






SLNet: A Hybrid Machine and Deep Learning Model for Sleep Apnea Episodes Detection From Single-Lead ECG Data

Charalampos Lamprou, *Student Member, IEEE* , Aamna Alshehhi, *Member, IEEE* , Thanos Stouraitis, *Life Fellow, IEEE* , Mohamed L. Seghier , and Leontios J. Hadjileontiadis, *Senior Member, IEEE* 

Abstract—Sleep Apnea (SA) is a common sleep disorder that causes breathing disturbances during sleep. Its diagnosis is often hindered by the high cost and the inconvenience of current diagnostic techniques, many involving multi-sensing from human body. Therefore, the development of automated tools that can effectively detect SA episodes using single-sensor data is crucial. In this vein, a comprehensive feature extraction and classification scheme for SA episodes detection using single-lead electrocardiogram (ECG) data, namely Single-Lead Network (SLNet), is proposed. SLNet involves ECG R-peak detection and heart rate variability estimation to extract time/frequency domain, and Poincaré plot features. Additionally, wavelet-based multi-resolution analysis (MRA) is employed to decompose the ECG signal and extract statistical and higher-order-crossings features from each MRA detail scale. Ultimately, to make predictions, SLNet utilizes hybrid machine/deep learning (ML/DL) models. The performance of SLNet was evaluated on 70 single-lead ECG recordings from 32 SA patients drawn from the open access “MIT Physionet Apnea-ECG Database”. Labels are provided for every minute of each recording resulting in a total of 34,243 annotated 1-min segments (21,201 non-SA and 13,042 SA). Stratified 5/10-fold nested and hold out cross-validation classification schemes were adopted for the ML and DL models, respectively. SLNet achieved high scores in terms of accuracy (up to 92.88%), sensitivity (up to 91.04%), specificity (94.02%) and Area Under the receiver operating characteristics Curve (AUC) (up to 98.09%), surpassing the performance of previous studies reported on the same dataset, using the same validation scheme. These results underscore the robustness of SLNet and indicate that highly accurate detection of SA episodes can be achieved, even through single-lead ECGs, scaffolding the use of efficient yet simple wearable SA episodes detectors.

Index Terms—Sleep Apnea (SA), Single-Lead Electrocardiogram (ECG), Automated SA Episode Detection, SLNet, Wavelet-based Multi-Resolution Analysis (MRA), Higher-Order Crossings, Machine/Deep Learning.

Manuscript received February 18, 2024. This work was supported by the Khalifa University of Science and Technology, under the Award No. CIRA-2020-031. (Corresponding author: C. Lamprou.)

Ch. Lamprou, A. Alshehhi, M. L. Seghier, and L. J. Hadjileontiadis are with the Department of Biomedical Engineering and the Health Engineering Innovation Center (HEIC), Khalifa University, P.O. Box 127788 Abu Dhabi, UAE (e-mail: {charalampos.lamprou, aamna.alshehhi, mohamed.seghier, leontios.hadjileontiadis}@ku.ac.ae).

T. Stouraitis is with the Department of Electrical Engineering & Computer Science, Khalifa University of Science and Technology, P.O. Box 127788 Abu Dhabi, UAE (e-mail: thanos.stouraitis@ku.ac.ae), and the Department of Electrical & Computer Engineering, University of Patras, GR 26504 Rio, Greece. (e-mail: thanos@upatras.gr)

L. J. Hadjileontiadis is also with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, GR 54124, Thessaloniki, Greece (e-mail: leontios@auth.gr).

I. INTRODUCTION

SLEEP constitutes a fundamental activity of living species, accounting for about one third of human life. Sleep quality is of high importance in maintaining a health status and poor sleep quality has been associated with several cardiovascular [1], mental [2], and neurological [3] deficits. In fact, sleep disorders have a significant negative impact on the quality of sleep and apart of affecting the human’s health, they also constitute a huge burden on the healthcare system. In particular, it is estimated that in the United States, 50 to 70 million adults suffer from a sleep disorder [4], while in Australia the estimated cost of sleep disorders during 2019 – 2020 was US\$35.4 billion [5].

Sleep Apnea (SA) constitutes a potentially serious sleep disorder with apneic episodes of repeated stops and starts of breathing. SA belongs to the category of Sleep Breathing Disorders (SBDs), which is the most prevalent category of sleep disorders and can be divided into Obstructive SA (OSA), Central SA (CSA), Mixed SA (MSA), sleep-related hypoventilation and hypoxemia [6]. OSA is the most common type of SA and it is caused by the incapability of the upper airway dilating muscles to oppose negative pressure in the airway during inhalation. As a result, the upper airway collapses, leading to breathing decrease (obstructive sleep hypopnea) or cessation (OSA) during sleep [7]. The severity of SA is estimated based on the Apnea Index (AI), Hypopnea Index (HI) and Apnea Hypopnea Index (AHI) that express the number of SA, hypopnea, SA and hypopnea events during one night, divided by the hours of sleep, respectively.

Early diagnosis and treatment of sleep apnea are essential to prevent long-term complications such as hypertension, atherosclerosis, and mental disorders, with various management options available, including lifestyle changes and medical interventions [8]–[10]. In clinical settings, the most widely used test for diagnosing SA is polysomnography (PSG). PSG uses electrocardiogram (ECG), Electroencephalogram (EEG), Electromyogram (EMG), Electrooculogram (EOG), respiratory effort, oxygen saturation (SaO₂) and airflow to simultaneously measure neurophysiologic, cardiopulmonary, and other physiologic parameters [11]. During the recordings, the subjects have to sleep in a clinical environment, with multiple sensing electrodes attached to their body and under

the presence of sleep experts who supervise the whole process. However, PSG is a highly inconvenient and expensive process. As a result, many people are deterred from visiting screening centers and consequently many SA cases remain undiagnosed [12]. Hence, the development of new SA detection methods that are more convenient and cost effective, while being highly accurate, is crucial.

In this vein, several studies have utilized various physiological signals and Machine/Deep Learning (ML/DL) algorithms in order to detect SA episodes. In [13], time domain features of EEG signals were extracted and fed to ML algorithms, in order to detect SA events. In [14], photoplethysmography and peripheral oxygen saturation signals were utilized to unveil differences between OSA, CSA and central hypopnea by means of a classification scheme. In [15], continuous wavelet transform and multilayer perceptrons were used for the detection of SA and hypopnea events from midsagittal jaw motion (mouth opening) recordings. Finally, in [16], the R-R intervals were extracted from single-lead ECG signals and presented to DL models in order to classify ECG segments as apneic or normal.

In this study, we introduce the SLNet framework for comprehensive analysis of single-lead ECG signals, aiming to uncover cardiovascular changes between SA episodes and healthy sleep. SLNet employs wavelet multi-resolution analysis (MRA) to extract information at different scales, followed by non-linear analysis using Higher-Order Crossings and approximate entropy. Additionally, widely used SA biomarkers, Heart Rate Variability (HRV) [17], and R-peak amplitudes [18] are incorporated for SA detection. HRV offers insights into cardiac autonomic control across sleep stages, while R-peak amplitudes reflect respiratory activity. To optimize feature utilization, sophisticated artificial intelligence schemes, combining ML and DL architectures, are designed. The proposed methodology's robustness is ensured through stratified 5/10-fold cross-validation (CV) classification schemes.

The experimental results from the application of SLNet on single-lead ECG data drawn from the open access "MIT Physionet Apnea-ECG Database" indicate that SLNet possesses high potential in efficiently detecting the SA episodes, thus contributing towards the development of convenient and cost-effective tools that can detect SA episodes using single-sensor data.

Overall, the contribution of this work can be summarized as follows:

- We suggested a comprehensive feature extraction pipeline to optimally capture the SA related information that is inherent in the single-lead ECG data.
- We introduced, for the first time, the use of Higher-Order Crossings to extract valuable features at different MRA scales.
- We performed feature characterization and model interpretation to identify a set of reliable biomarkers that can be used by future studies for SA episodes detection.
- We developed multiple ML and DL models in order to accurately detect SA episodes.

- We presented an additional evaluation of SLNet, with respect to anthropometric characteristics and SA-related indices.
- Finally, with the proposed SLNet pipeline, we achieved highly accurate results, surpassing the performance of previous studies that used the same dataset and evaluation method.

The structure of the paper is organized as follows. In Section II, related work on SA episodes detection is presented. The proposed SLNet methodology along with the performance evaluation setup are presented in Section III. Furthermore, in Section IV, experimental and implementation issues are described. Moreover, in Section V, the experimental results of the application of SLNet on single-lead ECG data from SA patients are presented, discussed and compared to the results of previous state-of-the-art studies; performance on different grouping scenarios, implications, limitations and future directions are also included. Ultimately, the paper is concluded in Section VI.

II. RELATED WORK

Several rigorous ML-/DL-based approaches already developed for SA episodes detection using single-lead ECG data are summarized here.

In an attempt to detect SA episodes, Janbakhshi *et al.* [18], employed features extracted from the time-domain and frequency-domain content of the ECG Derived Respiration (EDR). For the detection of SA episodes, different models, including Linear and Quadratic Discriminant Analysis (LDA and QDA), k-Nearest Neighbors (kNN), Support Vector Machines (SVM) and Artificial Neural Network (ANN) were used to achieve an accuracy score of 89.60 %. In another SA detection study [19], the HRV and EDR signals were derived from the ECG and subjected to a feature extraction procedure. The resulted features passed through a feature selection process and the 20 best were given as input to five different ML models. For their best performing model, an accuracy score of 82.12% was reported. In an effort to extract SA-related information at different frequency ranges, Fatimah *et al.* [20] employed Fourier decomposition to separate the ECG signals into nine frequency intervals. Consequently, some statistical and entropy features were extracted from these components. Finally, the authors compared the performance of four different classifiers (kNN, SVM, Bagging and LogitBoost) in detecting SA episodes and reported an accuracy of 92.59%. Another study that attempted to utilize an ECG decomposition approach in order to detect SA episodes was reported in [21]. In particular, the authors proposed a method that consists of decomposing the ECG signal into distinct frequency intervals using tunable-Q wavelet transform and calculating centered cross-entropies from the decomposed signals. The resulted features were presented to multiple ML models that achieved accuracy scores up to 92.78%. Recently, in an effort to explore the efficiency of different ML and DL architectures in predicting SA episodes, Bahrami *et al.* [22] extracted some widely used in ECG analysis features (e.g., HRV and R-peak amplitude). For the classification, 13 different ML models

