SLEEP-SEE-THROUGH: Explainable Deep Learning for Sleep Event Detection and Quantification From Wearable Somnography

Matteo Rossi, Davide Sala, Dario Bovio, Caterina Salito, Giulia Alessandrelli, Carolina Lombardi, Luca Mainardi, *Member, IEEE*, and Pietro Cerveri

***Abstract*—Evidence is rapidly accumulating that multifactorial nocturnal monitoring, through the coupling of wearable devices and deep learning, may be disruptive for early diagnosis and assessment of sleep disorders. In this work, optical, differential air-pressure and acceleration signals, acquired by a chest-worn sensor, are elaborated into five somnographic-like signals, which are then used to feed a deep network. This addresses a three-fold classification problem to predict the overall signal quality (normal, corrupted), three breathing-related patterns (normal, apnea, irregular) and three sleep-related patterns (normal, snoring, noise). In order to promote explainability, the developed architecture generates additional information in the form of qualitative (saliency maps) and quantitative (confidence indices) data, which helps to improve the interpretation of the predictions. Twenty healthy subjects enrolled in this study were monitored overnight for approximately ten hours during sleep. Somnographic-like signals were manually labeled according to the three class sets to build the training dataset. Both record- and subject-wise analyses were performed to evaluate the prediction performance and the coherence of the results. The network was accurate (0.96) in distinguishing normal from corrupted signals. Breathing patterns were predicted with higher accuracy (0.93) than sleep patterns (0.76). The prediction of irregular breathing was less accurate (0.88) than that of apnea (0.97). In the sleep pattern set, the distinction between snoring (0.73) and noise events (0.61) was less effective. The confidence index associated with the prediction allowed us to elucidate ambiguous predictions better. The saliency map analysis provided useful insights to relate predictions to the input signal content. While preliminary, this work supported the recent perspective on the use of deep learning to detect particular sleep events in multiple somnographic signals, thus representing a step towards bringing the use of AI-based tools for sleep disorder detection incrementally closer to clinical translation.**

***Index Terms*—Sleep disorders, wearable sensors, polysomnography, deep learning, interpretable/explainable AI.**

I. Introduction

APNEA episodes (cessation of breathing), intermittent hypoxia (low blood oxygen), snoring and ceaseless body movements sensibly affect the sleep quality and may contribute ultimately to the development of metabolic imbalance and cardiovascular diseases [1], [2], [3], [4], [5]. Multifactorial nocturnal monitoring is therefore recognized as being of paramount importance for early diagnosis and assessment of sleep disorders [6], [7]. Clinical polysomnography (PSG) is the gold standard technique to acquire biosignals describing multiple functions (e.g., respiration, brain activity, vascular and cardiac condition, muscular activation). It is performed in specialized sleep laboratory settings, and even at home, making use of bulky instrumentation, sensors, and cables [8], [9]. Typical clinical parameters describing sleep quality are the apnea–hypopnea index (AHI) and oxygen desaturation index (ODI). Despite proven efficacy and reliability, PSG acquisition setup is intrusive to sleep, leading the patient into an unwanted stressful state and disrupting the natural nocturnal rest. As a result, the recordings may be affected by substantial bias reducing the informative power of the exam [10], [11]. Another drawback of this technique is represented by the analysis of polysomnography signals. It is an intense, laborious and time-consuming manual activity, at least in part dependent on the qualitative judgment of the operator, with outcomes potentially conditioned by uncertainty [12]. The need for standardization in determining sleep-related events is clinically mandatory but is also useful in order to produce scientific and reproducible data for investigation purposes. Commercial and research tools have been recently proposed to add some degree of automatism in signal processing and sleep parameter extraction. SleepRT is a commercial software (OSG, Waarloos, BE) that provides functions to sleep quality score, identification of respiration events, as central and obstructive apnea, and body

Manuscript received 9 September 2022; revised 10 January 2023; accepted 12 April 2023. Date of publication 14 April 2023; date of current version 3 July 2023. This work was supported in part by Project SOUNDI under Grant 3156807, funded by Regione Lombardia within the framework POR FESR2014-2020, in part by PON “Ricerca e Innovazione” 2014–2020, in part by DM 1061/2021, in part by Italian Ministry of University and Research, and in part by PNRR-PE, Future Artificial Intelligence Research (FAIR) - Italian Ministry of University and Research. *(Corresponding author: Pietro Cerveri.)*

Matteo Rossi, Giulia Alessandrelli, and Pietro Cerveri are with the Department of Electronics, Information and Bioengineering, Politecnico di Milano, 32 20133 Milan, Italy, and also with the Istituto Auxologico Italiano IRCCS, 20021 Milan, Italy (e-mail: matteo2.rossi@polimi.it; giulia.alessandrelli@polimi.it; pietro.cerveri@polimi.it).

Davide Sala and Luca Mainardi are with the Department of Electronics, Information and Bioengineering, Politecnico di Milano, 32 20133 Milan, Italy (e-mail: davide8.sala@mail.polimi.it; luca.mainardi@polimi.it).

Dario Bovio and Caterina Salito are with the Biocubica srl, 2013691 Milan, Italy (e-mail: dario.bovio@biocubica.it; caterina.salito@polimi.it).

Carolina Lombardi is with the Istituto Auxologico Italiano IRCCS, 20145 Milan, Italy, and also with the Universita’ degli Studi di Milano, 20126 Bicocca, Milan, Italy (e-mail: c.lombardi@auxologico.it).

Digital Object Identifier 10.1109/JBHI.2023.3267087

positions. While the product offers semi-automatic labeling tools, the elaboration still requires advanced technical skills and extensive clinical expertise to refine the results. MNE toolbox is an open-source Python package for exploring, visualizing, and analyzing human neuro-physiological data, which provides advanced functions for polysomnography signal pre-processing and sleep stage classification [13]. Nonetheless, several manual interactions and heuristic selections are mandatory. Likewise, SLEEP is an open-source Python software for visualization, analysis, and staging of sleep data [14]. In spite of that, the tool is dedicated to electroencephalographic (EEG) and electromyographic (EMG) signal analysis, disregarding other vital signals. More recently, the coupling of wearable sensors and artificial intelligence tools has attracted a lot of attention to address non-invasive signal recordings and automatic quantification of parameters describing sleep disorders. Many different polysomnographic techniques using wearable devices have been documented, being based on cardiac, respiration, and actigraphy signals [15], [16]. As the extensive analysis of the sleep patterns requires the acquisition of several biosignals, multiple and heterogeneous sensors, distributed on the subject surface, are essential. Despite most sensors, like smartwatches and thorax bands, may track sleep phases and the respiratory rate, respectively, specific sleep patterns of clinical interest (e.g., apnea periods, snoring phases, oxygenation level, subject position changes) cannot be systematically measured unless additional devices are involved as nasal air-flow detectors and finger photoplethysmography (PPG) sensors. In this paper, we proposed the use of a novel chest-worn apparatus, called Soundi by Biocubica Srl (Milan - Italy), able to concurrently register photopletismogram, differential air-pressure (sound-like) and accelerogram [17], [18]. Capitalizing on such a device, the acquired signals were pre-processed to reconstruct respiratory, chest effort, body position, oxygenation and acoustic signals (surrogate somnographic signals), which in turn constitute the input to a deep network that classifies alterations of breathing, features of clinical relevance, during sleep. Feasibility tests were carried out on twenty healthy subjects, who were monitored overnight during sleep for approximately ten hours. Technically, we aimed to prove that the algorithm could detect sleep-related breathing events (e.g., apnea, breathlessness, snoring), of potential clinical relevance, with the same quality as an expert operator performing manual identification.

A. Background

1) Wearable Devices for Sleep Monitoring: Tools for collecting and analyzing sleep data can be grouped into two main classes, namely accelerometer actigraphs, shaped as wristwatch-size devices, and wearable signal recorders integrating one or multiple sensors to monitor respiration by direct nasal airflow or indirectly by acoustic signals, electrical cardiac activity, photoplethysmogram, electroocular activity to cite the most relevant data [15]. Wrist sensors mainly use body acceleration to assess sleep/wake epochs [19]. For instance, ActTrust is a wrist-based actigraph for continuous recording of subject movement, light conditions and temperature [20]. Although such devices read the activity-rest rhythm, actually measured rest may not match sleep. Likewise, Actiwatch by Philips Healthcare records sleep/wake data allowing only the reconstruction of sleep dynamics. Processing static and inertial acceleration led to the quantification of sleep position and/or changes between sleep positions using accelerometers on different body locations, such as the hip [21]. However, the comparison with wrist-worn sensors in wake detection was not favorable. A chest sensor was proposed to record the accelerogram and estimate sleep postural changes, which was demonstrated to be comparable with Actiwatch [22], using PSG as the gold standard [23]. However, the proposed sensor only provided postural data and the small testing cohort, considered in the study, did not allow to extend results to different abnormal sleep patterns. Wrist and chest actigram were compared in combination with heart rate variability, measured by ECG on 18 individuals with no previous history of sleep disorders [24]. Nonetheless, such a study was limited to analyzing sleep/wake patterns. The system named “Apnea MedAssist” performs sleep monitoring and recognizes obstructive apnea episodes by means of single-channel nocturnal ECG [25]. Despite a respiration signal is obtained, the body positions during sleep could not be detected. A micro-electro-mechanical system (MEMS) was proposed for measuring the nasal airflow, which was then processed to detect apnea [26]. While the miniaturization ensured very low intrusiveness, neither the body positions nor the oxygenation was measured. AcuPebble SA100 (Acurable, London - U.K.) is an acoustic sensor (CE mark), placed on the throat surface devoted to the automated diagnosis of obstructive sleep apnea. The device has recently undergone a clinical test confirming its reliability in detecting apnea events [27]. Nonetheless, more complex sleep patterns and cardiac activity cannot be monitored.

