7), 1532–1542.
22. Varon, C., Caicedo, A., Testelmans, D., Buyse, B., & Huffel, S. V. (2015). A novel algorithm for the automatic detection of sleep apnea from single-lead ECG. *IEEE Transactions on Biomedical Engineering, 62*(9), 2269–2278.
23. Martín-González, S., Navarro-Mesa, J. L., Juliá-Serdá, G., Kraemer, J. F., Wessel, N., & Ravelo-García, A. G. (2017). Heart rate variability feature selection in the presence of sleep apnea: An expert system for the characterization and detection of the disorder. *Computers in Biology and Medicine, 91*, 47–58.
24. Lomb, N. R. (1976). Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science, 39*(2), 447–462.
25. Delane, A., Bohórquez, J., Gupta, S., & Schiavenato, M. (2016). Lomb algorithm versus fast Fourier transform in heart rate variability analyses of pain in premature infants. In *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL* (pp. 944–947).
26. Shaffer, F., & Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. *Frontiers in Public Health, 5*, 1–17.
27. Clifford, G. D. (2002). Signal processing methods for heart rate variability, PhD thesis, Oxford University.
28. Sharan, R. V., Berkovsky, S., Xiong, H., & Coiera, E. (2020). ECG-derived heart rate variability interpolation and 1-D convolutional neural networks for detecting sleep apnea. In *42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Montréal* (pp. 637–640).
29. Pinho, A., Pombo, N., Silva, B. M. C., Bousson, K., & Garcia, N. (2019). Towards an accurate sleep apnea detection based on ECG signal: The quintessential of a wise feature selection. *Applied Soft Computing, 83*, 105568.
30. Zarei, A., & Asl, B. M. (2019). Automatic detection of obstructive sleep apnea using wavelet transform and entropy-based features from single-lead ECG signal. *IEEE Journal of Biomedical and Health Informatics, 23*(3), 1011–1021.
31. Keren Evangeline, I., Glory Precious, J., Pazhanivel, N., & Angeline Kirubha, S. P. (2020). Automatic detection and counting of lymphocytes from immunohistochemistry cancer images using deep learning. *Journal of Medical and Biological Engineering, 40*(5), 735–747.
32. Al Rahhal, M. M., Bazi, Y., Al Zuair, M., Othman, E., & BenJdira, B. (2018). Convolutional neural networks for electrocardiogram classification. *Journal of Medical and Biological Engineering, 38*(6), 1014–1025.
33. Shajil, N., Mohan, S., Srinivasan, P., Arivudaiyanambi, J., & Arasappan Murrugesan, A. (2020). Multiclass classification of spatially filtered motor imagery EEG signals using convolutional neural network for BCI based applications. *Journal of Medical and Biological Engineering, 40*(5), 663–672.

34. Wang, T., Lu, C., Shen, G., & Hong, F. (2019). Sleep apnea detection from a single-lead ECG signal with automatic feature-extraction through a modified LeNet-5 convolutional neural network. *PeerJ, 7*, e7731.
35. Wang, X., Cheng, M., Wang, Y., Liu, S., Tian, Z., Jiang, F., & Zhang, H. (2020). Obstructive sleep apnea detection using ecg-sensor with convolutional neural networks. *Multimedia Tools and Applications, 79*(23), 15813–15827.
36. Wang, L., Lin, Y., & Wang, J. (2019). A RR interval based automated apnea detection approach using residual network. *Computer Methods and Programs in Biomedicine, 176*, 93–104.
37. Keys, R. (1981). Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 29*(6), 1153–1160.
38. Kamaleswaran, R., Mahajan, R., & Akbilgic, O. (2018). A robust deep convolutional neural network for the classification of abnormal cardiac rhythm using single lead electrocardiograms of variable length. *Physiological Measurement, 39*(3), 035006.
39. Chen, T.-M., Huang, C.-H., Shih, E. S. C., Hu, Y.-F., & Hwang, M.-J. (2020). Detection and classification of cardiac arrhythmias by a challenge-best deep learning neural network model. *iScience, 23*(3), 100886.
40. Urtnasan, E., Park, J.-U., & Lee, K.-J. (2018). Multiclass classification of obstructive sleep apnea/hypopnea based on a convolutional neural network from a single-lead electrocardiogram. *Physiological Measurement, 39*(6), 065003.