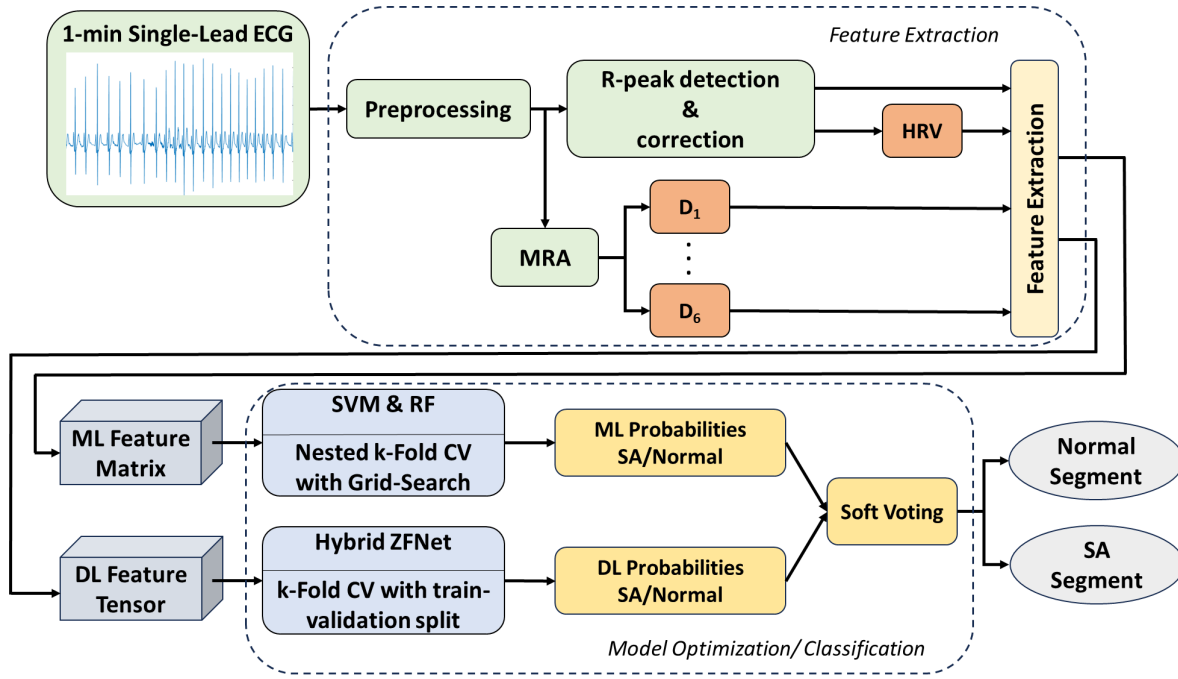


Fig. 1: Flowchart of the proposed SLNet. The single-lead ECG segment passes through two different pipelines. The first pipeline involves R-peak detection and HRV calculation, while in the second pipeline, the MRA is used to decompose the preprocessed ECG. Next, a feature extraction framework is applied to both pipelines and the result is a 2D feature matrix (ML Feature Matrix) and a 3D feature tensor (DL Feature Tensor). Then, the ML Feature Matrix is fed to a ML pipeline that integrates SVM and RF classifiers, while the DL Feature Tensor passes through a DL pipeline that includes a hybrid ZFNet model. Both pipelines include a common stratified k-fold CV and a hyperparameter optimization step (in case of ML using grid-search with CV and in case of DL using a validation set). The final result is obtained through a soft voting scheme.

and 19 DL models were compared and the reported accuracy scores ranged between 82.45% and 88.13%.

The recent advancements in the field of DL have prompted numerous researchers to adopt DL models for the detection of SA. Following the recent trend of employing DL architectures for classification tasks, Shen *et al.* [23] attempted to detect SA episodes by feeding the HRV to a multiscale dilation attention 1-D Convolutional Neural Network (CNN) and a weighted-loss-time-dependent classification model. The authors reported an accuracy score of 89.40%. Yang *et al.* [24] proposed a one-dimensional-squeeze-and-excitation residual group network. This model was designed to extract SA related patterns from both HRV and EDR and combine the extracted features to discriminate between Apnea and non-Apnea ECG segments. The authors reported an accuracy score of 89.70%. Moreover, Hu *et al.* [25] employed four ECG-derived sequences, i.e., raw ECG signal, R-peak amplitude, RR interval, and RR interval first-order difference and fed them to a hybrid transformer model with a multiperspective channel-attention block to achieve an accuracy score of 90.52%. Attempting to improve OSA detection, Shao *et al.* employed two different inputs, i.e., a 1-D sequence based on HRV and a 2-D time-frequency spectrum image, and presented them to a shared Bi-LSTM/Squeeze-Net model that achieved an accuracy score of 91.50%. Furthermore, in [26], Li *et al.* extracted the RR interval and the amplitude of the R-peaks and presented them to a time-frequency information fusion-based CNN-Transformer model with adaptive pruning, resulting in accuracy score of

91.68%. Finally, in an effort to optimize the feature extraction capabilities of CNNs, Abasi *et al.* [27] proposed a modified honey badger algorithm combined with quasi-opposition learning, arbitrary weighting agent, and adaptive mutation method to select the best set of hyperparameters for CNN. Their approach resulted to accuracy score of 91.30%.

Overall, these studies highlight the research efforts that are currently undertaken to improve the detection of SA episodes using single-lead ECG data, paving the way for more accurate and accessible diagnostic methods. By leveraging various signal processing techniques, feature extraction algorithms, and ML models, the related research efforts aim to develop reliable and efficient tools for identifying and classifying SA episodes. However, despite the wide variety of methods available, including ECG decomposition, HRV, EDR based methods, the potentiality of combining features from different analysis domains with different information representations, all handled by a hybrid ML/DL model, is still unexplored. In this vein, SLNet, a comprehensive feature extraction pipeline that integrates features from various domains and incorporates a hybrid ML/DL modeling is proposed here. By using information from diverse sources, SLNet aims to improve the accuracy and robustness of SA episode detection and classification. Moreover, the potential of combining DL with ML models is explored as a means to provide further advancements in SA episodes classification capabilities, leading to more reliable and precise outcomes. In the following section, the proposed SLNet is presented in detail.

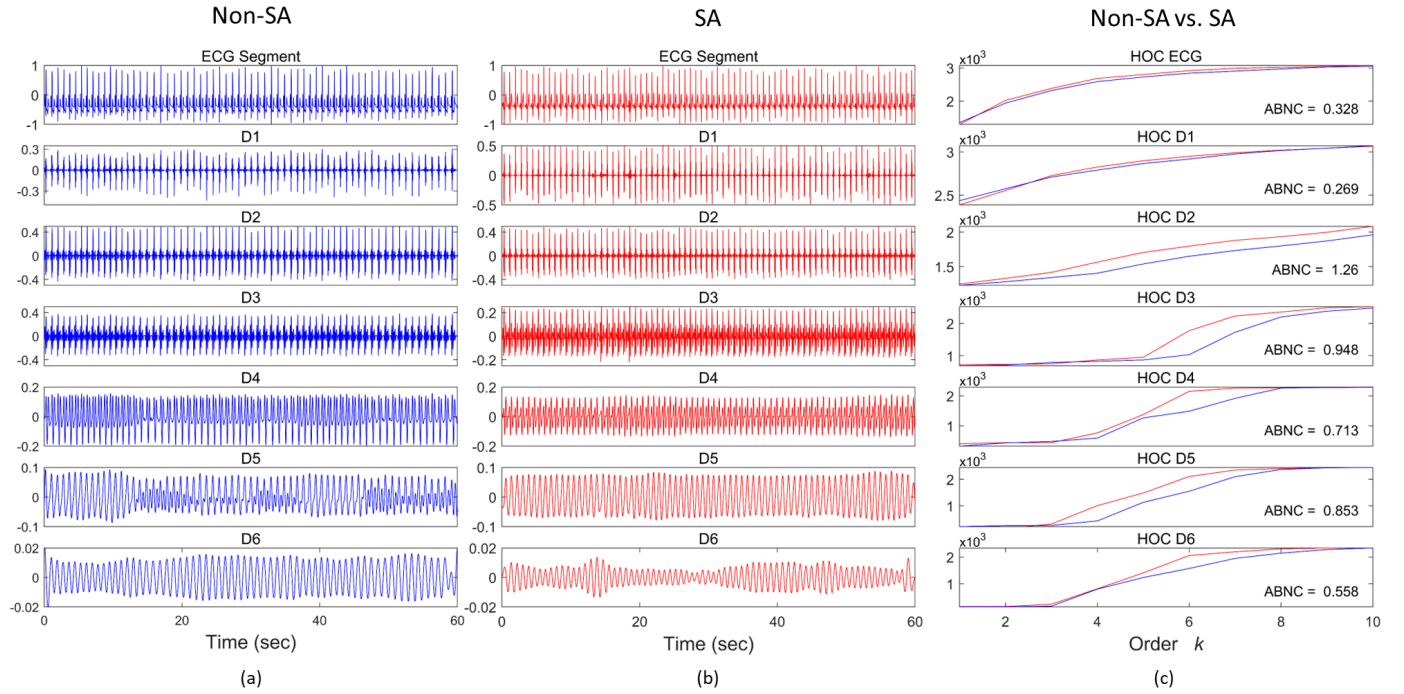


Fig. 2: Illustration of the 6-level MRA outcome using the sym4 wavelet, when applied to (a) a non-SA ECG segment (blue color) and (b) a SA ECG segment (red color). Moreover, in (c), a comparison between the estimated HOC values of the non-SA case (a, blue color) and the SA case (b, red color) ECG segments is presented, for both the original ECG segment (first row) and the obtained detail scales D_1, \dots, D_6 (rows two to seven). To highlight the amplitude-independent differences of the area between the non-SA and SA curves across the original ECG segment and the six detail scales, the Area Between Normalized Curves (ABNC) is also given.

III. METHODOLOGY

In order to classify ECG segments into the SA/non-SA classes, different features that are drawn from the single-lead ECG signal are proposed. The motivation behind is to accurately describe the morphology and the unique characteristics of the ECG signal, since it has been proven that SA affects the cardiovascular system, altering the morphology and the characteristics of the PQRST complex [28], [29].

A. SLNet Feature Extraction Scheme

Figure 1 depicts the flowchart of the propose SLNet approach. As shown in Fig. 1, the single-lead ECG segment is fed to two different feature extraction pipelines. Considering the upper branch of the feature extraction pipeline, the ECG signal is subjected to preprocessing, that includes a 3th-order IIR band-pass filter with a passband frequency range of [5-15] Hz. Next, the Hamilton R-peak detection method is used to extract the positions of R-peaks from the ECG signals [30]. Segments of 1-minute duration containing less than 25 or more than 100 peaks are discarded. Furthermore, a median filter, as proposed by *Chen et al.* [31], is applied using an 11-samples window to correct physiologically uninterpretable segments (RR intervals outside the range: [0.6-1.2], corresponding to heart rate: [50 bpm-100 bpm]). Finally, the amplitude of the estimated R-peaks and the HRV are calculated, and forwarded to the feature extraction procedure.