2) AI Tools for Somnography Signal Analysis: Sleep-related parameters were computed using tracheal sounds and deep learning techniques and comparing the results with PSG with the tracheal sounds derived from a 3D accelerometer attached to the subject suprasternal notch [28]. To classify apnea events and calculate the AHI, convolutional neural networks (CNN), long short-term memory (LSTM), and fully connected networks (FCN) have been fused. In particular, CNN was used to extract relevant features from the windowed time series, and LSTM was used to encode information across time and for temporal data analysis. FCN was then used to predict the probability of the event. Harnessing a 1D-CNN, chest and abdomen accelerations were processed to discriminate among different breath events such as normal breathing, apnea, coughing, yawning, and sighing [29]. Deep recurrent neural networks (RNN) like LSTM and bidirectional LSTM (BiLSTM) to perform sleep apnea recognition using oronasal thermal airflow, nasal pressure, and abdominal respiratory inductance plethysmography signals [30]. The target was a binary classification between normal and apnea. A dense flexible sensor array for pressure detection and deep residual networks were coupled to estimate positions during sleep [31]. Despite the high rate of identification, the approach was not based on a wearable device, as the sensor was placed on the bed. In addition, respiration and other sleep variables, such as respiration variability or snoring events, were not considered.

B. Work Contributions

Despite several wearable devices and AI-based techniques for somnography signal processing have been recently proposed, the identification of sleep disorders to a larger extent, in home-based settings, has remained subtle so far. Firstly, because of the low level of integration in modern wearables, the acquisition of multiple bio-signals still requires different sensors attached to the body surface. Secondly, deep learning approaches to process all the required signals are not integrated and cannot still describe the quality of the predictions. Adopting the multi-modal Soundi wearable system, the present work describes a novel deep learning architecture, which harnesses the surrogate somnographic signals to automatically classify sleep-related breathing events (e.g., apnea, breathlessness, snoring). Three independent event sets were identified: 1) set #1 (2 classes), named "Signal quality" to distinguish clinically reliable signal from low-quality signal corrupted by systematic noise like body position changes or mechanical interference on the device; 2) set #2 (3 classes), named "Breathing pattern", to distinguish normal breathing from apneas and irregular respiratory/oxygenation patterns, this last featuring some potential clinical interest; 3) set #3 (3 classes), named "Sleep pattern", to distinguish snoring episodes from other sleep noises (e.g., wheeze, rhoncus, groan, teeth grinding), with respect to regular sleep. The network intrinsically provides an index to quantify the prediction reliability and a scalar map that can be visualized to qualitatively verify the coherence of the detected events with the somnographic signals. The following main innovations outline the work scope:

- sensor-fusion approach to robustly compute surrogate somnographic signals from the sensor recordings;
- deep neural architecture to analyze in bundle somnographic signals and identify automatically sleep-related breathing events, along with an overall signal quality;
- integration into the network of a dual (qualitative and quantitative) framework to promote the explainability of the predictions.

II. Materials and Methods

A. Chest-Worn Sensor

The acquisition technology concerns a non-invasive apparatus positioned on the chest surface close to the heart, featuring a simultaneous recording of optical, acoustic, accelerometer, 1-lead electrical, bioimpedance, body, and ambient temperature signals. The apparatus has been patented recently (Patent No. EP3248541A1) on a European scale and is currently under CE marking procedure as a medical device (class II). It features a circular shape of diameter not as much as 6 cm and a thickness of approximately 1 cm, with a weight not exceeding 40 grams. Medically certified double-sided tape ensures that it sticks securely to the chest. The apparatus ensures continuous signal recording for a maximum duration of approximately 24 hrs with all sensors enabled (see [17], [32] for further technical details).

B. Signal Acquisition Protocol

Twenty volunteering participants (4 females and 16 males) who had no previous clinical conditions of the cardiovascular system and no apparent sleep disorder were recruited for this study. On average, they were 45 years old (24–67), with height and weight in the range of 55–95 kg and 160–190 cm, respectively. Subject enrollment and data acquisition were performed according to the descriptive rules reported in the experimental protocol (Opinion 3/2019, dated February 19th, 2019) that received approval from the Politecnico di Milano Ethical Committee. All participants were provided with all the required information about the experimental sessions, and they were asked to sign an informed consent before the tests. The acquisition protocol consisted of a single ten-hour phase: the subject was required to wear the Soundi system, closely placed on the left second intercostal space, before falling asleep. For this study, optical, acoustic, and accelerometer signals were continuously recorded during the night, sampled at frequencies of 25, 400 and 100 Hz, respectively. ECG and bioimpedance electrical signals were not acquired. Following signal acquisitions, recorded raw data were downloaded on a PC, resampled to a common frequency of 400 Hz and then underwent pre-processing.

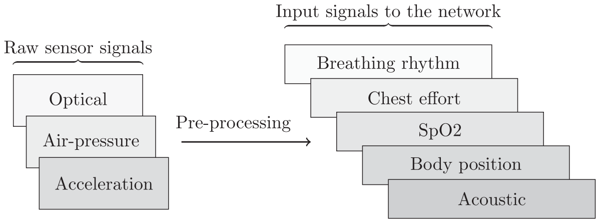

Fig. 1. Data pre-processing: optical, acoustic and accelerometer signals were elaborated to attain surrogate somnographic signals, namely breathing rhythm, chest effort, body position, SpO2 and mechanical vibrations, which constitute the input to the classifier network.

C. Signal Pre-Processing

The pre-processing stage aimed to derive a signal set describing breathing rhythm, chest effort, body position, oxygen saturation and sounds from the raw acquired signals (optical, differential air pressure and acceleration). Especially, the body position was taken into account according to the previous literature that documented positional dependency of both obstructive apnea and snoring [33], [34]. In the output, this stage generated five continuous signals sampled at the same frequency (400 Hz) as the raw signals (Fig. 1). Before pre-processing, the first half hour of the acquisition was removed, considering the subject was not yet asleep. This choice was in agreement with the well-known psychological "first-night effect" of taking longer to fall asleep, especially on the first night of experimentation, compared to the subsequent nights [35].

1) Breathing Rhythm and Chest Effort Signals: Concerning respiration, both differential air pressure and accelerometer signals were taken into account. The pressure signal can be regarded as a multivariate signal conveying different physiological information. Specifically, the respiratory rhythm, featuring a very low-frequency range, was obtained by applying 4th order pass-band (0.05 and 0.6 Hz) Butterworth filter and then standardization based on Z-score transform (zero mean and unitary variance). Likewise, the respiratory effort signal was derived from the displacement of the chest wall sensed by the accelerometer. The signals of the three recorded axes underwent filtering and standardization, identical to the one applied to the acoustic signal. To avoid the dependency on the sleeping position, and retain mostly the chest-wall movement, the absolute value of X, Y, and Z accelerations were computed, and then the maximum value along the timeline was taken, deriving a single time series (Fig. 2).

2) SpO2 Signal: Oxygen saturation by the optical measure of pulsatile blood flow is traditionally quantified as the amount of red ($\lambda = 660\,\mathrm{nm}$) and infrared (IR, $\lambda = 880\,\mathrm{nm}$) light detected by the sensor photodiode. This measure is represented by the modulation factor R [36], obtained as the ratio between baseline (DC) and differential (AC) absorbance components. For the optical sensor used in the Soundi device, namely MAXM86161 (Maxim Integrated Products, Inc., CA, USA), the following

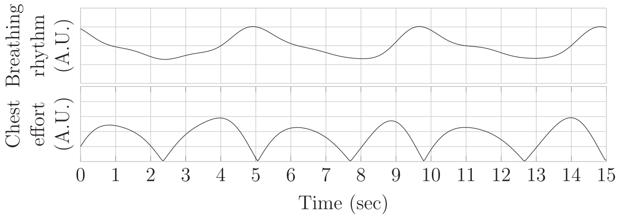

Fig. 2. Example of computed breathing and chest-effort signals for a 15 s long record.

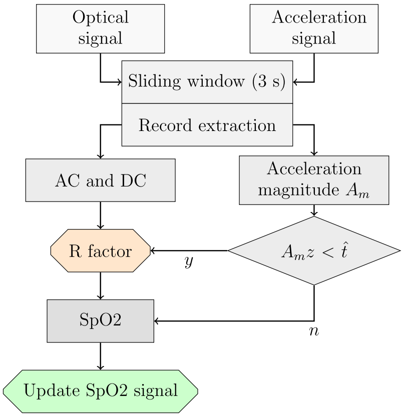

Fig. 3. Computational pipeline to reconstruct the SpO2 signal.

calibration model maps R into the $SpO2$ as:

$$SpO2 = -16.6R^2 + 8.33R + 100. \tag{1}$$

In principle, the R factor can be calculated over the cardiac cycle or better averaged over a series of pulses to increase the reliability of the measure. In wearable setups, due to motion artifacts that distort the signal waveform, this approach may be inadequate. In this work, the acceleration signal was used to monitor sudden subject movements to increase accuracy and ameliorate resiliency against distortions of the photoplethysmographic signal. To the aim, a time window of 3 s, sliding sample per sample, was adopted to compute the DC and AC components of both RED and IR optical responses. Both components were obtained as an average over the time window, according to the number of included cardiac cycles. Whether the acceleration value was smaller than an empirically predefined threshold $\hat{t}$ then a new value of R factor was computed, retained and used to compute the SpO2 value. Otherwise, the earlier value of SpO2 was recovered to update the signal (Fig. 3). Finally, the signal underwent standardization based on the Z-score transform.