41. Dey, D., Chaudhuri, S., & Munshi, S. (2018). Obstructive sleep apnoea detection using convolutional neural network based deep learning framework. *Biomedical Engineering Letters, 8*(1), 95–100.
42. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV* (pp. 770–778).
43. Li, Z., Zhou, D., Wan, L., Li, J., & Mou, W. (2020). Heartbeat classification using deep residual convolutional neural network from 2-lead electrocardiogram. *Journal of Electrocardiology, 58*, 105–112.
44. He, R., Liu, Y., Wang, K., Zhao, N., Yuan, Y., Li, Q., & Zhang, H. (2019). Automatic cardiac arrhythmia classification using combination of deep residual network and bidirectional LSTM. *IEEE Access, 7*, 102119–102135.
45. Wu, H., Zhan, X., Zhao, M., & Wei, Y. (2016). Mean apnea–hypopnea duration (but not apnea–hypopnea index) is associated with worse hypertension in patients with obstructive sleep apnea. *Medicine, 95*(48), e5493.
46. Rosenberg, R. S., & Hout, S. V. (2014). The American academy of sleep medicine inter-scorer reliability program: Respiratory events. *Journal of Clinical Sleep Medicine, 10*(4), 447–454.
47. McNames, J. N., & Fraser, A. M. (2000). Obstructive sleep apnea classification based on spectrogram patterns in the electrocardiogram. In *Computers in Cardiology, Cambridge, MA* (pp. 749–752).
48. Ho, Y., & Wookey, S. (2020). The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE Access, 8*, 4806–4813.
49. Doke, P., Shrivastava, D., Pan, C., Zhou, Q., & Zhang, Y.-D. (2020). Using CNN with Bayesian optimization to identify cerebral micro-bleeds. *Machine Vision and Applications, 31*(5), 36.
50. Snoek, J., Larochelle, H., & Adams, R. P. (2012*in Neural Information Processing Systems* (pp. 2951–2959).
51. Penzel, T., Moody, G. B., Mark, R. G., Goldberger, A. L., & Peter, J. H. (2000). The apnea-ECG database. In *Computers in Cardiology, Cambridge, MA* (pp. 255–258).
52. Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C. K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation, 101*(23), e215–e220.
53. Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint*. https://arxiv.org/abs/1502.03167
54. Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In *27th International Conference on Machine Learning, Haifa* (pp. 807–814).
55. Jarrett, K., Kavukcuoglu, K., Ranzato, M. A., & LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? In *IEEE 12th International Conference on Computer Vision, Kyoto* (pp. 2146–2153).
56. Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
57. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint*. https://arxiv.org/abs/1412.6980
58. Engelse, W. A. H., & Zeelenberg, C. (1979). A single scan algorithm for QRS-detection and feature extraction. *Computers in Cardiology, 6*, 37–42.
59. Cramer, J. S. (2002). The origins of logistic regression, Tinbergen Institute, Discussion Paper 2002–119/4.
60. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.
61. Task Force of the European Society of Cardiology and the North American Society of Pacing Electrophysiology. (1996). Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *Circulation, 93*(5), 1043–1065.
62. Wyatt, K. D., Poole, L. R., Mullan, A. F., Kopecky, S. L., & Heaton, H. A. (2020). Clinical evaluation and diagnostic yield following evaluation of abnormal pulse detected using Apple Watch. *Journal of the American Medical Informatics Association, 27*(9), 1359–1363.
63. Walia, H. K., Thompson, N. R., Katzan, I., Foldvary-Schaefer, N., Moul, D. E., & Mehra, R. (2017). Impact of sleep-disordered breathing treatment on quality of life measures in a large clinic-based cohort. *Journal of Clinical Sleep Medicine, 13*(11), 1255–1263.
64. Lisan, Q., Van Sloten, T., Marques Vidal, P., Haba Rubio, J., Heinzer, R., & Empana, J. P. (2019). Association of positive airway pressure prescription with mortality in patients with obesity and severe obstructive sleep apnea: The sleep heart health study. *JAMA Otolaryngology-Head & Neck Surgery, 145*(6), 509–515.
65. Dumitrache-Rujinski, S., Calcaianu, G., Zaharia, D., Toma, C. L., & Bogdan, M. (2013). The role of overnight pulse-oximetry in recognition of obstructive sleep apnea syndrome in morbidly obese and non obese patients. *Maedica (Bucur), 8*(3), 237–242.