Regarding the second feature extraction pipeline (lower part of the feature extraction framework, Fig. 1), the single-lead ECG segment is preprocessed by first being normalized through maximum absolute scaling and then subjected to a 5th-order IIR high-pass filter, followed by a 5th-order IIR low-pass filter with cut-off frequencies of 1Hz and 40Hz, respectively. Next, the wavelet-based MRA is employed to extract features from the ECG at different resolution scales (up to six) and unveil characteristics of the ECG that could not be conventionally exposed. Regarding the choice of wavelet, there are several families of wavelets available. Among these families, Daubechies and Symlets are considered gold standard in ECG analysis [32]. After inspecting the different wavelets of these families, the 4th-order wavelet of the Symlet family (sym4) was chosen, as it closely resembles the QRS complex [32]. Although in each decomposition level the input is decomposed into the approximate (representing low frequency content) and the detail (representing high frequency content) coefficients, only the detail coefficients are used for feature extraction here, following the work in [33]. Ultimately, the six detail coefficients are forwarded to a feature extraction procedure. Figure 2 illustrates an example of a decomposed ECG segment with the corresponding detail scales ($D_k, k = 1, \dots, 6$) for the case of a non-SA (Fig. 2(a)) and a SA (Fig. 2(b)) 1-min ECG segment. Following, the two separate feature extraction procedures that were applied to the HRV and the $D_k, k = 1, \dots, 6$, respectively, are elaborated.

1) *HRV Features*: HRV is a reliable SA biomarker, as it has been widely used in ECG characterization and SA detection [17]. Here, HRV is utilized to extract a set of time-domain, frequency-domain, and Poincaré plot features.

Time Domain. The most prominent categories of time-domain features derived from the HRV are statistical, long-term, and short-term features [34], i.e.:

a) *Statistical Features*: The mean, median, minimum, range, standard deviation, skewness, and kurtosis are the statistical features extracted from the HRV.

b) *Long-Term Features*: The heart rate and the standard deviation of successive differences between adjacent R–R Intervals (*SDSD*) are the long-term features extracted from the HRV.

c) *Short-Term Features*: The total number of successive R–R differences greater than 50 ms (*NN50*), the total number of successive R–R differences greater than 20 ms (*NN20*), the percentage of adjacent R–R intervals that differ from each other by more than 50 ms (*pNN50*), and the percentage of successive R–R differences greater than 20 ms (*pNN20*) are used. Furthermore, the root mean square of successive differences between heart beats (*rMSSD*) is extracted from the HRV.

Frequency Domain. The frequency content of the HRV can be divided into three primary bands [34], i.e.: a) the Very Low-Frequency (VLF) band ([0.0033-0.04]Hz), that relates with the thermoregulation mechanisms, b) the Low-Frequency (LF) band ([0.04-0.15]Hz), that mainly reflects the sympathetic activity, and c) the High-Frequency (HF) band ([0.15-0.4]Hz), that relates with the parasympathetic activity. Although VLF features were previously used [22], the VLF band was not used here, as SA episodes classification was based on short (i.e. 1-minute length) ECG segments (see Section IV-A); hence, this ECG length is not sufficiently large to accurately capture the frequency content of the VLF band. After calculating the power spectrum of the HRV (PHRV), the frequency of maximum power (*maxF*), the skewness and the kurtosis of the PHRV (*PHRVskew*, *PHRVkurt*), as well as the skewness and the kurtosis of the PHRV within the 2 predominant bands (*PLFskew*, *PLFkurt*, *PHFskew*, *PHFkurt*), are calculated. Moreover the Shannon entropy of the PHRV (*PHRVent*) and of the PHRV within the 2 predominant bands (*PLFent*, *PHFent*) are also extracted. Finally, the Total Power (*TP*), LF Power (*LFP*), HF Power (*HFP*), percentage of LF Power (*LFN*), percentage of HF Power (*HFN*) and LF to HF ratio (*LF2HF*), are estimated.

2) *Poincaré plot features*: Poincaré plot features have been widely used in analyzing ECG recordings before, during, and after SA [22], [35]. Poincaré plot features including *SD1* and *SD2* as well as their ratio (*SD1/SD2*) [34], cardio sympathetic index (*CSI*) [36], modified CSI (*mCSI*) [36], and cardiovagal index (*CVI*) [36] are extracted from the HRV. Ultimately, the permutation entropy [37] is also calculated, to quantify the time-domain complexity of the HRV. Furthermore, the power spectrum of the HRV is also kept. In total, 36 features are extracted from the HRV.

3) *MRA Features*: Following the MRA analysis resulting in D_k , $k = 1, \dots, 6$, the following features are calculated for

TABLE I: DESCRIPTION OF FEATURES USED IN THIS STUDY.

Feature	Description
meanHRV	Mean value of HRV.
medianHRV	Median value of HRV.
minHRV	Minimum value of HRV.
rangeHRV	Range of HRV values.
stdHRV	Standard deviation of HRV.
skewHRV	Skewness of HRV.
kurtHRV	Kurtosis of HRV.
HR	Heart rate.
SDSD	Std of successive differences between adjacent R–R intervals.
NN50	Number of successive R–R differences greater than 50 ms.
NN20	Number of successive R–R differences greater than 20 ms.
pNN50	Percentage of adjacent R–R intervals that differ by more than 50 ms.
pNN20	Percentage of adjacent R–R intervals that differ by more than 20 ms.
rMSSD	Root mean square of successive differences between heart beats.
maxF	Peak frequency of HRV.
PHRV	Power spectrum of PHRV.
PHRVskew	Skewness of PHRV.
PHRVkurt	Kurtosis of PHRV.
PLFskew	Skewness of PHRV within the LF band.
PLFkurt	Kurtosis of PHRV within the LF band.
PHFskew	Skewness of PHRV within the HF band.
PHFkurt	Kurtosis of PHRV within the HF band.
PHRVent	Shannon entropy of PHRV.
PLFent	Shannon entropy of PHRV within the LF band.
PHFent	Shannon entropy of PHRV within the HF band.
TP	Total power.
LFP	Power within the LF band.
HFP	Power within the HF band.
LFN	Percentage of LF power.
HFN	Percentage of HF power.
LF2HF	LF to HF ratio.
SD1	Standard deviation of short-term R-R interval variability.
SD2	Standard deviation of long-term R-R interval.
CSI	Cardio sympathetic index.
mCSI	Modified cardio sympathetic index.
CVI	Cardiovagal index.
std_k	Standard deviation of the k^{th} MRA component.
$skew_k$	Skewness of the k^{th} MRA component.
$kurt_k$	Kurtosis of the k^{th} MRA component.
$ApEn_k$	Approximate entropy of the k^{th} MRA component.
$HOC_{n,k}$	Higher-order-crossings at order n and MRA level k.

each D_k component:

- Standard deviation (std_k)
- Skewness ($skew_k$)
- Kurtosis ($kurt_k$)
- Approximate Entropy ($ApEn_k$) [38]
- Higher-Order-Crossings ($HOC_{n,k}$) [39] (see Appendix A),

where $n = 1, 2, \dots, 10$ denotes the HOC order (see Appendix A). Figure 2 (c) illustrates an example of the estimated HOCs from the 1-min ECG segments and their MRA detail scales of non-SA (Fig. 2(a)) and SA (Fig. 2(b)) cases. To highlight the amplitude-independent differences of the area between the non-SA (blue color) and SA (red color) curves across the original ECG segment and the six detail scales, the Area Between Normalized Curves (ABNC) is given (see Fig. 2(c)).

A total of 84 features are extracted from the MRA-related feature extraction. Moreover, the histograms of the D_2 and D_3 were calculated using 180 bins. The selection of these particular detail coefficients is justified retrospectively through the feature characterization analysis (see Section V-B).

Overall, a total of 120 features (84 from the MRA of the ECG and 36 from the HRV) together with four vectors, i.e., the amplitude of R-peaks, the power spectrum of the HRV (PHRV) and the two histograms of the detail coefficients D_2, D_3 are estimated. In order to instill equal size, the PHRV and the R-peaks amplitude vectors are interpolated at 180 samples, following [22], using cubic interpolation. Thus, after the feature extraction pipelines of SLNet are complete, there are two different set of feature representations, i.e., a $(N, 120)$ feature matrix and a $(N, 4, 180)$ feature tensor, that are used as input to ML and DL models, respectively (where N the number of samples). Following the employed classification schemes used for performance evaluation, are presented.

B. SLNet Model Optimization/Classification Scheme

As shown in the lower part of Fig. 1, the obtained feature matrix (ML Feature Matrix) and feature tensor (DL Feature Tensor) are pushed through two different classification pipelines. In order to assess the ability of the proposed features in detecting SA episodes, stratified 5/10-fold CV classification schemes are used. In order to enable a fair comparison and be able to merge the outcomes of the two pipelines through a voting scheme, the two pipelines share the exact same CV scheme. Considering the ML models, hyperparameter tuning is performed in each fold of the outer k-fold CV using grid search with CV. On the other hand, for DL models, in each fold of the outer k-fold CV, 90% of the data are used to train the model and 10% of the data are used as a validation set to fine-tune the hyperparameters. In both the ML and DL pipelines the trained models are evaluated against the hold-out fold that does not participate in the training and hyperparameter tuning processes. Ultimately, the calculated probabilities of the ML models i.e., Support Vector Machine (SVM) and Random Forest (RF) (see Section IV-B), and DL model, i.e., Hybrid ZFNet (see Section IV-B), are merged through a soft voting scheme. Hence, after tested against the test set, every model provides a probability value that a specific 1-min ECG segment belongs either to SA or non-SA class. Consequently, the predictions are weighted based on the validation results and summed up. Here, the probabilities of the DL model were weighted by 0.6 while the two ML models were weighted by 0.2 each.

For the evaluation of the SLNet performance, the well established classification metrics of Accuracy, Sensitivity, Specificity, F1-score, and Area Under Curve (AUC) of the Receiver Operating Characteristics (ROC) metrics are used. In addition, for explaining the outputs of the DL models, the SHAP Gradient Explainer was used [40], assigning importance values to input features. The SHAP values fairly distribute the contribution of each feature to the prediction for a specific instance by considering all possible feature combinations. The Gradient Explainer approximates SHAP values using the gradients of the model's output with respect to its input features. Thus, it provides an explanation of the model's local behavior around a specific input instance by combining the gradients with the expected values of the model output. Features with higher absolute SHAP values have a greater impact on the

model's output, while those with lower absolute SHAP values have a less significant impact [40].