3) Body Position Signal: Information about body position was reconstructed as a step-wise continuous signal, each level representing either a body position or a transition between two positions. Assuming the repeatability of the sensor placement on the chest, the tri-axial acceleration signal was used to decode the body position. Six different steady positions, namely stand, supine with the tilted trunk, supine, lateral right, prone and lateral left, were considered fully representative of the body position. The dynamic transition between two positions, or a subject movement in general, was considered the seventh condition. Two stages were put in place: 1) training of a decision-tree classifier, able to discriminate between the most likely sleeping position and a position transition by processing a short record of the acceleration signal; 2) creating the step-wise continuous time series by linear interpolation of the classified records. In order to train the classifier, ten subjects were recorded for 25 minutes, with a protocol involving the change among the six positions approximately every 2 minutes sequentially, back and forth. Transitions between two following positions lasted less than 10 s. The accelerometer sequence was split into 15 s long records and the DC components of the three accelerations were obtained by applying a low-pass 4th order Butterworth filter with a cut-off frequency of 0.05 Hz. For each record, the three median values of DC components were computed. In order to detect transitions between two positions, the acceleration derivatives were taken into account and summarized into a single jerk factor J_a, calculated as:

$$J_a = \sum_{i=1}^{3} mean(|\nabla(a_{\mathrm{i}})|) \tag{2}$$

so that four parameters were used to discriminate among the seven conditions (six positions and motion). Considering the 10 subjects and some 100 records, 15 s long each, about 1000 records were utilized overall to train and test the classifier. In the second stage, the accelerometer sequences of the overnight acquisitions underwent 15-s long record subdivision, DC component filtering, parameter computation, position classification and body position signal reconstruction by joining step-wise records (Fig. 4). The final position signal was attained by remapping the discrete steps (0.7) into the range -1 and $+1$.

4) Acoustic Signal: The acoustic signal was computed by properly processing the signal recorded by the differential air-pressure sensor. This served as a surrogate for snoring episodes, expected to be scattered over the sequence of normal breathing patterns. First, the air-pressure signal was filtered using a 4th order high-band Butterworth filter, featuring a cut-off frequency of 60 Hz. Then, the signal was standardized, and the envelope's absolute value was taken using the Hilbert transform. Savitzky–Golay filtering (4th order) allowed to smooth the resulting signal.

D. Signal Labeling and Class Priority

Labeling the five pre-processed signals was mandatory to create a training dataset used to learn weights in the deep network. In order to enable the network to identify heterogeneous conditions in sleep (e.g., regular and irregular breathing, temporary cessation of breathing, snoring patterns), the identification was reframed into a triple-set classification problem as introduced above. Summarizing, the first class set was intended to identify useless signal records, considered irrelevant from a diagnostic point of view because affected by noise and motion artifacts. For this class set, breathing, body position and acoustic noise signals were taken into account. The second set discriminated apnea events and abnormal/irregular respiratory/oxygenation patterns, from normal breathing. For this class set, breathing, chest-wall motion and SpO2 signals were considered. For clarity, obstructive, central and mixed sleep apnea patterns, the three main conditions for temporary cessation of breathing, were represented in the label "Apnea". The third label was designated to represent respiration and oxygenation events featuring doubtful nature, which could be of interest from a clinical point of view. Non-periodic breathing patterns, sudden decrease in SpO2, repetitive desaturation periods along with sinusoidal trend present in breathing and chest-effort signals,

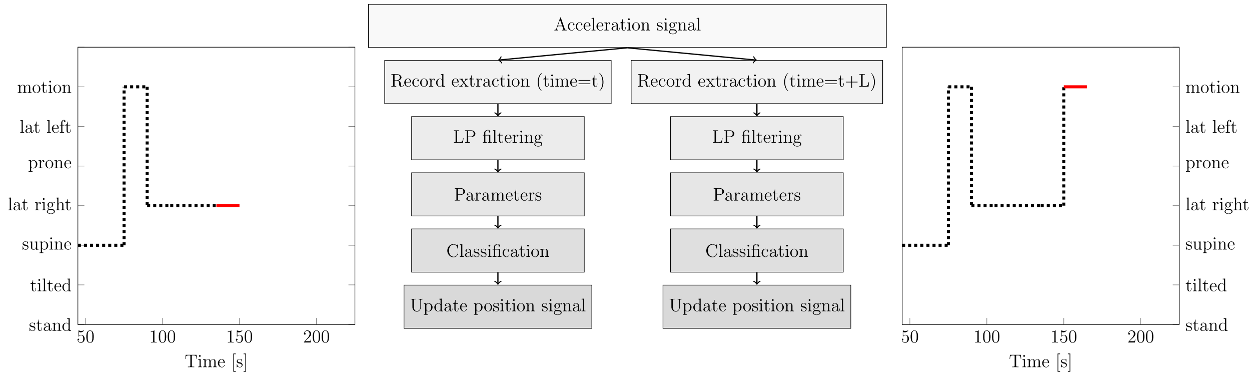

Fig. 4. Computational pipeline to reconstruct the body position signal: example of the analysis of two following records of length 15 s. In the example, in the first period t, the subject was lying down on the right side, while in the next period the subject had moved.

TABLE I
Class Priority Rule. L is the Time Length of the Record

Signal Quality	*Corrupted*	*Normal*		Priority
Corrupted	$\geq 30\% \cdot$ L	-		1^{st}
Normal	-	-		2^{nd}

Breathing Pattern	*Apnea*	*Abnormal*	*Normal*	Priority
Apnea	$\geq 25\% \cdot$ L	-	-	1^{st}
Abnormal	-	$\geq 50\% \cdot$ L	-	2^{nd}
Normal	-	-	-	3^{th}

Sleep Pattern	*Noise*	*Snoring*	*Normal*	Priority
Snoring	-	$\geq 50\% \cdot$ L	-	1^{st}
Irregular	$\geq 30\% \cdot$ L	-	-	2^{nd}
Normal	-	-	-	3^{th}

short periods in which the amplitude of the chest-wall motion decreased following desaturation, and sudden decrease in RR interval, the time elapsed between two successive R-waves of the QRS signal on the electrocardiogram, were acknowledged in the label. Especially identified repetitive desaturation was reported to match the characteristics of the Cheyne-Stokes respiration event [37], [38], featuring decrease and increase of both SpO2 and respiratory effort, but even swinging alteration of the RR interval. The third class set focused specifically on sleep quality exploiting the acoustic noise signal to differentiate snoring events and irregular noise (e.g., groan, teeth grinding) from normal sleep. We acknowledge that snoring is traditionally interpreted as a breathing disorder. Nonetheless, snoring can also be regarded as a respiratory sound-related pattern occurring during sleep. The labeling task was performed by two operators, supervised by an expert clinician in sleep medicine, exploiting semi-automatic signal processing tools available in a custom interactive dashboard, developed using the python-based suite called Voilá. In the dashboard, the operators visualized the five signals, selected the time window of interest (record), identified events and classified the signal records according to the three described sets. First of all, the labeling involved the qualitative interpretation of the overall signal quality (first class set) detecting the presence of artifacts that may disrupt the signal waveforms, corresponding to either body position changes or mechanical interference on the chest-worn device. In this class set, the priority was given the low-quality signal labeled as "corrupted" (Table I). Then, breathing patterns were recognized. The event interpretation was first helped by recognizing potential apnea computing the variability of the breathing signal. Whether this value was less than a predefined threshold B_s, then the record was set as a potential candidate for apnea label (first class set), confirmed after expert revision. Likewise, potential snoring events were detected by the automatic identification of acoustic patterns exploiting the energy of the signal record. Whether the energy value exceeded a predefined threshold A_s, then the record was set as a potential candidate for snoring label (third class set), confirmed again by expert revision. Finally, the labeling underwent careful check to remove inconsistent records severely affected by noise, which leads to retaining about eight and a half hours of labeled records, on average across all 20 acquisitions. As a result, some 41000 15 s long records were retained. Heuristic rules for priority assignment were developed to address the labeling in breathing and sleep pattern class sets. In the class set "Breathing Pattern", 25% of continuous apnea across the record duration L was considered enough to set the corresponding label. Otherwise, if at least 50% of the overall record length L was considered "irregular", the label was set to "abnormal". In the third class set "Sleep Pattern", the snore label (”snoring") was set whether at least 50% of the record length of the acoustic noise signal was identified with a snoring pattern. Otherwise, the "noise" label was set whether at least 30% of the acoustic signal length was noised.

