INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Obstructive sleep apnea screening with breathing sounds and respiratory effort: a multimodal deep learning approach

Hector E. Romero[1], *Ning Ma*[1], *Guy J. Brown*[1], *Sam Johnson*[2]

[1]Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
[2]PFL Healthcare Limited, HQ 5th Floor, 58 Nicholas Street, Chester CH1 2NP, UK
{h.e.romero.ramirez, n.ma, g.j.brown}@sheffield.ac.uk, sam.johnson@passionforlife.com

Abstract

Obstructive sleep apnea (OSA) is a chronic and prevalent condition with well-established comorbidities. Due to limited diagnostic resources and high cost, a significant OSA population lives undiagnosed, and accurate and low-cost methods to screen for OSA are needed. We propose a novel screening method based on breathing sounds recorded with a smartphone and respiratory effort. Whole night recordings are divided into 30-s segments, each of which is classified for the presence or absence of OSA events by a multimodal deep neural network. Data fusion techniques were investigated and evaluated based on the apnea-hypopnea index estimated from whole night recordings. Real-world recordings made during home sleep apnea testing from 103 participants were used to develop and evaluate the proposed system. The late fusion system achieved the best sensitivity and specificity when screening for severe OSA, at 0.93 and 0.92, respectively. This offers the prospect of inexpensive OSA screening at home.

Index Terms: obstructive sleep apnea, acoustic analysis, respiratory effort, multimodal, deep learning, sleep-disordered breathing

1. Introduction

Sleep-disordered breathing (SDB) is a debilitating condition that affects a significant proportion of the population; a typical US study shows that 24% of middle-aged men and 9% of middle-aged women are affected [1]. Obstructive sleep apnea (OSA) is the most common sleep-related breathing disorder, which repeatedly interrupts breathing during sleep, leading to desaturation in blood oxygen level. As a result, OSA is often associated with fatigue, daytime sleepiness, and increased risk of stroke, heart attack, high blood pressure and diabetes [1, 2]. There is also recognised prevalence of OSA in patients recovering from COVID-19 [3]. However, a significant proportion of OSA patients are not diagnosed until these other medical problems become apparent [2, 4]. This is in part because current OSA diagnosis using polysomnography (PSG) is uncomfortable, time-consuming and expensive, requiring patients to stay overnight in a sleep clinic with multiple wired sensors attached to their head and body. The availability of sleep clinics has been further strained by the COVID-19 pandemic, which has placed significant demand on respiratory wards [5]. Rapidly rising patient numbers means it is now much more likely that a sleep test takes place at home using home sleep apnea testing (HSAT) equipment, but this equipment is similarly uncomfortable and expensive – as well as being less accurate than the hospital-based study. Effective home screening methods using inexpensive and less invasive equipment are therefore needed in order to allow earlier interventions (such as lifestyle changes) to be made and identify those that require PSG to fully assess their condition, thus making better use of scarce resources.