IV. EXPERIMENTAL AND IMPLEMENTATION ISSUES

A. Dataset Characteristics

The dataset used in this work is the "MIT Physionet Apnea-ECG dataset" contributed by Penzel *et al.* [41]. It is the most widely used dataset for ECG-based SA detection, as it is highly comprehensive and possess large sample size. The dataset includes a total of 70 nocturnal recordings in length from slightly less than 7 hrs to nearly 10 hrs (average \pm std: 8.2 ± 0.52 hrs), from 32 individuals (7/25 female/male, age: 44 ± 11 yrs, weight: 86 ± 22 kg, Body Mass Index (BMI): 27.87 ± 7.02). Out of the 32 individuals, 22 were recorded twice, two were recorded three times each, and four were recorded four times each. Each recording includes a continuous digitized ECG signal along with a set of apnea annotations (derived by human experts on the basis of simultaneously recorded respiration and related signals) [41]. Within this dataset, there are 13 individuals (25 recordings) classified as normal, indicated by an $AHI\leq5$. Six individuals (14 recordings) are categorized in the Mild and Moderate (M-M) SA class ($5<AHI\leq30$). Additionally, 13 individuals (31 recordings) are classified in the severe SA class ($AHI>30$). In addition to the AHI, the dataset also provides the Apnea and Hypopnea indices (AI and HI, respectively), as well. The sampling frequency is 100 Hz. Apart from the original annotation based on the PSG analysis, an additional binary annotation regarding the presence of SA ("disordered breathing") or not ("healthy breathing") was performed by a sleep expert. It should be noted that this annotation did not differentiate between SA and hypopnea events; hence, the annotated disordered breathing may contain one single SA or hypopnea or may contain a longer sequence of SAs and hypopneas. All annotations were mapped to 1-min time resolution, resulting in an overall of 34,243 annotated, non-overlapping 1-min ECG segments. After a pre-selection process for excluding some noisy ECG segments, a total of 34,000 (20,983 non-SA and 13,017 SA) 1-min ECG segments are used for the SLNet training and testing.

B. SLNet ML/DL Architectures

1) *Machine Learning*: Regarding the ML procedure within the SLNet classification scheme (Fig. 1), the SVM and RF classifiers were adopted. As Fig. 1 illustrates, in each fold of the outer k-fold CV scheme, the grid-search with CV method is used to select the best set of hyperparameters. Regarding the SVM implementation, a radial basis function kernel was utilized. The grid search approach was employed to determine optimal values for the regularization parameter C and the kernel coefficient γ . The grid for both C and γ consisted of the following values: $\{0.001, 0.01, 0.1, 1, 10, 100\}$. Furthermore, in case of RF, a grid search approach was employed, considering different values for the number of estimators and the maximum depth. The grid search involved testing combinations of $\{50, 100, 200\}$ for the number of estimators and $\{5, 10, 20\}$ for the maximum depth. By

systematically testing the different combinations, the SVM and RF models could identify the most suitable parameters for the given problem.

Moreover, in each fold of the outer k-fold CV scheme the obtained feature matrix is subjected to standardization, followed by an ANOVA feature selection analysis [42], from where the 40 best features are kept.

2) *Deep Learning*: CNNs and Recurrent Neural Networks (RNNs) are well-known deep neural networks that have been proven extremely useful in complex classification problems due to their ability to learn complex feature maps of the input [43], [44]. Regarding the DL procedure within the SLNet classification scheme (Fig. 1), the widely used CNN architecture of ZFNet, is employed, after being modified according to the dimension of the input data, as proposed in [22], and combined with two widely used Deep RNNs (DRNNs), the Gated Recurrent Unit (GRU) [45] and the Bi-Directional Long Short-Term Memory (BiLSTM) [46].

In particular, the developed hybrid ZFNet consists of five layers. In the first layer, 96 kernels of size 7×1 are stacked to a 3×1 max-pooling, followed by a batch normalization. In the second layer, a convolution layer with 256 kernels of size 5×1 is followed by 3×1 max-pooling and batch normalization. The last three layers consist of three convolution and three 3×1 max-pooling layers. The convolution layers have 512, 1024 and 512 filters with kernel sizes of 3×1 , respectively. Consequently, the output of the ZFNet is reshaped to 2×4608 and fed to a two-layer two-cell stacked DRNN (BiLSTM, or GRU). The output dimension of the DRNN is set to 64×2 and 128×2 for the GRU and the BiLSTM, respectively. Finally, a global average pooling layer is added to flatten the data, followed by two fully connected layers with 37 and two nodes.

For compiling the model, the categorical cross entropy loss and the Adam optimizer [47] were used. Moreover, a learning rate scheduler that exponentially decreases the learning rate after 30 epochs was adopted, together with an early stopping criterion that terminates the training process if the validation loss does not improve for 10 consecutive epochs. The maximum number of epochs was set to 100.

The SHAP values were extracted using the *GradientExplainer* method of the SHAP Python library, for the test set during both the 5-fold and the 10-fold CV schemes. Consequently, their absolute value is calculated and averaged across all the folds and samples to produce a single value, indicative of the significance of each one of the four inputs (PHRV, histogram of the D_2 and D_3 detail coefficients and amplitude of R-peaks). Following, the obtained values are normalized in order to highlight the percentage contribution of each feature to the model's decision.

The performance evaluation metrics were estimated across each classification path (ML/DL) and without ({SVM}, {RF}, {ZFNet-GRU}, {ZFNet-BiLSTM}) and with ({SVM, RF, ZFNet-GRU}, {SVM, RF, ZFNet-BiLSTM}) soft voting, to evaluate the contribution of each classification path and showcase the efficiency of ML/DL combination in the SLNet classification pipeline. Furthermore, to assess the effect of various anthropometric and SA-related factors on the efficacy

TABLE II: CLASSIFICATION PERFORMANCE OF SLNET FOR 5-/10-CROSS-VALIDATION (CV) SCHEMES. $\pm 95\%$ CI IS GIVEN WHERE APPLICABLE. BOLDFACE INDICATES THE MAXIMUM PERFORMANCE

Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
5-fold CV				
SVM	91.54	88.25	93.58	88.87
RF	91.22	88.20	93.09	88.50
ZFNet-GRU	91.93 \pm 0.13	89.51 \pm 0.35	93.43 \pm 0.12	89.46 \pm 0.18
ZFNet-BiLSTM	91.91 \pm 0.14	89.84 \pm 0.45	93.19 \pm 0.17	89.47 \pm 0.20
{SVM, RF, ZFNet-GRU}	92.78 \pm 0.08	90.59 \pm 0.23	94.15\pm0.09	90.58 \pm 0.11
{SVM, RF, ZFNet-BiLSTM}	92.81\pm0.07	90.89\pm0.31	94.00 \pm 0.12	90.64\pm0.10
10-fold CV				
SVM	91.69	88.44	93.71	89.07
RF	91.36	88.50	93.13	88.69
ZFNet-GRU	92.06 \pm 0.14	89.85 \pm 0.24	93.43 \pm 0.14	89.65 \pm 0.18
ZFNet-BiLSTM	92.01 \pm 0.07	90.06 \pm 0.33	93.22 \pm 0.18	89.62 \pm 0.10
{SVM, RF, ZFNet-GRU}	92.91\pm0.10	90.86 \pm 0.18	94.18\pm0.11	90.75\pm0.13
{SVM, RF, ZFNet-BiLSTM}	92.88 \pm 0.03	91.04\pm0.21	94.02 \pm 0.13	90.73 \pm 0.05

of SLNet, we perform non-parametric statistical tests using, the Mann-Whitney U test.

For the implementation of the ML procedures the scikit-learn Python library [42] was used. Additionally, the DL architectures were designed and compiled using the Keras [48] and Tensorflow [49] libraries, while for model interpretation, the SHAP Python library [50] was employed. All the analyses were performed using an HP Victus 16 laptop with the following specifications: 11th Gen Intel Core i7 - 2.3 GHz processor, 32 GB RAM and Windows 11 operating system. The computations were accelerated by an NVIDIA GeForce RTX 3060 GPU with 6GB VRAM.

V. RESULTS AND DISCUSSION

A. Classification results

Table II and Fig. 3 present the classification results yielded by SLNet. More specifically, Table II tabulates the Accuracy, Sensitivity, Specificity and F1 scores, along with the 95% Confidence Interval (CI), for all classifiers (without and with soft voting) and for both the 5-/10-fold CV schemes. Furthermore, Fig. 3 illustrates the ROC curves of the examined models, for the 5-fold (Fig. 3a) and the 10-fold (Fig. 3b) CV schemes, respectively. From both Table II and Fig. 3 it is evident that SLNet provides a robust classification outcome, with steady performance under both CV schemes, without and with soft voting. Moreover, according to Table II, the DL models achieve slightly better overall performance than ML models. Additionally, both ML and DL models are characterized by a sensitivity/specificity imbalance, with the latter being higher in all cases. Clearly, the combination of the ML and DL models through the soft voting scheme boosted the results (all metrics $>90\%$) and achieved the best performance for both DL models (ZFNet-GRU and ZFNet-BiLSTM) when combined with the ML models (SVM, RF). This highlights the efficiency of SLNet in accurately discriminating between the SA/non-SA classes by combining knowledge from both ML and DL domains.

B. Feature Characterization

1) *Machine Learning*: For the feature selection and characterization, an ANOVA feature selection process took place

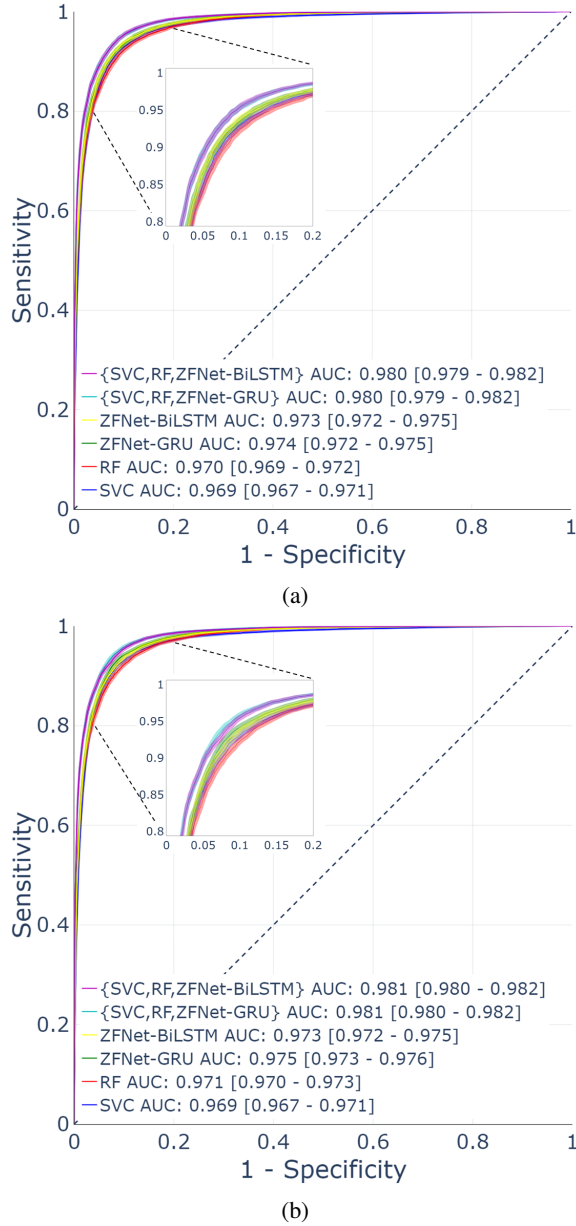


Fig. 3: Comparison of ROC curves of all classifiers for (a) 5-fold and (b) 10-fold CV. Bold lines represent the mean ROC curve, whereas shadowed intervals denote the 95% CI, computed over 500 bootstraps. In the legend the AUC score and the 95% confidence interval are shown for each model.