E. Deep Network and Explainability

1) Architecture:

The developed network, implementing multiple classification tasks, performed multi-modal and multi-scale processing. Multi-modality was ensured by setting the input as the bundle of the 5 signals. Without losing generality, the signals were down-sampled from the original 400 samples per second to 60 samples per second, corresponding to the maximum cut-off frequency in the acoustic signal. Considering this new sampling frequency and the duration L of the record equal to 15 s, the input of the network was 900 × 5 samples (Fig. 5). Multi-scale was achieved by designing three parallel paths, in the encoder branch, each composed of three convolutional (activation function: Relu) and max-pooling layers in cascade and featuring filter sizes of 3, 10, and 25 samples, respectively. In the next layer, the feature maps coming from each of the three paths were stacked on the depth dimension. To further reduce the number of feature maps, an additional convolutional layer (bottle-neck), with a one-dimensional kernel, was added. The output of the bottle-neck layer was then flattened to a

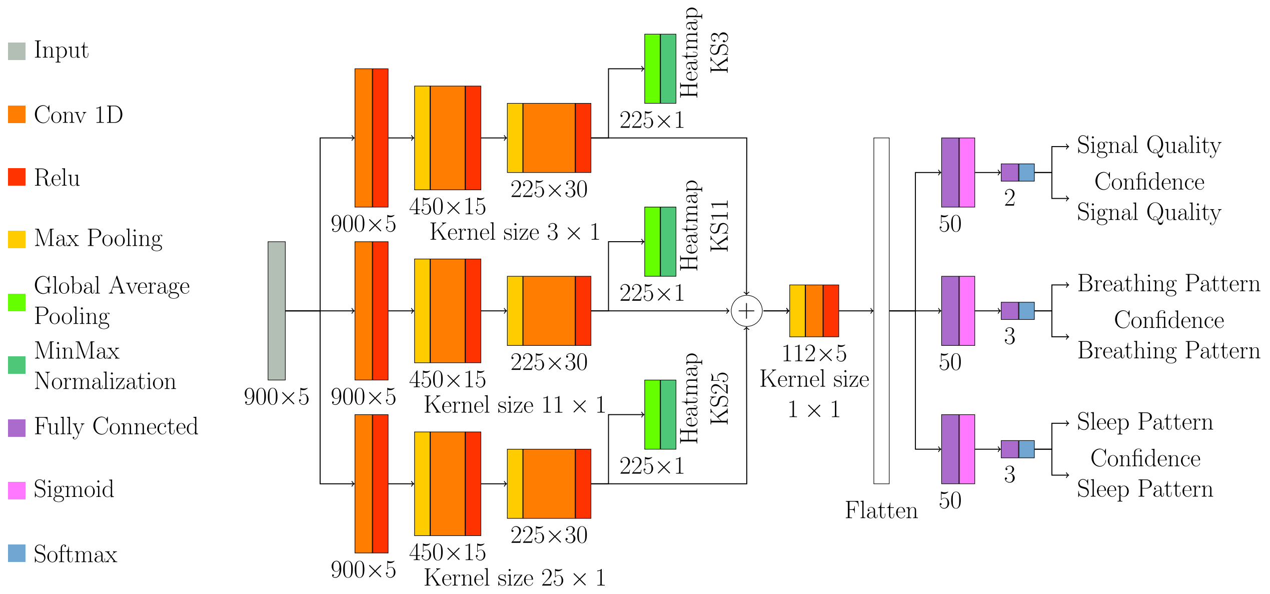

Fig. 5. SLEEP-SEE-THROUGH (SST) neural architecture used both for classification and result explanation. Qualitative explainability is represented by the heatmaps KS3, KS11 and KS25. Quantitative explainability is represented by the confidence indices of the three predicted class sets.

mono-dimensional array of features that entered in parallel three fully connected (FC) layers (sigmoidal neurons). The output of each FC layer was the input to the specific Softmax layer designed to predict one of the three class sets. Three losses, based on categorical cross-entropy, were used as cost functions. To take into account the disparity of each of the classes in each class set, the entropy was opportunely weighted. In order to enable the explainability (XAI) of the network, two strategies, one quantitative and one qualitative, were implemented, the first one based on a confidence index (CI) of the predictions and the second one based on visual heatmaps.

2) Confidence Index: The first XAI's strategy involved the computation of the confidence of the prediction quality. To take into account the different number of classes in each class set j, a new index was proposed:

$$C_j = \frac{N-1}{N} \cdot \left(\max\{p_{\mathrm{i}} \text{ with } i \ \in [0; N_j]\} - \frac{1}{N_j} \right) \qquad (3)$$

where N_j is the number of classes of the class set j, and p_i is the probability output of the i-th class. C_j ranges from 1 to 0, being the best and the worst confidence value, respectively. This computation was made explicit into the network by cascading a 1D max-pooling layer, computing the highest predicted probability in the class set, to the output of the Softmax layer (Fig. 5) and a custom layer to compute (3).

3) Heatmaps: Heats were built from the output of the last convolutional layer in each path of the encoder branch. First, all the activation maps (e.the g. $n = 30$) arising from one path were averaged (global-average-pooling layer) along the depth dimension to remove the dependency on the feature maps. Then, a *MinMax* standardization, implemented by a custom layer, was applied to re-scale the amplitude into the range 0,1. The resulting data vector was graphically plotted as a 2D map, stacking the vector n times, with n equal to one-quarter of the vector size. The final visualization was composed of three heatmaps, the output of the three encoder paths, namely KS3, KS11 and KS25, according to the corresponding kernel size of the convolution (cfr Fig. 5). The heatmaps displayed the occurring events, with the corresponding colormap continuously ranging from dark blue (zero value) to yellow (unitary value).

III. Experimental Tests and Results

A. Class Balancing Strategies

Following the annotation phase, occurrences in the three class sets were unbalanced in the labeled dataset. Particularly, events of interest, especially apneas, were underrepresented. In the signal quality class set, the ratio between corrupted and normal records was about 4%. In the breathing pattern set, the ratios of apnea and abnormal events with respect to normal breathing were about 1% and 16%, respectively. In the sleep pattern set, the ratios of snoring and irregular events with respect to normal sleep were about 25% and 5%, respectively. In order to favor class balancing, and reduce potential overfitting in the network training, down-sampling and data augmentation strategies were put in place. First, the records in which all three class sets were labeled as normal were identified. Half of such records were randomly chosen and dropped from the dataset. Then, records labeled as apnea class in the sleep class set were up-sampled by a factor of about ten. One new record was obtained by scaling breathing rhythm and chest effort signals using a random amplitude selected in the range ±0.01 and ±0.05, respectively. The overall procedure increased the ratio between apnea and normal breathing up to 15%, and the ratio between irregular and normal sleep events up to 10% (Table II).

In order to further address unbalance in the class sets, a weighted version of the categorical cross entropy L was adopted for each of the three class sets as:

$$L_c = -\sum_{i=1}^{n} \left(y_{i,c} \cdot \log \hat{y}_{i,c} \cdot w_{i,c} \right) \qquad (4)$$

where L_c, $y_{i,c}$, $\hat{y}_{i,c}$, n and $w_{i,c}$ were the loss function of the class set c, the i-th scalar value in the model output, the corresponding target value, the number of scalar values in the model output,

TABLE II
Amount of Events Across the Three Class Sets Before and After Data Augmentation

	Signal quality		Breathing			Sleep		
	Corrupted	Normal	Apnea	Abnormal	Normal	Snoring	Irregular	Normal
Original	1630	39170	350	5580	34870	7845	1575	31380
Data augmentation	2330	27970	3020	7100	20180	7200	2100	21000

and the corresponding weight, respectively. The weight $w_{i,c}$ was computed as:

$$w_{i,c} = \frac{k_{i,c}}{K} \tag{5}$$

where $k_{i,c}$ and K were the number of events i (e.g., records labeled as apnea) for class set c (in this case Sleep class set) and the overall number of records, respectively.

B. Testing Strategies

The labeled dataset was used in three different testing strategies, namely network architecture ablation, sensor-fusion advantage and subject-wise cross-validation. In the ablation test, data were processed in a record-wise mode so that all 20 subjects were gathered into a unique dataset of records. The 10% of the records were randomly extracted to serve as the test set. The remaining records were further split into training (70%) and validation (30%) subsets. Accuracy, recall, precision and F1-score, computed across the test set, quantified record-wise prediction performance. In detail, the ablation analysis was developed to assess the dependability of the classification results upon the architecture of the encoding branch of the network and the time length of the signal record. In the first trial, the record length was set arbitrarily to 15 s and four variants of the encoder path were envisioned, namely: 1) single path with kernel size 3×1; 2) single path with kernel size 11×1; 3) single path with kernel size 25×1; 4) triple path (cfr. Fig. 5). The Statistical difference among the four setups was computed by Kruskal-Wallis non-parametric test and posthoc comparison (p-value=0.05) of the distribution of the classification errors $\{E_c\}$, accumulated across the three class sets. For each 15 s long record (225 samples × 5 signals), E_c was computed as:

$$E_c = \sqrt{\sum_{i=1}^{2} (e_{c_1,i})^2 + \sum_{i=1}^{3} (e_{c_2,i})^2 + \sum_{i=1}^{3} (e_{c_3,i})^2} \tag{6}$$

where k ranges the three class sets (Signal quality, Breathing Pattern and Sleep Pattern), i ranges the elements of the corresponding class set and $e_{c_k,i}$ is the difference between the predicted probability $y_{c_k,i}$ and the corresponding nominal probability $\hat{y}_{c_k,i}$, respectively. In the second trial (dependability on the record length), the best network architecture was selected according to the previous test and three different record lengths, namely 15, 30 and 45 s were taken into account and compared. In the second testing strategy, the role of sensor-fusion was explored, by comparing the prediction quality obtained using all five input signals against single and pair signal combinations. In the third testing strategy, leave-one-out validation (LOOV) was performed by carrying out 20 training sessions. In each training, the dataset, composed of 19 subjects, was split into training validation, this last used to stop the convergence according to a patience criterion, namely 10 epochs. The test was performed on the subject excluded from the actual training set. In the LOOV, accuracy, recall, precision and F1-score, computed across the 20 training sessions, independently quantified the prediction performance.