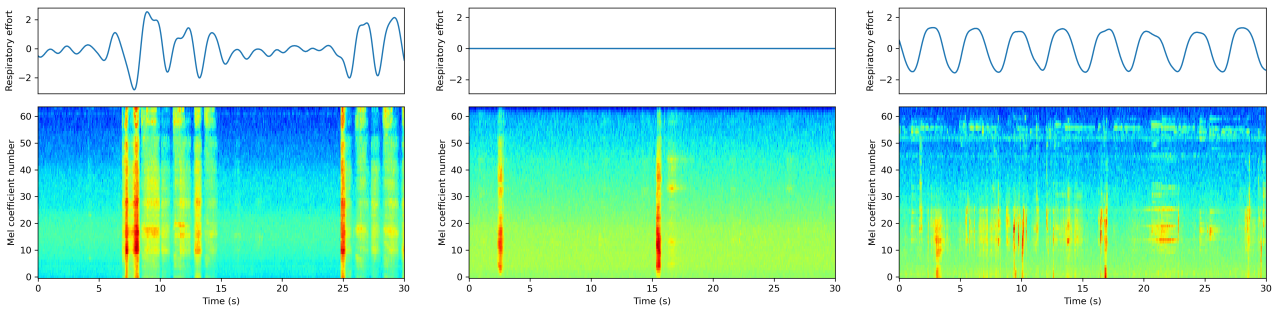
OSA events are often associated with unique acoustic characteristics including a sequence of acoustic events such as snores, chokes, loud gasps and absence of breathing. There have been a number of studies that investigated low-cost audio-based solutions for SDB assessment at home [6–10]. Other studies employed different data modalities for OSA screening. Memis et al. [11] proposed a feature-based or *early fusion* approach to screen for OSA. Features from electrocardiography (ECG) and blood oxygen saturation (SpO_2) were combined to train a support vector machine (SVM) classifier. Prabha et al. [12] developed a decision-based or *late fusion* method to screen for OSA. They trained one SVM classifier from heart rate variability features (i.e., ECG) and another using respiratory effort. The outputs of both classifiers were combined to obtain the final decision. Yadollahi et al. [13] used SpO_2 and acoustic features to classify tracheal audio recording segments into 'normal' and 'apneic'. SpO_2 was employed to extract audio segments from 10 seconds before the start of a desaturation. Then acoustic and SpO_2 features computed from the extracted segment were passed through sigmoid functions and summed to obtain the final decision. A similar approach was proposed by Saha et al. [13] using tracheal audio recordings, SpO_2 and body movement. Castillo-Escario et al. [6] also made use of SpO_2 to extract audio recording segments and looked for silent regions with a duration of at least 6 seconds based on sample entropy.

There are however two major challenges for realistic home-monitoring: 1) audio may be corrupted by background noise or interfered by snoring from the bed partner [14]; 2) sensors may be unreliable or missing (e.g., incorrectly fitted or falling off during sleep). To overcome these, this study proposes a multimodal approach through a combination of audio recordings and abdominal respiratory effort. The abdominal respiratory effort was selected because the sensor can be placed unobtrusively on the body and is less prone to falling off compared to other typically used sensors (e.g., nasal cannulas and pulse oximeters). It can also be implemented using small and low-cost accelerometers.

The novelty of this paper is twofold. First, different fusion methods of the acoustic signal and the abdominal respiratory effort signal are investigated in a multimodal deep learning framework. Second, previous methods have largely been focused on acoustically well controlled conditions in a clinical setting, whereas the signals used in this study were collected in over 100 real home scenarios. The results demonstrated realistic performance of the methods investigated.

The rest of this paper is organised as follows. Section 2 describes the OSA data used in this study. The proposed late fusion and early fusion systems for combining audio and respi-

10.21437/Interspeech.2023-209



(a) *Both data modalities available* (b) *No respiratory effort available* (c) *Background noise in audio*

Figure 1: *Abdominal respiratory effort and mel spectrogram for 30-second segments*

ratory effort data are described in Section 3. The experimental settings and the evaluation framework are introduced in Section 4. Section 5 presents and discusses the results and conclusions are given in Section 6.

2. Data

Participants undertook HSAT for one or two nights in a real home setting using a SOMNOtouch RESP device [15], which consists of several attached sensors that record SpO_2, heart rate, nasal airflow, respiratory effort, and body position. HSAT data was annotated for apnea events by a technologist with the RPSGT qualification certified by the Board of Registered Polysomnologists (USA, AASM). During each HSAT session, audio recordings were made simultaneously using a smartphone (iOS or Android) placed next to the bed at the head level. Audio recordings were collected using a purpose-built app, which records sound with a sampling frequency of 16 kHz and 16-bit resolution. In total 157 nights of recordings from 103 participants (Table 1) were included in this study. Data collection and storage protocols were subjected to the ethical review procedures of the University of Sheffield.

Table 1: *Demographics of the participants included in this study. The percentages of data groups, data ranges, and averages with standard deviations are also given.*

Total Participants	**103**	
Male	67	65%
Female	36	35%
Age (years)	45 ± 13	25 – 71
BMI (kg/m^2)	31 ± 7	19 – 48
Total Nights	**157**	
Night duration (hours)	7.0 ± 1.4	3.0 – 9.8
Healthy: AHI $<$ 5 (nights)	13	8%
Mild: 5 $\leq$ AHI $<$ 15 (nights)	67	43%
Moderate: 15 $\leq$ AHI $<$ 30 (nights)	38	24%
Severe: AHI $\geq$ 30 (nights)	39	25%

The scored HSAT data (inhalations, exhalations, desaturations, snores, and apnea-hypopnea events) was used as a reference for the acoustic recordings. Because the HSAT device clock may not be accurate or tightly synchronised with the smartphone, the HSAT data needed to be synchronised to the audio recordings. A 20-minute segment of the audio recordings after the HSAT recording started was used to compute the synchronisation delay with the snore channel signal from HSAT based on cross-correlation [10].