inside each fold of the 5/10-fold CV schemes and applied to the 120 employed features. From them, 43 were selected at least once, with 36 of them being selected across all folds. Consequently, the importance of each of these features, calculated during the feature selection, was averaged across the folds of both the 5-fold and the 10-fold CV schemes, in order to get the final importance score of each feature. Figure 4 depicts the importance of the best selected features, produced by the ANOVA feature selection process. From Fig. 4 it is evident that both MRA and HRV analyses provide valuable features. In particular, 26 out of the 43 features come from

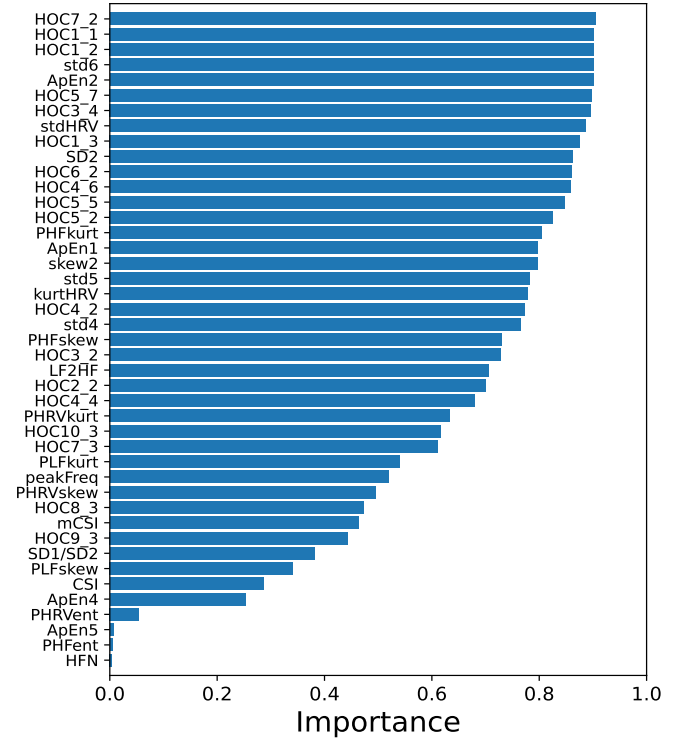


Fig. 4: Scores of the best ML features (descending order) produced by the ANOVA feature selection process.

the MRA accounting for a total importance score of 15.27. On the other hand, the remaining 21 features produced by the HRV analysis, account for a total importance score of 12.22.

Furthermore, Fig. 4 indicates that the D_2 and D_3 MRA detail coefficients are the most valuable across the rest of D_i , justifying their selection as input to the DL model. Moreover, Fig. 4 highlights the effectiveness of HOCs in the SA episode detection (see also Fig. 2(c)). In particular, HOCs account for 17 out of the 43 features considered, with six of them ranking among the top 10 best features. This is of high importance, as to the best of our knowledge, this is the first study that utilizes HOCs for analyzing ECG signals at different MRA levels.

Moreover, the frequency content of the HRV constitutes a good biomarker of SA. More specifically, 11 out of the 44 features presented in Fig. 4 originate from the power spectrum of the HRV. The HF band is found to be the most important with a total feature importance score of 3.11. This can be attributed to the fact that the HF band of the HRV, which is associated with the parasympathetic activity of the autonomous nervous system, reflects the heart rate variations associated with the respiratory cycle [34]. Additionally, apart from the HF band, the LF band of the HRV also provides features of high importance, that account for a total feature importance score of 2.53. The LF band of the HRV relates with both sympathetic and parasympathetic activity and it has been previously associated with SA [51].

Finally, Poincaré plot and time-domain features of the HRV have a lower participation in the most important features, contributing with four and two features, respectively.

TABLE III: DL FEATURE IMPORTANCE (%) THROUGH THE SHAP GRADIENT EXPLAINER ($\pm 95\%$ CI IS ALSO PROVIDED)

Classifier	PHRV	D_2	D_3	R-peaks Amplitude
5-fold CV				
ZFNet-GRU	9.41 ± 3.84	16.63 ± 4.06	11.10 ± 1.51	62.86 ± 8.79
ZFNet-BiLSTM	11.51 ± 3.67	19.4 ± 4.87	10.56 ± 1.50	58.53 ± 7.36
10-fold CV				
ZFNet-GRU	9.81 ± 1.77	18.30 ± 3.34	11.93 ± 0.89	59.96 ± 3.58
ZFNet-BiLSTM	7.80 ± 0.95	15.70 ± 1.73	11.59 ± 0.86	64.91 ± 3.16

2) *Deep Learning*: Table III tabulates the results of the SHAP analysis for the case of DL. From Table III, it is evident that the R-Peak amplitude is by far the most significant input to the DL models, followed by the D_2 and D_3 detail coefficients, while the least significant input seems to be the PHRV.

C. Comparison with Other Approaches

Recently, multiple studies have attempted to build models that can detect SA episodes using single-lead ECG recordings. Table IV compares the performance of the proposed SLNet scheme to the one from some recent, state-of-the-art SA detection methods that have been applied to the same dataset and have been evaluated under a k-fold CV classification scheme. As it can be seen from Table IV, SLNet performs similarly to [20], [21], [27], yet exhibiting higher Accuracy and Specificity, and superior to [18], [19], [22]–[24], [52]–[54]. However, unlike the majority of these studies ([18]–[21], [23], [52], [53]) that used only half of the available recordings (35), all 70 recordings were included in the performance evaluation of the SLNet, providing a more comprehensive and robust evaluation. Comparing the results of SLNet to the ones from [24], [22], [54] and [27], that have also used all the available recordings, significant improvements are observed. In particular, compared to the results of [24], SLNet achieves accuracy, sensitivity, specificity scores that are approximately 3%, 4% and 3% higher, respectively. Moreover, comparing the results of SLNet to the ones reported in [22], an approximate improvement of 4%, 10% and 1.5% is observed for the accuracy, sensitivity and specificity scores, respectively. Furthermore, compared to [54] SLNet resulted to 1.3% higher accuracy and 2% higher specificity. Finally, compared to [27], our approach yielded 1.5% higher accuracy and 1% higher sensitivity and specificity. Overall, the proposed SLNet matches or outperforms previous methods that utilized the same dataset and performed their classification under the same scheme (k-fold CV).

D. Performance Across Different Groupings

Anthropometric characteristics, including age, gender, and BMI, as well as SA-related indices, such as AHI, AI and HI, can significantly influence the manifestation and severity of SA episodes [55]. Therefore, it is crucial to assess the performance of SLNet considering different groupings of these parameters.

1) *Anthropometric Parameters Grouping*: Figure 5 reports the Accuracy, Sensitivity, Specificity, F1-score and AUC score, specifically focusing on age (age <45 and age ≥ 45), gender and BMI (Healthy: $18.5 \leq \text{BMI} \leq 25$, Overweight: $25 \leq \text{BMI} \leq 30$, and Obese: >30). Moreover, Fig. 5

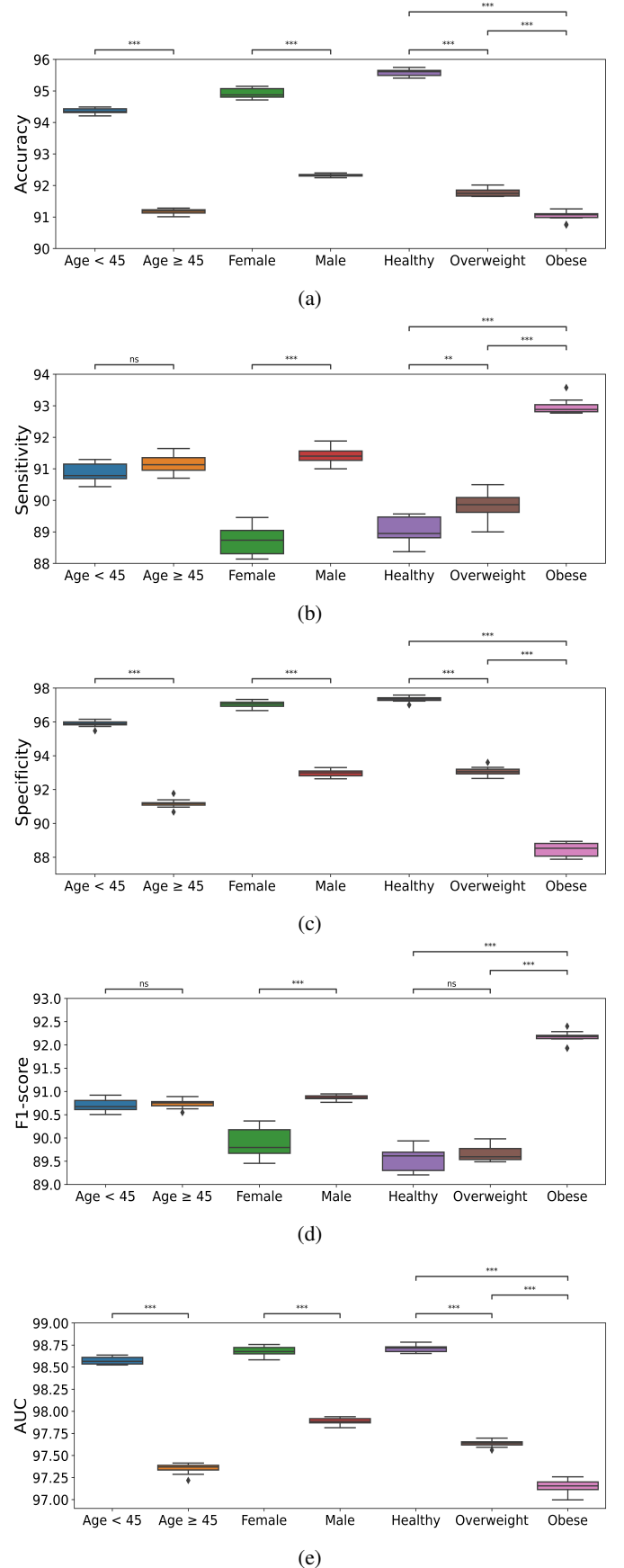


Fig. 5: (a) Accuracy, (b) Sensitivity, (c) Specificity, (d) F1, and (e) AUC scores with respect to age, gender, and BMI. Statistically evaluated differences, estimated through the Mann–Whitney U test, are superimposed (not significant (ns): $p \geq 0.05$, *: $0.01 < p < 0.05$, **: $p \leq 0.01$).