TABLE III
Ablation Test: Encoder Architecture. Label p1, p2, p3 and All Correspond to Single Path (Ks: 3×1), Single Path (Ks: 11×1), Single Path (Ks: 25×1), Triple-Parallel Path, Respectively

	Signal			Breathing			Sleep		
	Rec	Prec	F1	Rec	Prec	F1	Rec	Prec	F1
p1	0.88	0.91	0.89	0.79	0.88	0.82	0.60	0.70	0.63
p2	0.87	0.91	0.89	0.85	0.87	0.86	0.61	0.72	0.64
p3	0.87	0.92	0.89	0.89	0.89	0.89	0.63	0.74	0.67
all	0.91	0.92	0.92	0.84	0.89	0.87	0.70	0.74	0.71

C. Results of Record-Wise Ablation

The ablation analysis was devoted to assessing the dependability of the classification results upon the architecture of the encoding branch of the network and the length of the signal record. In this analysis, all the statistical data were used in bundle, exploiting a record-wise approach. In the first test, (record length: 15 s) four variants of the encoder path were considered, namely: 1) single-path encoder with kernel size 3×1; 2) single-path encoder kernel size 11×1; 3) single-path encoder kernel size 25×1; 4) triple-path encoder (cfr. Fig. 5). The overall accuracy values for the three class sets, averaged across the four architectures, were 0.97, 0.92, 0.82, respectively, highlighting the worst results for the sleep pattern class set. The architecture featuring the triple-path encoder was statistically compared to the previous three ones by evaluating the error distributions, computed as the root mean squared error between the predicted and nominal values. The difference was significant in all three cases ($p < 0.001$, $p < 0.01$, $p < 0.03$). The analysis of the recall, precision and F1-score metrics confirmed that the best results were provided when using the encoder architecture with the three paths in parallel (Table III). According to the reported results, the second test considered three different record lengths, namely 15, 30 and 45 s, applied to the triple-path encoder architecture. Accuracy results were 0.97, 0.93 and 0.82 on average across the three setups. The statistical comparison among the three error distributions, computed as done before, provided no statistical differences ($p = 0.18$, $p = 0.82$, $p = 0.14$), thus demonstrating that the record length was not a critical feature for the analysis.

D. Results of Sensor-Fusion Test

The network featuring the three parallel paths in the encoder was taken into account, and its input was tailored in agreement with the used signals. The bundle of all the five signals was compared against breathing rhythm alone, chest effort alone, acoustic signal alone, and bundle of breathing rhythm and chest

TABLE IV
SENSOR-FUSION TEST. A) BREATHING RHYTHM; B) CHEST EFFORT; C) ACOUSTIC; D) BREATHING + EFFORT; E) COMPLETE 5-SIGNAL BUNDLE

	Signal			Breathing			Sleep		
	Rec	Prec	F1	Rec	Prec	F1	Rec	Prec	F1
A	0.78	0.87	0.82	0.76	0.79	0.78	0.56	0.67	0.59
B	0.86	0.91	0.88	0.69	0.76	0.72	0.53	0.66	0.56
C	0.82	0.90	0.85	0.81	0.86	0.83	0.57	0.69	0.61
D	0.87	0.90	0.89	0.83	0.83	0.83	0.61	0.69	0.64
E	0.91	0.92	0.92	0.84	0.89	0.87	0.70	0.74	0.71

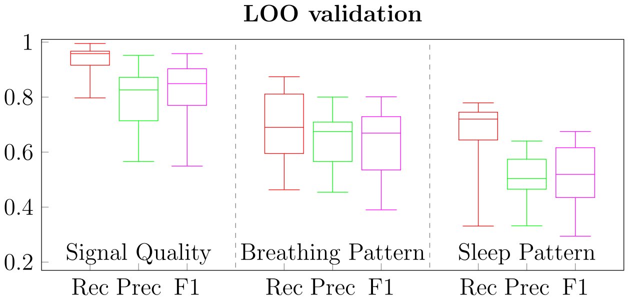

Fig. 6. Subject-wise analysis. Results (recall, precision and F1-score) of the LOO validation across the three class sets.

effort (Table IV). Breathing rhythm (A) and chest effort (B) alone were both inferior in all the three class sets with respect to full bundle setup (E) of about 14, 10, 20% and 5, 18, 25% (recall), respectively. Likewise, the acoustic signal alone (C) provided less accurate predictions than that of the full bundle. Despite joining breathing rhythm and chest effort (D) slightly increased prediction results of breathing patterns with respect to single-signal input, the prediction of sleep patterns did not improve though. Such differences were confirmed by the statistical analysis of the error distributions between the predicted and nominal values, which provided p values (E against A, B, C and D) smaller than 0.001. Overall, such results confirmed that multifactorial classification may take advantage of processing multiple signal in bundle.

E. Results of Subject-Wise Validation

For subject-wise validation throughout LOOV, the same network of the previous test was considered. Overall accuracy results demonstrated meaningful ability of the network in the rating of the signal quality (0.96) (Table V). Remarkable accuracy results were achieved in the breathing class set as well, namely 0.92, 0.96 and 0.88 in the detection of normal pattern, apnea event and irregular pattern, respectively. The prediction quality of the three sleep patterns featured more uncertainty, especially in the distinction between snoring (0.76) and noise (0.61) patterns. The analysis of recall, precision and F1 score across the three class sets (Fig. 6) supported the preliminary analysis of the overall accuracy. As a result, the recall was greater than the precision in all three class sets. The results for the specific breathing pattern highlighted high quality detection of normal breathing patterns, namely recall: 0.86, precision: 0.97 and F1-score: 0.89 (Fig. 7). Slightly less accurate results were found for the abnormality pattern, featuring 0.84, 0.75 and 0.78 for recall, precision and F1-score, respectively. The detection of apnea events was the most critical. While the recall was 0.71, the precision was small in the range of 0.4. As far as sleep pattern is concerned, the recall of snoring and noise records was 0.65 and 0.93, respectively. However, the uncertainty of the precision was higher, with both values less than 0.5. Cohen's kappa coefficients of agreement, between the manual and automatic prediction of the three breathing patterns, were 0.76 (normal), 0.64 (abnormal) and 0.42 (apnea), featuring a substantial agreement for the first two patterns, while the apnea pattern was in moderate agreement only [39]. As far as sleeping patterns were concerned, Kappa coefficients were 0.23 (weak agreement), 0.25 (weak agreement) and 0.90 (very high agreement) for normal, snoring and noise patterns, respectively.

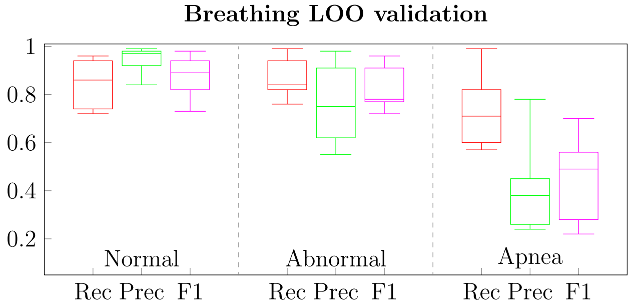

Fig. 7. Subject-wise analysis. Results (recall, precision and F1-score) of the breathing pattern detection.

F. Snoring and Apnea Episodes: Relation With Body Position and Comparison With Competitive Approaches in the Literature

The relation of apnea and snoring episodes with the body position was analyzed by computing the time percentage spent in the different body positions (Table VI). As expected, being the population substantially nonapneic, a higher positional dependency of the snoring with the supine position than lateral positions was found, this in agreement with [33]. Likewise, rare apnea episodes occurred in prone and lateral positions, being much higher than the occurrence of the supine position. As far as apnea episode detection is concerned, the obtained accuracy (cfr. Table V) was compared with representative contributions in the literature (Table VII) reporting the dataset numerosity, the AI model, and the classification type. Regarding the model, they spanned support vector machine (SVM), CNN and traditional and bidirectional LSTM. The classification difference was in terms of single-class (e.g., apnea against no apnea) and multi-class (e.g., apnea, obstructive apnea, no apnea). While attained on a different dataset and using a specific AI model, results were substantially competitive with those ones of the other works. Considering that we performed a multi-set classification, the only similar paper reported a substantially lower accuracy (0.70).

G. Quantitative and Qualitative Explainability

The confidence indices (cfr. (3)) were consistent in the quantification of the prediction quality (Table VIII). Without losing generality, for the record in the first row, both signal quality and breathing pattern were predicted normal, featuring network output values of 0.90 and 0.80, respectively. Conversely, the prediction of the sleep pattern was uncertain between normal (0.50) and snoring (0.4). The corresponding confidence indices

TABLE V
LOO Analysis: Event Count and Classification Accuracy of the Three Class Sets. Median Values and Interquartile Range (IRQ) Computed Across the Subject Population

	Signal quality		**Breathing pattern**			**Sleep pattern**		
	Normal	Corrupted	Normal	Apnea	Irregular	Normal	Snoring	Noise
count	1417.5(461.5)	104.5(95.5)	1122(498.5)	72(126)	166.5(465)	1222(333.75)	191.5(274.5	59(45.25)
accuracy	0.96(0.03)	0.96(0.03)	0.92(0.05)	0.97(0.05)	0.88(0.07)	0.96(0.04)	0.73(0.1)	0.61(0.1)

TABLE VI
Relation of Apnea and Snoring Episodes With the Body Position

Event	motion	lat left	prone	lat right	supine	tilted
Apnea	28%	10%	1%	12%	45%	4%
Snoring	10%	12%	3%	15%	54%	6%

TABLE VII
Apnea Detection Accuracy: Comparison With Relevant Literature Results. CNN: Convolutional Neural Network, DNN: Deep Neural Network, GRU: Gated Recurrent Unit, k-NN: K-Nearest Neighbors Algorithm, LDA: Linear Discriminant Analysis, LSTM: Long-Short Term Memory, BiLSTM: Bidirectional Long-Short Term Memory, SVM: Support Vector Machine, SST: SLEEP-SEE-TRHOUGH - This Work