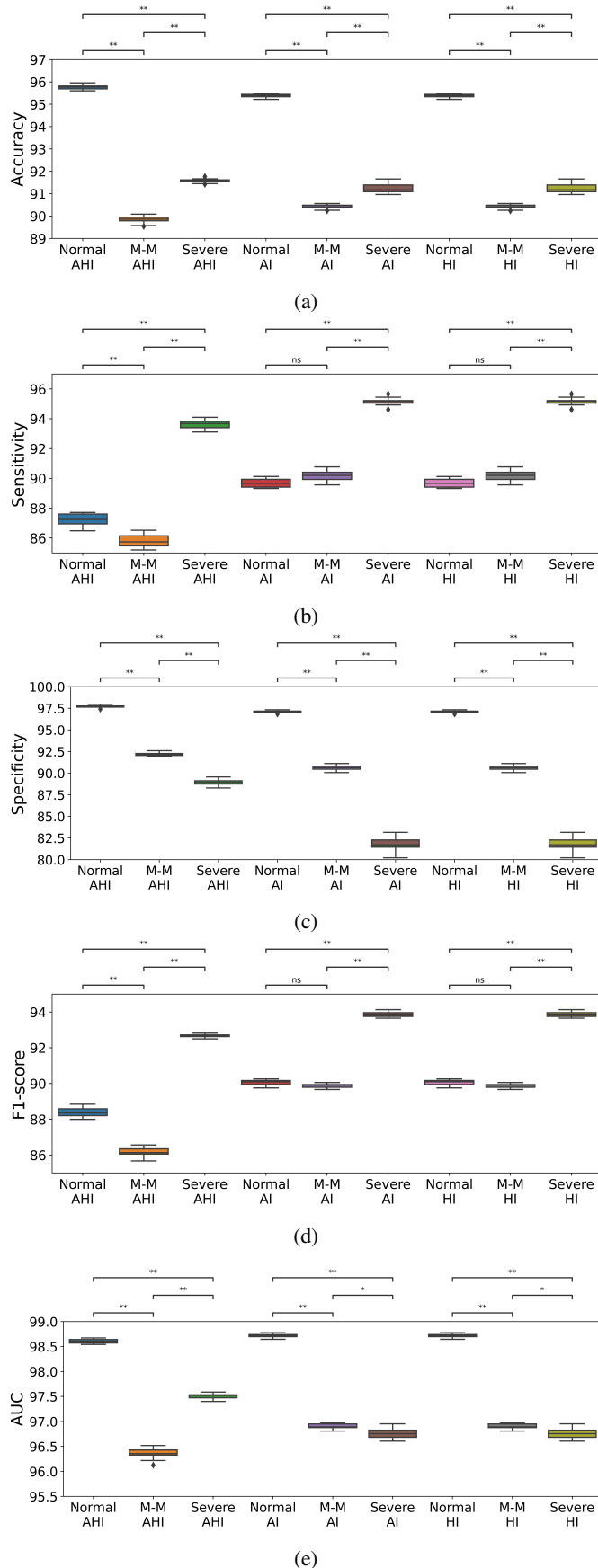


Fig. 6: (a) Accuracy, (b) Sensitivity, (c) Specificity, (d) F1 and (e) AUC scores with respect to the AHI, AI and HI. Statistically evaluated differences, estimated through the Mann–Whitney U test, are superimposed (M-M: Mild and Moderate, ns: $p \geq 0.05$, *: $0.01 < p < 0.05$, **: $p \leq 0.01$).

TABLE IV: PERFORMANCE COMPARISON OF SLNET WITH OTHER APPROACHES APPLIED ON THE SAME APNEA-ECG DATABASE. BOLDFACE INDICATES THE MAXIMUM PERFORMANCE

Study	ML/DL	CV Folds (k)	Accuracy (%)	Sensitivity (%)	Specificity (%)
Nishad et al. (2018) [21]*	ML	10	92.78	93.91	90.95
Janbakhshi et al. (2018) [18]*	ML	35	89.60	85.4	91.90
Pombo et al. (2020) [19]*	ML	10	82.12	88.41	72.29
Fatimah et al. (2020) [20]*	ML	10	92.59	89.70	94.67
Faai et al. (2021) [52]*	ML	5	81.43	76.63	84.40
Shen et al. (2021) [23]*	DL	10	89.40	89.80	89.10
Rajesh et al. (2021) [53]*	ML	10	90.30	86.6	92.59
Yang et al. (2022) [24]	DL	10	89.70	87.00	90.90
Bahrami et al. (2022) [22]	ML&DL	5	88.13	81.49	92.27
Shao et al. (2022) [54]	DL	10	91.5	91.00	91.90
Abasi et al. (2023) [27]	DL	10	91.30	90.10	93.6
SLNet (2024)	ML&DL	5	92.81	90.89	94.00
SLNet (2024)	ML&DL	10	92.88	91.04	94.02

*50% of the available recordings have only been employed in the study

illustrates the results of statistical tests performed using the Mann–Whitney U test to assess the effect of each one of the examined anthropometric factors on the efficacy of SLNet. For simplicity, only the results from the 10-fold cross-validation of our best performing model ($\{SVM, RF, ZFNet-BiLSTM\}$) are presented.

It is evident from Fig. 5 that the influence of all the examined anthropometric factors on accuracy, specificity and AUC is statistically significant (see Figs. 5a, 5c and 5e). Considering sensitivity, only the effect of age was found not significant (see Fig. 5b). Furthermore, regarding F1-score, all factors, except age and BMI (between healthy and overweight), were found to be statistically significant (see Fig. 5d). However, despite the statistical significance observed for the majority of anthropometric factors influencing the performance of SLNet, it is important to note that the obtained scores are sustained high in all cases.

Figure 5 demonstrates that SLNet maintains, in general, robust performance across all anthropometric grouping classification scenarios. However, some notable differences can be identified. Considering the factor of age, Fig. 5 suggests that SLNet achieves higher specificity and slightly higher AUC score in the lower age range. No significant differences can be identified for sensitivity and F1-score. Regarding gender classification, SLNet demonstrates similar F1-score and AUC score for male and female classes. However, there is a significant difference in sensitivity and specificity scores. According to Fig. 5, SLNet achieved higher sensitivity scores when tested in the male subgroup, whereas the opposite trend was observed for the female one.

Moreover, an interesting trend is observed when examining the classification score with regards to BMI (see Fig. 5). Particularly, as the BMI scores increases, the sensitivity score of SLNet increases, while the specificity score decreases. This trend can be attributed to the association of BMI with the severity of SA [56]. Therefore, with increasing BMI, the severity of SA episodes escalates and, as a result, the model's ability in detecting these episodes improves. On the other hand, the decreasing specificity scores can be explained by the exact opposite pattern. Given the different ratios of SA to non-SA segments that exist in the three BMI subgroups, F1-score is also an important metric. According to Fig. 5, the F1-score achieved by SLNet improves with increasing BMI, as

expected. It is important to interpret results related to gender and age cautiously, as they are not entirely independent of BMI. The dataset shows that subjects under 45 years old tend to have a lower average BMI than those older than 45, and males have a higher average BMI than females. Therefore, the observed trends in age and gender in Fig. 5 may be influenced by this co-dependence.

2) *SA Severity Level Grouping*: Figure 6 summarizes the classification results with respect to AHI (normal: $AHI \leq 5$, M-M: $5 < AHI \leq 30$, and severe: $AHI > 30$), AI (normal: $AI \leq 2.5$, M-M: $2.5 < AI \leq 15$, and severe: $AI > 15$) and HI (normal: $HI \leq 2.5$, M-M: $2.5 < HI \leq 15$, and severe: $HI > 15$). Furthermore, Fig. 6 depicts the results of the Mann–Whitney U tests that were performed to evaluate the effects of SA severity on the efficacy of SLNet. As in Fig. 5, only the results from the 10-fold cross-validation of our best performing model are shown.

Considering the SA-related indices of AHI, AI and HI, Fig. 6 illustrates that for the metrics of accuracy, specificity and AUC score, all the statistical tests performed between normal, M-M, and severe SA showed statistically significant differences (see Figs. 6a, 6c and 6d). Not statistically significant differences were only found between the pairs (normal and M-M AI) and (normal and M-M HI) for the sensitivity and F1-score metrics (see Figs. 6b and 6d, respectively). Notwithstanding the statistical significance found for the majority of SA-related factors, it is crucial to emphasize that SLNet achieved high scores across all cases.

Furthermore, since AHI, AI and HI directly describe the severity of SA episodes, we expect the results of Fig. 6 to have a trend similar to that of BMI. Specifically, we anticipate that the sensitivity will improve with increasing AHI, AI, and HI values, while the specificity will exhibit the opposite trend. Although this pattern is indeed observed for AI and HI, it is only partially true in the case of AHI. More specifically, the sensitivity score of the normal AHI group is approximately 1.5% higher than that of the M-M group, which is unexpected, yet it can possibly be attributed to the unbalanced combination in the number of apnea and hypopnea episodes involved in the estimation of AHI. However, as anticipated, the sensitivity score of the severe AHI group is the highest. Another significant outcome arising from Fig. 6 is that the F1-score of the severe AHI, AI and HI groups is consistently greater than the normal and M-M groups. Finally, the obtained AUC scores are constantly above 96%, showcasing the efficiency of SLNet in detecting SA episodes.

Overall, with this analysis we aimed to isolate and examine the effects of the various anthropometric and SA severity level factors on the effectiveness of SLNet. The results of this analysis demonstrated that although statistically significant differences exist between various groupings, SLNet consistently maintains high performance across all examined scenarios. This underscores the robustness and efficiency of the proposed SLNet in detecting episodes of SA.

E. Implications, Limitations and Future Directions

Implications. Various implications can be drawn from the proposed work. From a practical standpoint, the presented

SLNet provides a comprehensive framework that combines advanced feature extraction from single-lead ECG signals with sophisticated AI-based techniques. This integration enables the development of predictive models capable of accurately detecting episodes of SA, achieving state-of-the-art performance.

Moreover, although the developed SLNet framework has been employed in this study for SA detection, it is essential to note that SLNet is a versatile single-lead ECG analysis framework that can be applied to various other single-lead ECG analyses, as well. Its effectiveness lies in its ability to process and extract meaningful information from the ECG signals. Furthermore, by leveraging the power of ML and DL models and the interpretability of single-lead ECG signals, SLNet holds potential for applications beyond SA, enabling researchers and clinicians to explore its utility in detecting and monitoring other cardiac abnormalities or even broader health monitoring scenarios.

From a clinical perspective, our study demonstrates the feasibility of an end-to-end system that can be seamlessly integrated into commercial devices, such as smartbelts, to enable an unobtrusive and continuous monitoring of individuals during their sleep [57]. By utilizing such a system, valuable information regarding the presence, severity, and frequency of SA episodes can be collected. Furthermore, the frequency of SA events recorded by the system can provide insights into the temporal patterns and possible triggers of SA episodes. This information holds significant potential for clinicians to gain deeper insights into the condition.

Moreover, the presented SLNet could also serve as part of a digital twin framework, which is rapidly advancing, leading to rapid improvements in personalized healthcare [58]. In particular, considering SA, digital twins technology holds great potential to further improve the management of SA [59]. By creating virtual replicas of individuals based on their physiological and sleep-related data, clinicians can simulate and predict the effects of various treatment interventions on each individual patient. This innovative approach empowers healthcare professionals to explore different scenarios and optimize treatment plans before implementing them in real-life settings [60]. Thus, from a clinical perspective, the integration of SLNet into a digital twin system could add a feature-based dimension of information to the framework, enabling clinicians to gain a more comprehensive understanding of each patient's condition.