Study	Cohort	Model	Classification	Accuracy	Sen/Spec
[40]	28	SVM	single-class	0.92	0.87/0.95
[41]	2100	LSTM	single-class	0.77	0.62/0.80
[42]	86	DNN	single-class	0.93	0.93/0.94
[42]	86	GRU	single-class	0.99	0.99/0.99
[28]	20	CNN	single-class	0.84	0.81/0.87
[43]	70	LDA	single-class	0.82	0.82/0.90
[29]	100	CNN	single-class	0.95	0.97/0.79
[29]	100	CNN	multi-class	0.55	0.50/0.52
[44]	10	k-NN	single-class	0.82	0.78/0.75
[30]	17	BiLSTM	single-class	0.86	0.90/0.83
[35]	103	CNN	single-class	0.80	0.78/0.93
SST	20	CNN	multi-class	0.68	0.71/0.4

TABLE VIII
Predicted Output Results and Their Relative Confidence Indexes (CI) in Four Different Predictions

Prediction			CI
Signal	Breathing	Sleep	
[0.90, 0.10]	[0.80, 0.17, 0.03]	[0.50, 0.40, 0.10]	[0.80, 0.70, 0.25]
[0.51, 0.49]	[0.96, 0.02, 0.02]	[0.60, 0.40, 0.00]	[0.02, 1.0, 0.40]
[0.99, 0.10]	[0.20, 0.20, 0.60]	[0.33, 0.27, 0.4]	[1.00, 0.40, 0.10]
[0.21, 0.79]	[0.70, 0.20, 0.10]	[0.6, 0.19, 0.21]	0.58, 0.55, 0.40]

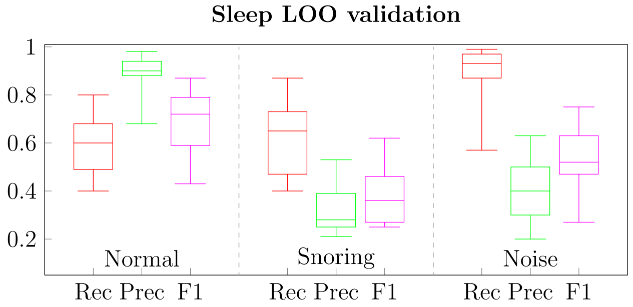

Fig. 8. Subject-wise analysis. Results (recall, precision and F1-score) of the sleep pattern detection.

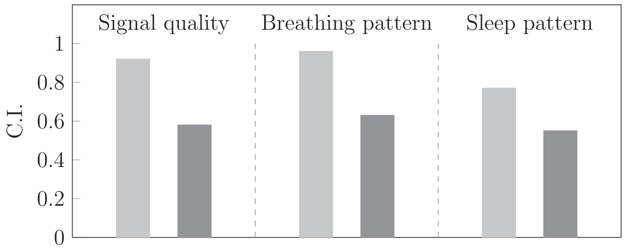

Fig. 9. Bar chart of the confidence indexes (CI) for each of the three class sets, subdivided between correct (light gray bar) and wrong (dark gray bar) classifications.

synthesized that the first two predictions were reliable (0.80 and 0.70) while the third one was unreliable (0.25), proving the effectiveness of such measure. In the 2nd row of the table, the uncertainty in the signal quality was very high testified by a corresponding CI of 0.02. The distribution of the confidence indices (Fig. 9), computed over the test dataset (record-wise), subdivided between misclassified and correctly classified samples confirmed less confidence (about 0.50) as soon as wrong prediction occurred, across all the three class sets. In case of correct predictions, the confidence approaches 1 in the first two class sets. For sleep pattern, the confidence was lower close to 0.75. For sake of clarity, qualitative XAI results were reported only in three paradigmatic cases, namely normal breath and sleep condition (Fig. 10), snoring event (Fig. 11) and apnea episode (Fig. 12). In the first case, the 15 s long record was correctly classified as normal both in breathing and sleep patterns (Fig. 10). The second case (Fig. 11) represents one different 15 s long record that shows a correctly classified snoring event, while the third one (Fig. 12) depicts an apnea episode. As far as the first case is concerned, the vibration signal (high-frequency acoustic signal) was practically flat and the three heatmaps showed periodic patterns related to the breathing rhythm and chest effort signals. Interestingly, the three maps were not synchronous, with patterns in KS3 and KS11 mainly in phase with the breathing rhythm and chest effort signals, respectively. Likewise, K25 map showed a periodic pattern although less appreciable. Differently, the heatmaps of snoring episodes (Fig. 11) echoed mostly the patterns in the acoustic signal, especially in KS3 and KS11 maps, being the pattern in the KS25 map still coherent but more fading. It can be argued that the amplitude of snoring episode signals was that high to erase the effects of breathing and chest waveforms. Consistently, the snoring episode was pretty much aligned with chest effort signal. In the case of apnea (Fig. 12), the duration of the breathless period (approximately the first 8 s of the record) was represented in the maps as a flattened pattern. Interestingly, while KS3 and KS25 maps featured a low-value pattern (blue), KS11 map showcased a high-value pattern, this probably related to the variation, although slow, in the oxygenation signal. As soon as the subject breathed again, the heatmap showed patterns synchronous with the respiration peaks.

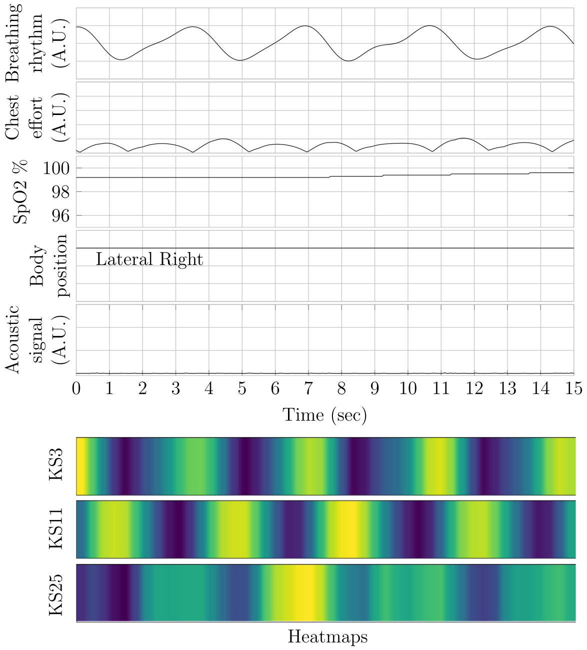

Fig. 10. 15 s record with the corresponding heatmap (record correctly classified as normal both in breathing and sleep patterns).

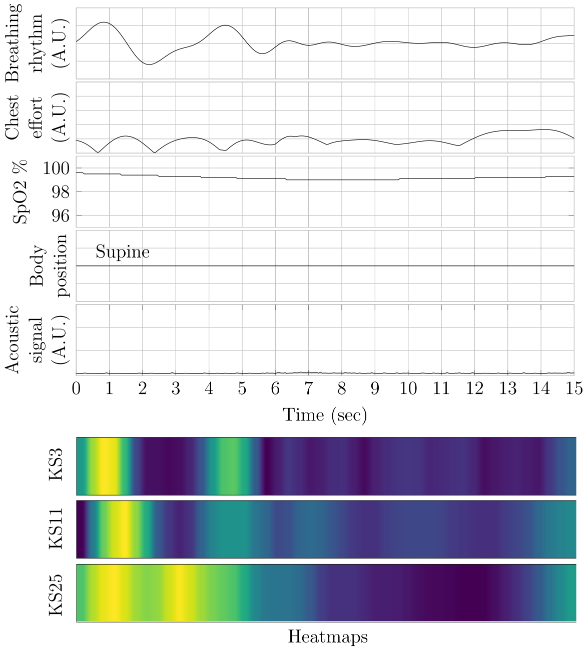

Fig. 12. 15 s record with the corresponding heatmap (record classified as apnea in the breathing pattern and normal in the sleep pattern).

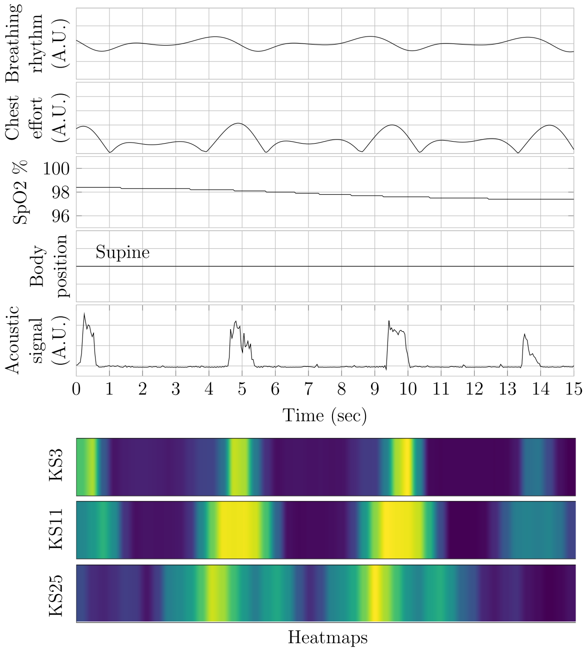

Fig. 11. 15 s record with the corresponding heatmap (record correctly classified as snoring in the sleep pattern).

IV. Discussion

Occurring debate in the sleep science community is about tools and methods supporting the transition from traditional lab-based PSG to easier and home-based respiratory polygraphy [45], [46]. Nonetheless, its effectiveness has yet to be assessed across a broad clinical spectrum [9], [27]. Wearables, largely advocated for their intrinsic invasiveness, may record different multi-modal signals, but the knowledge of what is the ideal combination to mimic the entirety of PSG remains elusive [16], [47], [48]. Likewise, deep learning models have been extensively investigated for their ability to automate event detection in biosignals [29], [32], however the sleep pattern is intrinsically complex and multifactorial, and agreement on architectures is still lacking [35], [38], [49], [50]. Saying that, in this work, a fully-integrated chest-worn device blended seamlessly with a novel deep neural network, which performed three multiple classifications, to detect the overall signal quality, recognize breathing patterns, and identify alterations of breathing, featuring clinical relevance, during sleep. Despite traditional PSG, the proposed solution made use of optical, acoustic (by using a differential air pressure sensor), and accelerometer signals. Unlike most works in the literature that used only respiratory data, this study exploited sensor-fusion. By pre-processing the recorded signals, breathing, chest effort, oxygenation, subject position and sound data were obtained, constituting the network input. Especially, for breathing rhythm assessment, despite traditional PPG-based measurements [15], we exploited the bundle of surrogate acoustic and accelerometer signals. As an additional and unprecedented technical contribution, the prediction confidence index and heatmaps were readily integrated into the network to increase the explainability of the results. The network showcased an excellent ability to label the goodness of the signals. Results on breathing pattern classification were globally promising, being the accuracy of apnea detection comparable to the results previously reported in the literature (cfr. Table VII). Sleep pattern classification was less effective than that one of the breathing pattern, likely due to the intrinsic difficulty in distinguishing snoring and noise events (cfr. Fig. 8). Actually, while the recall of the two events was favorable, the precision was actually sensibly poorer, arguing that disturbing sounds may be detrimental to the accurate breathing alteration detection. Nonetheless, the snoring waveform in the acoustic signal features a typical pattern (cfr. Fig. 11), and the network was generally effective in detecting it.

The analysis of the advantage of sensor-fusion confirmed that using all the five signals in bundle provided the best results (cfr. Table IV). From an architectural point of view, two innovative technical contributions provided the confidence index and three maps of saliency. An algorithmic module, in parallel with the Softmax output, computed the CI whose role was effective in elucidating the prediction quality. Three parallel convolutional paths were implemented in the encoder branch, featuring intermediate network outputs corresponding to graphical descriptive patterns, shaped as 2D heatmaps. They were showcased to be linked to signal content, especially in case of events of interest such as snoring and apnea (cfr. Figs. 11 and 12). We remark that joining these two pieces of information into a verification tool can be regarded as a step forward in making deep networks explainable. As far as sensor-fusion is concerned, the position signal revealed interesting correlations with both apnea and snoring events, occurring with higher frequency when the subjects were supine. Likewise, the oxygenation signal revealed fundamental to discriminate the apneas from irregular breathing.

Despite much literature, extensive comparison with other documents was essentially impractical due to the specificity of the acquisition device and the signal processing developed. However, concerning the assessment of breathing patterns, the obtained results were quite in agreement with recent findings [29]. The author proposed the use of accelerometers to achieve identification of normal breathing, central sleep apnea and obstructive sleep apnea 92%, 87% and 70% (F1-score), respectively. LSTM networks were proposed to detect sleep apnea from potentially different sources such as respiratory bands or ECG [41]. Similar to our findings, the paper reported precision results in the range of 40%, recognizing the difficulty in discriminating respiratory disturbances from apnea. In addition, the networks made decisions on a per-second basis introducing further false positives on a shorter time basis, requiring a lot of post-processing. In [35], the authors exploited sounds recorded by the smartphone at home and reported detection accuracy of moderate and severe OSA in the range of 0.80 and 0.85, respectively, again in agreement with our results.

Some shortcomings of the present work deserve attention. The subjects were monitored overnight for 10 hrs, and following the pre-processing approximately 8 and half hours of useful data were attained on average. Despite an overall availability of about 170 hrs of recording, the present work lacked to deal with pathological sleep conditions. Therefore, the attained performances cannot be extended to a general population of patients affected by sleep disorders. In addition, data analysis from the signal acquisition in a domestic environment is to be tested against portable PSG protocol. Nonetheless, a lot of effort was spent in labeling data according to the three defined class sets, enabling the focus not only on breathing patterns but also on the overall signal quality and sleep events such as snoring and general acoustic noise. As far as the apnea event is concerned, the acquisition protocol was unfeasible in describing its wide phenomenology. This resulted in grouping all the different types under a general apnea event. However, along with regular breathing, we enabled an additional class for including abnormal breathing, which could be used to label varying respiratory rhythms. We may argue however that some hypopnea events, featuring short duration, were not detected as apnea and were probably treated as either abnormal or normal respiration. Generally speaking, the developed prediction framework was conceived to be potentially scalable to enclose many several states so that the detection of different apnea typologies could be easily enabled. As far as the imbalanced dataset is concerned, apnea events were systematically underrepresented featuring a potential negative impact on the performance of the model. Nonetheless, a custom procedure was put in place to increase the balance of the label distribution. At the same time, the use of neural networks allowed us to tailor weighted loss functions [cfr. (4)]. The annotation procedure, while supervised by expert personnel, was affected by subjectivity in defining the occurrence and length of the events of interest in the record. This affected the class attribution and, as a consequence, the results. Nonetheless, the use of heuristic rules for class priority (cfr. Table I) allowed to make easier the task. Finally, we remark that, as a feasibility study, no clinical somnographic support was adopted to validate the methodology, which cannot be regarded therefore as an algorithm for staging/diagnosis. Rather, it can be used as a technical tool capable of separating normal from abnormal respiratory/sleep events. Prospectively, potential benefits towards clinical use can be synthesized in: a) fast screening of alterations of breathing during sleep; b) unobstructive monitoring leading to less interference with the sleep; c) less patient cooperation/adherence because of integration of sensors into a unique device.

V. Conclusion

In this paper, we contributed to support the recent perspective [28], [29], [31] that is feasible to use a reduced set of somnographic signals, coupled to deep neural networks, to automatically identify breathing alterations during sleep featuring potential clinical relevance. The study also demonstrated the relevance of equipping neural networks with intrinsic algorithms enabling the assessment of prediction quality. While results are still preliminary, we remark that this work may bring the use of AI-based tools for sleep disorder detection, with special attention to apnea screening tools, incrementally closer to clinical translation.

References

[1] J. Corral-Peñafiel, J.-L. Pepin, and F. Barbe, "Ambulatory monitoring in the diagnosis and management of obstructive sleep apnoea syndrome," *Eur. Respir. Rev.*, vol. 22, no. 129, pp. 312–324, 2013.

[2] R. Heinzer et al., "Prevalence of sleep-disordered breathing in the general population: The Hypnolaus study," *Lancet Respir. Med.*, vol. 3, pp. 310–318, Apr. 2015.

[3] S. B. Tufik, L. F. Berro, M. L. Andersen, and S. Tufik, "Prevalence and classification of sleep-disordered breathing," *Lancet Respir. Med.*, vol. 3, pp. 263–264, Apr. 2015.

[4] G. Medic, M. Wille, and M. E. Hemels, "Short- and long-term health consequences of sleep disruption," *Nature Sci. Sleep*, vol. 9, pp. 151–161, 2017.

[5] A. V. Benjafield et al., "Estimation of the global prevalence and burden of obstructive sleep apnoea: A literature-based analysis," *Lancet Respir. Med.*, vol. 7, pp. 687–698, Aug. 2019.

[6] A. Qaseem et al., "Management of obstructive sleep apnea in adults: A clinical practice guideline from the American College of Physicians," *Ann. Intern. Med.*, vol. 159, pp. 471–483, Oct. 2013.

[7] E. Anitua, J. Duran-Cantolla, G. Z. Almeida, and M. H. Alkhraisat, "Predicting the night-to-night variability in the severity of obstructive sleep apnea: The case of the standard error of measurement," *Sleep Sci.*, vol. 12, no. 2, pp. 72–78, 2019.

[8] H.-L. Tan, D. Gozal, H. M. Ramirez, H. P. R. Bandla, and L. Kheirandish-Gozal, "Overnight polysomnography versus respiratory polygraphy in the diagnosis of pediatric obstructive sleep apnea," *Sleep*, vol. 37, pp. 255–260, Feb. 2014.

[9] J. Corral et al., "Conventional polysomnography is not necessary for the management of most patients with suspected obstructive sleep apnea. Noninferiority, randomized controlled trial," *Amer. J. Respir. Crit. Care Med.*, vol. 196, pp. 1181–1190, Nov. 2017.

[10] F. Siddiqui, E. Osuna, A. S. Walters, and S. Chokroverty, “Sweat artifact and respiratory artifact occurring simultaneously in polysomnogram,” *Sleep Med.*, vol. 7, pp. 197–199, Mar. 2006.

[11] M. Younes, M. Younes, and E. Giannouli, “Accuracy of automatic polysomnography scoring using frontal electrodes,” *J. Clin. Sleep Med.: Official Pub. Amer. Acad. Sleep Med.*, vol. 12, pp. 735–746, May 2016.

[12] M. Hirshkowitz, “Polysomnography challenges,” *Sleep Med. Clin.*, vol. 11, pp. 403–411, Dec. 2016.

[13] A. Gramfort et al., “MEG and EEG data analysis with MNE-python,” *Front. Neurosci.*, vol. 7, Dec. 2013, Art. no. 267.