In the same vein, by leveraging the insights gained from the ML/DL models used for SA detection, the identified features can serve as potential biomarkers that could continuously be monitored throughout the day to assess the effects of SA on individuals' daily lives [61]. These biomarkers can be utilized as inputs to personalized models that can detect states of extreme fatigue or drowsiness, providing valuable information to guide patients in taking proactive measures [62]. For instance, if the model detects high levels of fatigue or drowsiness, it can prompt individuals to take short breaks, engage in physical activity, or avoid potentially hazardous activities, such as driving. Moreover, it could provide some evidence for potential risk in somnolence-associated pathologies, such as Parkinson's Disease [63].

Finally, another potential clinical application of the SLNet framework is monitoring for SA episodes in infants. Some researchers have found an association between SA and Sudden Infant Death Syndrome (SIDS) [64], while others argue that SA and SIDS are unrelated [65]. Nonetheless, if SA does contribute to SIDS, early monitoring could be highly beneficial in preventing fatal consequences. However, accurately measuring and analyzing multimodal physiological signals in infants pose challenges. Thus, utilizing a single-lead ECG based analysis could be a possibly efficient solution. It is important to note that there are inherent differences between adult and infant ECGs [66], so fine-tuning SLNet using the appropriate data would be necessary.

Limitations. Two main limitations are acknowledged here. The first limitation arises from the fact that only one dataset was used. Testing the proposed methodology on more than one datasets, collected from different medical centers would render the results more robust, with high generality. Secondly, our study utilized the PhysioNet Apnea-ECG Database, which has certain shortcomings, including a small number of M-M SA patients and the absence of CSA and MSA episodes. Testing on different types of SA is crucial to ensure that the proposed methodology can maintain its efficiency in detecting CSA and MSA episodes too.

Future Directions. In subsequent studies, efforts will be directed towards overcoming the aforementioned limitations and enhancing the performance of SLNet. Therefore, to further assess the capabilities of SLNet, future work will focus on testing its performance on additional datasets that encompass multiple types of SA. Moreover, the analysis and results presented in this study were based on 1-min ECG segments. To enhance the classification performance, future investigations could explore the inclusion of information from previous segments into the analysis of each segment. This consideration may offer further improvements in the overall classification performance. Finally, using more sophisticated, data-driven decomposition techniques, such as swarm decomposition [67], is one additional potential improvement that could enhance the performance of SLNet.

VI. CONCLUSION

A comprehensive feature extraction procedure for SA detection using single-lead ECG signals, namely SLNet, has been presented here. When SLNet was applied to experimental data drawn from patients with different levels of SA severity, highly accurate results were obtained. Comparing the performance of the proposed SLNet approach to previous studies that employed the same dataset, similar or superior classification results were obtained. These findings highlight the potential contribution of SLNet towards the development of convenient and cost-effective tools for detecting SA episodes using single-lead ECG data, scaffolding the efficiency of low-cost, wearable devices for ambulatory SA detection.

APPENDIX A

HIGHER-ORDER-CROSSINGS

The backward difference operator, denoted by ∇ and defined as $\nabla Z_t = Z_t - Z_{t-1}$, acts as a high-pass filter. To

create a sequence of high-pass filters, we define $J_k = \nabla^{k-1}$ for $k = 1, 2, 3, \dots$. Then, the simple HOC can be estimated as $D_k = NZC_k\{Z_t\}$ for $k = 1, 2, 3, \dots$, and $t = 1, \dots, N$, where $NZC_k\{\cdot\}$ denotes the estimation of the number of zero-crossings at order k and

$$\nabla^{k-1} Z_t = \sum_{j=1}^k \binom{k-1}{j-1} (-1)^{j-1} Z_{t-j+1}. \quad (1)$$

In practice, time series have a finite number of data points and, as a result, a sample is lost after each difference. Thus, to cope with this phenomenon, the data must be indexed by moving to the right, i.e., for the evaluation of k simple HOC, the index $t = 1$ should be given to the k_{th} or a later sample. To calculate the number of zero-crossings in (1), a binary time series $X_t(k)$ is initially formed, as follows:

$$X_t(k) = \begin{cases} 1, & \text{if } J_k(Z_t) \geq 0 \\ 0, & \text{if } J_k(Z_t) < 0 \end{cases} \quad k = 1, 2, \dots; t = 1, \dots, N. \quad (2)$$

Finally, the desired simple HOC of order k is estimated by counting the symbol changes in $X_1(k), \dots, X_N(k)$, i.e.,

$$D_k = \sum_{t=2}^N [X_t(k) - X_{t-1}(k)]^2. \quad (3)$$

REFERENCES

- [1] L. Drager, R. McEvoy, F. Barbe, G. Lorenzi-Filho, and S. Redline, "Sleep apnea and cardiovascular disease: lessons from recent trials and need for team science," *Circulation*, vol. 136, no. 19, pp. 1840–1850, 2017.
- [2] C. N. Kaufmann, R. Susukida, and C. A. Depp, "Sleep apnea, psychopathology, and mental health care," *Sleep Health*, vol. 3, no. 4, pp. 244–249, 2017.
- [3] M. Grigg-Damberger and N. Foldvary-Schaefer, "Bidirectional relationships of sleep and epilepsy in adults with epilepsy," *Epilepsy & Behavior*, vol. 116, p. 107735, 2021.
- [4] E. K. Seng, C. Cervoni, J. L. Lawson, T. Oken, S. Sheldon, M. D. McKee, and K. A. Bonuck, "The burden of sleep problems: A pilot observational study in an ethnically diverse urban primary care setting," *Journal of Primary Care & Community Health*, vol. 7, no. 4, pp. 276–280, 2016.
- [5] J. Streatfeild, J. Smith, D. Mansfield, L. Pezzullo, and D. Hillman, "The social and economic cost of sleep disorders," *Sleep*, vol. 44, no. 11, pp. 132–142, 2021.
- [6] D. Burman, "Sleep Disorders: Sleep-Related Breathing Disorders," *FP Essentials*, vol. 460, pp. 11–21, 2017.
- [7] F. Ralls and L. Cutchen, "A contemporary review of obstructive sleep apnea," *Current Opinion in Pulmonary Medicine*, vol. 25, no. 6, pp. 578–593, 2019.
- [8] G. Paidi, A. Beesetty, M. Jean, F. P. A. Greye, T. Siyam, M. F. Fleming, J. Nealy, L. Kop, R. Sandhu, and R. S. Sandhu, "The management of obstructive sleep apnea in primary care," *Cureus*, vol. 14, no. 7, 2022.
- [9] W. Randerath, J. Verbraecken, S. Andreas, M. Arzt, K. E. Bloch, T. Brack, B. Buyse, W. De Backer, D. J. Eckert, L. Grote *et al.*, "Definition, discrimination, diagnosis and treatment of central breathing disturbances during sleep," *European Respiratory Journal*, vol. 49, no. 1, 2017.
- [10] M. H. Araghi, Y.-F. Chen, A. Jagielski, S. Choudhury, D. Banerjee, S. Hussain, G. N. Thomas, and S. Taheri, "Effectiveness of lifestyle interventions on obstructive sleep apnea (OSA): systematic review and meta-analysis," *Sleep*, vol. 36, no. 10, pp. 1553–1562, 2013.
- [11] J. V. Rundo and R. Downey III, "Polysomnography," *Handbook of Clinical Neurology*, vol. 160, pp. 381–392, 2019.
- [12] M. Santilli, E. Manciocchi, G. D'Addazio, E. Di Maria, M. D'Attilio, B. Femminella, and B. Sinjari, "Prevalence of obstructive sleep apnea syndrome: a single-center retrospective study," *International Journal of Environmental Research and Public Health*, vol. 18, no. 19, p. 10277, 2021.