[14] E. Combrisson et al., “Sleep: An open-source python software for visualization, analysis, and staging of sleep data,” *Front. Neuroinform.*, vol. 11, 2017, Art. no. 60.

[15] P. H. Charlton et al., “Breathing rate estimation from the electrocardiogram and photoplethysmogram: A review,” *IEEE Rev. Biomed. Eng.*, vol. 11, pp. 2–20, 2018.

[16] R. Lazazzera et al., “Detection and classification of sleep apnea and hypopnea using PPG and SPO_2 signals,” *IEEE Trans. Biomed. Eng.*, vol. 68, no. 5, pp. 1496–1506, May 2021.

[17] D. Marzorati, D. Bovio, C. Salito, L. Mainardi, and P. Cerveri, “Chest wearable apparatus for cuffless continuous blood pressure measurements based on PPG and PCG signals,” *IEEE Access*, vol. 8, pp. 55424–55437, 2020.

[18] D. Marzorati, A. Dorizza, D. Bovio, C. Salito, L. Mainardi, and P. Cerveri, “Hybrid convolutional networks for end-to-end event detection in concurrent PPG and PCG signals affected by motion artifacts,” *IEEE Trans. Bio-Med. Eng.*, vol. 69, no. 8, pp. 2512–2523, Aug. 2022.

[19] A. Sadeh, “The role and validity of actigraphy in sleep medicine: An update,” *Sleep Med. Rev.*, vol. 15, pp. 259–267, Aug. 2011.

[20] B. Gonçalves, T. Adamowicz, F. M. Louzada, C. R. Moreno, and J. F. Araujo, “A fresh look at the use of nonparametric analysis in actimetry,” *Sleep Med. Rev.*, vol. 20, pp. 84–91, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1087079214000665

[21] J. A. Slater, T. Botsis, J. Walsh, S. King, L. M. Straker, and P. R. Eastwood, “Assessing sleep using hip and wrist actigraphy,” *Sleep Biol. Rhythms*, vol. 13, pp. 172–180, 2015.

[22] L. J. Meltzer, L. S. Hiruma, K. Avis, H. Montgomery-Downs, and J. Valentin, “Comparison of a commercial accelerometer with polysomnography and actigraphy in children and adolescents,” *Sleep*, vol. 38, pp. 1323–1330, Aug. 2015.

[23] J. Razjouyan, H. Lee, S. Parthasarathy, J. Mohler, A. Sharafkhaneh, and B. Najafi, “Improving sleep quality assessment using wearable sensors by including information from postural/sleep position changes and body acceleration: A comparison of chest-worn sensors, wrist actigraphy, and polysomnography,” *J. Clin. Sleep Med.: Official Pub. Amer. Acad. Sleep Med.*, vol. 13, pp. 1301–1310, Nov. 2017.

[24] M. Aktaruzzaman et al., “Performance comparison between wrist and chest actigraphy in combination with heart rate variability for sleep classification,” *Comput. Biol. Med.*, vol. 89, pp. 212–221, Oct. 2017.

[25] M. Bsoul, H. Minn, and L. Tamil, “Apnea MedAssist: Real-time sleep apnea monitor using single-lead ECG,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 3, pp. 416–427, May 2011.

[26] J. Jin and E. Sánchez-Sinencio, “A home sleep apnea screening device with time-domain signal processing and autonomous scoring capability,” *IEEE Trans. Biomed. Circuits Syst.*, vol. 9, no. 1, pp. 96–104, Feb. 2015.

[27] N. Devani, R. X. A. Pramono, S. A. Imtiaz, S. Bowyer, E. Rodriguez-Villegas, and S. Mandal, “Accuracy and usability of AcuPebble SA100 for automated diagnosis of obstructive sleep apnoea in the home environment setting: An evaluation study,” *BMJ Open*, vol. 11, Dec. 2021, Art. no. e046803.

[28] M. Hafezi et al., “Sleep apnea severity estimation from tracheal movements using a deep learning model,” *IEEE Access*, vol. 8, pp. 22641–22649, 2020.

[29] K. McClure, B. Erdreich, J. H. T. Bates, R. S. McGinnis, A. Masquelin, and S. Wshah, “Classification and detection of breathing patterns with wearable sensors and deep learning,” *Sensors*, vol. 20, no. 22, 2020, Art. no. 6481.

[30] H. ElMoaqet, M. Eid, M. Glos, M. Ryalat, and T. Penzel, “Deep recurrent neural networks for automatic detection of sleep apnea from single channel respiration signals,” *Sensors*, vol. 20, 2020, Art. no. 5037. [Online]. Available: https://www.mdpi.com/1424-8220/20/18/5037

[31] H. Diao et al., “Deep residual networks for sleep posture recognition with unobtrusive miniature scale smart mat system,” *IEEE Trans. Biomed. Circuits Syst.*, vol. 15, no. 1, pp. 111–121, Feb. 2021.

[32] M. Rossi et al., “Identification of characteristic points in multivariate physiological signals by sensor fusion and multi-task deep networks,” *Sensors*, vol. 22, no. 7, 2022, Art. no. 2684.

[33] H. Nakano, T. Ikeda, M. Hayashi, E. Ohshima, and A. Onizuka, “Effects of body position on snoring in apneic and nonapneic snorers,” *Sleep*, vol. 26, no. 2, pp. 169–172, 2003.

[34] W.-C. Chen et al., “Treatment of snoring with positional therapy in patients with positional obstructive sleep apnea syndrome,” *Sci. Rep.*, vol. 5, Dec. 2015, Art. no. 18188.

[35] H. E. Romero, N. Ma, G. J. Brown, and E. A. Hill, “Acoustic screening for obstructive sleep apnea in home environments based on deep neural networks,” *IEEE J. Biomed. Health Inform.*, vol. 26, no. 7, pp. 2941–2950, Jul. 2022.

[36] E. D. Chan, M. M. Chan, and M. M. Chan, “Pulse oximetry: Understanding its basic principles facilitates appreciation of its limitations,” *Respir. Med.*, vol. 107, no. 6, pp. 789–799, 2013.

[37] K. Basic, H. Fox, J. Spießhöfer, T. Bitter, D. Horstkotte, and O. Oldenburg, “Improvements of central respiratory events, cheyne-stokes respiration and oxygenation in patients hospitalized for acute decompensated heart failure,” *Sleep Med.*, vol. 27–28, pp. 15–19, 2016.

[38] H. Korkalainen et al., “Detailed assessment of sleep architecture with deep learning and shorter epoch-to-epoch duration reveals sleep fragmentation of patients with obstructive sleep apnea,” *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, pp. 2567–2574, Jul. 2021.

[39] J. Sim and C. C. Wright, “The kappa statistic in reliability studies: Use, interpretation, and sample size requirements,” *Phys. Ther.*, vol. 85, pp. 257–268, Mar. 2005.

[40] B. L. Koley and D. Dey, “Real-time adaptive apnea and hypopnea event detection methodology for portable sleep apnea monitoring devices,” *IEEE Trans. Biomed. Eng.*, vol. 60, no. 12, pp. 3354–3363, Dec. 2013.

[41] T. Van Steenkiste, W. Groenendaal, D. Deschrijver, and T. Dhaene, “Automated sleep apnea detection in raw respiratory signals using long short-term memory neural networks,” *IEEE J. Biomed. Health Inform.*, vol. 23, no. 6, pp. 2354–2364, Nov. 2019.

[42] U. Erdenebayar, Y. J. Kim, J. Park, E. Y. Joo, and K. Lee, “Deep learning approaches for automatic detection of sleep apnea events from an electrocardiogram,” *Comput. Methods Programs Biomed.*, vol. 180, 2019, Art. no. 105001.

[43] H. Sharma and K. K. Sharma, “Sleep apnea detection from ECG using variational mode decomposition,” *Biomed. Phys. Eng. Exp.*, vol. 6, no. 1, 2020, Art. no. 015026.

[44] F. Bozkurt, M. K. Uçar, M. R. Bozkurt, and C. Bilgin, “Detection of abnormal respiratory events with single channel ECG and hybrid machine learning model in patients with obstructive sleep apnea,” *IRBM*, vol. 41, pp. 241–251, 2020.

[45] K. W. To, T. O. Chan, W. C. Chan, K. L. Choo, and D. S. C. Hui, “Using a portable monitoring device for diagnosing obstructive sleep apnea in patients with multiple coexisting medical illnesses,” *Clin. Respir. J.*, vol. 15, pp. 1104–1112, Oct. 2021.

[46] D. Huysmans et al., “Sleep diagnostics for home monitoring of sleep apnea patients,” *Front. Digit. Health*, vol. 3, 2021, Art. no. 685766.

[47] K. Qian et al., “Can machine learning assist locating the excitation of snore sound? a review,” *IEEE J. Biomed. Health Inform.*, vol. 25, no. 4, pp. 1233–1246, Apr. 2021.

[48] M. H. Rahmani, R. Berkvens, and M. Weyn, “Chest-worn inertial sensors: A survey of applications and methods,” *Sensors*, vol. 21, 2021, Art. no. 2875. [Online]. Available: https://www.mdpi.com/1424-8220/21/8/2875

[49] F. Vaquerizo-Villar et al., “A convolutional neural network architecture to enhance oximetry ability to diagnose pediatric obstructive sleep apnea,” *IEEE J. Biomed. Health Inform.*, vol. 25, no. 8, pp. 2906–2916, Aug. 2021.

[50] M. Awais et al., “A hybrid DCNN-SVM model for classifying neonatal sleep and wake states based on facial expressions in video,” *IEEE J. Biomed. Health Inform.*, vol. 25, no. 5, pp. 1441–1449, May 2021.