- [13] X. Zhao, X. Wang, T. Yang, S. Ji, H. Wang, J. Wang, Y. Wang, and Q. Wu, "Classification of sleep apnea based on eeg sub-band signal characteristics," *Scientific Reports*, vol. 11, no. 1, pp. 1–11, 2021.
- [14] R. Lazazzera, M. Deviaene, C. Varon, B. Buyse, D. Testelmans, P. Laguna, E. Gil, and G. Carrault, "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.
- [15] F. Senny, J. Destin  , and R. Poirrier, "Midsagittal jaw movement analysis for the scoring of sleep apneas and hypopneas," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 1, pp. 87–95, 2007.
- [16] L. Wang, Y. Lin, and J. Wang, "A RR interval based automated apnea detection approach using residual network," *Computer Methods and Programs in Biomedicine*, vol. 176, pp. 93–104, 2019.
- [17] S. Ucak, H. U. Dissanayake, K. Sutherland, P. de Chazal, and P. A. Cistulli, "Heart rate variability and obstructive sleep apnea: Current perspectives and novel technologies," *Journal of Sleep Research*, vol. 30, no. 4, p. e13274, 2021.
- [18] P. Janbakhshi and M. Shamsollahi, "Sleep apnea detection from single-lead ECG using features based on ECG-derived respiration (EDR) signals," *IRBM*, vol. 39, no. 3, pp. 206–218, 2018.
- [19] N. Pombo, B. M. Silva, A. M. Pinho, and N. Garcia, "Classifier precision analysis for sleep apnea detection using ECG signals," *IEEE Access*, vol. 8, pp. 200 477–200 485, 2020.
- [20] B. Fatimah, P. Singh, A. Singhal, and R. B. Pachori, "Detection of apnea events from ECG segments using Fourier decomposition method," *Biomedical Signal Processing and Control*, vol. 61, p. 102005, 2020.
- [21] A. Nishad, R. B. Pachori, and U. R. Acharya, "Application of TQWT based filter-bank for sleep apnea screening using ECG signals," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–12, 2018.
- [22] M. Bahrami and M. Forouzanfar, "Sleep apnea detection from single-lead ECG: a comprehensive analysis of machine learning and deep learning algorithms," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022.
- [23] Q. Shen, H. Qin, K. Wei, and G. Liu, "Multiscale deep neural network for obstructive sleep apnea detection using RR interval from single-lead ECG signal," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2021.
- [24] Q. Yang, L. Zou, K. Wei, and G. Liu, "Obstructive sleep apnea detection from single-lead electrocardiogram signals using one-dimensional squeeze-and-excitation residual group network," *Computers in Biology and Medicine*, vol. 140, p. 105124, 2022.
- [25] S. Hu, W. Cai, T. Gao, and M. Wang, "A hybrid transformer model for obstructive sleep apnea detection based on self-attention mechanism using single-lead ecg," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022.
- [26] C. Li, Z. Shi, L. Zhou, Z. Zhang, C. Wu, X. Ren, X. Hei, M. Zhao, Y. Zhang, H. Liu *et al.*, "Ttformer: A time frequency information fusion based cnn-transformer model for osa detection with single-lead ecg," *IEEE Transactions on Instrumentation and Measurement*, 2023.
- [27] A. K. Abasi, M. Aloqaily, and M. Guizani, "Optimization of cnn using modified honey badger algorithm for sleep apnea detection," *Expert Systems with Applications*, vol. 229, p. 120484, 2023.
- [28] M. Khalil and O. Rifaie, "Electrocardiographic changes in obstructive sleep apnoea syndrome," *Respiratory Medicine*, vol. 92, no. 1, pp. 25–27, 1998.
- [29] L. Bacharova, E. Triantafyllou, C. Vazaios, I. Tomeckova, I. Paranicova, and R. Tkacova, "The effect of obstructive sleep apnea on QRS complex morphology," *Journal of Electrocardiology*, vol. 48, no. 2, pp. 164–170, 2015.
- [30] P. Hamilton, "Open source ECG analysis," in *Computers in Cardiology*. IEEE, 2002, pp. 101–104.
- [31] L. Chen, X. Zhang, and C. Song, "An automatic screening approach for obstructive sleep apnea diagnosis based on single-lead electrocardiogram," *IEEE Transactions on Automation Science and Engineering*, vol. 12, no. 1, pp. 106–115, 2014.
- [32] C.-C. Chen and F. R. Tsui, "Comparing different wavelet transforms on removing electrocardiogram baseline wanders and special trends," *BMC Medical Informatics and Decision making*, vol. 20, pp. 1–10, 2020.
- [33] A. Zarei and B. M. Asl, "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*, vol. 23, no. 3, pp. 1011–1021, 2018.
- [34] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in Public Health*, p. 258, 2017.
- [35] C. Mermigkis, D. Mermigkis, G. Varouchakis, S. Schiza, and P. Panagou, "Poincare plot in obstructive sleep apnoea patients before and after CPAP treatment," *European Respiratory Journal*, vol. 34, no. 5, pp. 1197–1198, 2009.
- [36] J. Jeppesen, S. Beniczky, P. Johansen, P. Sidenius, and A. Fuglsang-Frederiksen, "Using Lorenz plot and Cardiac Sympathetic Index of heart rate variability for detecting seizures for patients with epilepsy," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2014, pp. 4563–4566.
- [37] O. A. Rosso, H. Larrondo, M. T. Martin, A. Plastino, and M. A. Fuentes, "Distinguishing noise from chaos," *Physical Review Letters*, vol. 99, no. 15, p. 154102, 2007.
- [38] S. Pincus, "Approximate entropy (ApEn) as a complexity measure," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 5, no. 1, pp. 110–117, 1995.
- [39] P. C. Petrantonas and L. J. Hadjileontiadis, "Emotion recognition from EEG using higher order crossings," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 186–197, 2009.
- [40] M. Sundararajan and A. Najmi, "The many Shapley values for model explanation," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9269–9278.
- [41] T. Penzel, G. B. Moody, R. G. Mark, A. L. Goldberger, and J. H. Peter, "The apnea-ECG database," *Computers in Cardiology*, vol. 27, pp. 255–258, 2000.
- [42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.
- [43] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai *et al.*, "Recent advances in convolutional neural networks," *Pattern recognition*, vol. 77, pp. 354–377, 2018.
- [44] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, and S. Valaee, "Recent advances in recurrent neural networks," *arXiv preprint arXiv:1801.01078*, 2017.
- [45] K. Cho, B. Van Merri  nboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.
- [46] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
- [47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [48] F. Chollet *et al.* (2015) Keras. [Online]. Available: <https://github.com/fchollet/keras>
- [49] M. A. *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," 2015, software available from tensorflow.org. [Online]. Available: <http://tensorflow.org/>
- [50] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
- [51] V. C. C. Sequeira, P. M. Bandeira, and J. C. M. Azevedo, "Heart rate variability in adults with obstructive sleep apnea: a systematic review," *Sleep Science*, vol. 12, no. 3, p. 214, 2019.
- [52] M. Faal and F. Almasganj, "Obstructive sleep apnea screening from unprocessed ECG signals using statistical modelling," *Biomedical Signal Processing and Control*, vol. 68, p. 102685, 2021.
- [53] K. N. Rajesh, R. Dhuli, and T. S. Kumar, "Obstructive sleep apnea detection using discrete wavelet transform-based statistical features," *Computers in Biology and Medicine*, vol. 130, p. 104199, 2021.
- [54] S. Shao, G. Han, T. Wang, C. Song, C. Yao, and J. Hou, "Obstructive sleep apnea detection scheme based on manually generated features and parallel heterogeneous deep learning model under iomt," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 12, pp. 5841–5850, 2022.
- [55] I. E. Gabbay and P. Lavie, "Age-and gender-related characteristics of obstructive sleep apnea," *Sleep and Breathing*, vol. 16, pp. 453–460, 2012.
- [56] A. Romero-Corral, S. M. Caples, F. Lopez-Jimenez, and V. K. Somers, "Interactions between obesity and obstructive sleep apnea: implications for treatment," *Chest*, vol. 137, no. 3, pp. 711–719, 2010.
- [57] P. Fontana, N. R. A. Martins, M. Camenzind, M. Boesch, F. Baty, O. D. Schoch, M. H. Brutsche, R. M. Rossi, and S. Annaheim, "Applicability of a textile ECG-belt for unattended sleep apnoea monitoring in a home setting," *Sensors*, vol. 19, no. 15, p. 3367, 2019.

- [58] M. Liu, S. Fang, H. Dong, and C. Xu, "Review of digital twin about concepts, technologies, and industrial applications," *Journal of Manufacturing Systems*, vol. 58, pp. 346–361, 2021.
- [59] J. Guo and Z. Lv, "Application of Digital Twins in multiple fields," *Multimedia tools and applications*, vol. 81, no. 19, pp. 26 941–26 967, 2022.
- [60] Y. Liu, L. Zhang, Y. Yang, L. Zhou, L. Ren, F. Wang, R. Liu, Z. Pang, and M. J. Deen, "A novel cloud-based framework for the elderly healthcare services using digital twin," *IEEE access*, vol. 7, pp. 49 088–49 101, 2019.
- [61] C. Rotariu, C. Cristea, D. Arotaritei, R. G. Bozomitu, and A. Pasarica, "Continuous respiratory monitoring device for detection of sleep apnea episodes," in *2016 IEEE 22nd international symposium for design and technology in electronic packaging (SIITME)*. IEEE, 2016, pp. 106–109.
- [62] E. Bjornsdottir, B. T. Keenan, B. Eysteinsdottir, E. S. Arnardottir, C. Janson, T. Gislason, J. F. Sigurdsson, S. T. Kuna, A. I. Pack, and B. Benediktsdottir, "Quality of life among untreated sleep apnea patients compared with the general population and changes after treatment with positive airway pressure," *Journal of Sleep Research*, vol. 24, no. 3, pp. 328–338, 2015.
- [63] S. Du, Y. Qin, M. Han, Y. Huang, J. Cui, H. Han, X. Ge, W. Bai, X. Zhang, and H. Yu, "Longitudinal mediating effect of depression on the relationship between excessive daytime sleepiness and activities of daily living in parkinson's disease," *Clinical Gerontologist*, pp. 1–10, 2022.
- [64] R. Mathur and N. Douglas, "Relation between sudden infant death syndrome and adult sleep apnoea/hypopnoea syndrome," *The Lancet*, vol. 344, no. 8925, pp. 819–820, 1994.
- [65] L. R. Blackmon, D. G. Batton, E. F. Bell, and W. A. Engle, "Apnea, sudden infant death syndrome, and home monitoring," *Pediatrics*, vol. 111, no. 4, pp. 914–914, 2003.
- [66] V. Palodeto and J. Marques, "Methodology for classification and analysis of neonate and adult ecg," in *World Congress on Medical Physics and Biomedical Engineering 2006: August 27–September 1, 2006 COEX Seoul, Korea "Imaging the Future Medicine"*. Springer, 2007, pp. 1214–1217.
- [67] G. K. Apostolidis and L. J. Hadjileontiadis, "Swarm decomposition: A novel signal analysis using swarm intelligence," *Signal Processing*, vol. 132, pp. 40–50, 2017.

Charalampos Lamprou received his integrated Master's degree in electrical and computer engineering from Aristotle University of Thessaloniki (AUTH), Thessaloniki, Greece, in 2022. Currently, he is conducting his Ph.D. with the Healthcare Engineering Innovation Center, Khalifa University of Science and Technology, Abu Dhabi, UAE. His research interests include image processing, advanced signal processing, artificial intelligence, functional brain connectivity and brain mapping.

Aamna Al Shehhi received the bachelor's degree in software engineering from United Arab Emirates University, Abu Dhabi, United Arab Emirates, in 2009, and the master's degree in computing and information science and the Ph.D. degree in interdisciplinary engineering from the Masdar Institute of Science and Technology, Abu Dhabi, in collaboration with the Massachusetts Institute of Technology (MIT), in 2013 and 2017, respectively. During her Ph.D., she was a part of an exchange program for one semester with MIT, in 2015. Currently, she is an Assistant Professor with the Department of Biomedical Engineering, Khalifa University. Her research interests include causal inference, machine learning, and artificial intelligence for the medical domain.

Thanos Stouraitis received the B.S. degree in physics from the University of Athens, Athens, Greece, in 1979, the M.S. degree in electronic automation from the University of Athens, Athens, Greece, in 1981, the M.Sc. degree in electrical and computer engineering from the University of Cincinnati, Cincinnati, OH, in 1983, and the Ph.D. degree in electrical engineering from the University of Florida, Gainesville, in 1986 (for which he received the Outstanding Ph.D. Dissertation Award). Currently, he is a Professor with the Department of Electrical and Computer Engineering Department, Khalifa University. His current research interests include signal and image processing systems, application-specific processor technology and design, computer arithmetic, and design and architecture of optimal digital systems.

Mohamed Seghier received his PhD degree in neuroimaging and brain mapping from Joseph Fourier University of Grenoble, France. He did a post-doc at Geneva University Hospitals (Switzerland) and later he held the position of Senior Research Fellow at the interdisciplinary Wellcome Centre for Human Neuroimaging, University College London (UCL) in the United Kingdom. Currently, he is a Professor with the Department of Biomedical Engineering, Khalifa University. His research interests include investigating the different facets of inter-subject variability in brain activity and how this can be used as a proxy for the different available systems and strategies used by subjects when performing a given task.

Leontios J. Hadjileontiadis received a Diploma degree in electrical engineering and a Ph.D. degree in electrical and computer engineering from Aristotle University of Thessaloniki (AUTH), Thessaloniki, Greece, in 1989 and 1997, respectively, the Ph.D. degree in music composition from the University of York, York, U.K., in 2004, and the Diploma degree in musicology from AUTH, in 2011. He is currently Chair of the department of Biomedical Engineering, Khalifa University and Professor with the department of Electrical and Computer Engineering, Aristotle University of Thessaloniki. His research interests include advanced signal processing, machine learning, biomedical engineering, affective computing, and active and healthy ageing.