3. System description

We hypothesise that the integration of these data modalities can result in more robust screening systems in comparison with those that only use one modality, since both streams of data are complementary. Unlike audio recordings, the respiratory effort signal is not affected by background noise, but it is more prone to signal loss, as it is measured with a sensor attached to the body that can fall off or become disconnected from the HSAT device. Our analysis of the data from 157 nights of recordings shows that around 7% of the recordings have signal loss in the respiratory effort sensor. In contrast, audio recordings using a smartphone placed on a bedside table do not have this issue, but may capture background noise or interfering breathing sounds from a bed partner in addition to those of the subject under the study. Therefore, combining both data modalities could overcome the disadvantages of individual modalities [16].

Examples of both streams of data are presented in Fig. 1. In Fig. 1a, the respiratory effort is present with clean recordings of breathing sounds. In Fig. 1b, the respiratory effort signal is missing; whereas, in Fig. 1c, background noise is present in the sound recording.

This study investigates different multimodal deep learning approaches to screen for OSA based on the analysis of breathing sounds and abdominal respiratory effort during sleep. Three mechanisms for fusing the data streams from the two modalities are considered – *early feature fusion*, *latent space fusion*, and *late decision fusion*.

3.1. Early feature fusion

For early feature-based fusion, features are combined immediately after they are extracted from the raw signals, and a single classifier is trained on the combined features [16]. This is illustrated in Fig. 2.

The audio and respiratory effort signals are first divided into 30-s overlapping segments with a shifting window of 10 s. A long context window is employed so that an apnea event, which could last up to 30 seconds, can be captured within the window. Mel spectrograms are computed for the audio segments using a frame size of 50 ms, a frame rate of 50 Hz, a Hann window, and a bank of 64 mel filters spaced between 75 Hz and 7.5 kHz.

Each segment is labeled as either having apnea-hypopnea events inside it or not according to the scored HSAT data.

The raw respiratory effort signal, and its first and second order differences (i.e., delta and delta-delta) are appended as additional features to the mel spectrogram features. Since the raw respiratory effort signal is sampled at 32 Hz, it is upsampled to 50 Hz first in order to match the frame rate of mel spectrogram features. The combined features are provided as input to a convolutional neural network (CNN) system similar to the one employed in our previous study [10]. This network has the same architecture and parameters that we previously used, since the dimensionality of the resulting features (1,500 $\times$ 67) is close to that of mel spectrograms (1,500 $\times$ 64). It is made of three 2-dimensional convolutional layers of 16, 32, and 64 filters with a kernel size of 3 $\times$ 3. Each convolutional layer is followed by a 4 $\times$ 3 max-pooling layer, and a batch normalisation layer. All convolutional layers use ReLU activation, and a dropout rate of 0.3. The output of the convolutional layers is flattened and passed to a dense layer with 512 ReLU activation units. Lastly, a dense layer with one sigmoid unit performs the classification. The network has 0.7 million parameters.

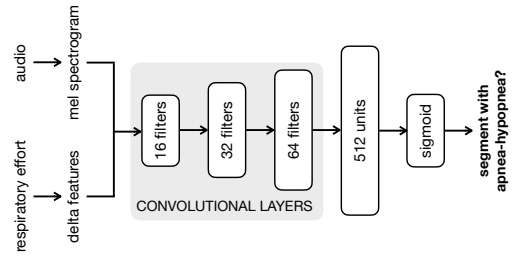


Figure 2: *System architecture for early feature fusion*

3.2. Latent space fusion

For latent space fusion (Fig. 3), the audio mel spectrograms features and the respiratory effort signal are fed to separate CNN filters. The extracted features from the two pathways of convolutional layers (i.e., latent space) are flattened, concatenated, and passed to a dense layer of 512 ReLU units and a dense layer of one sigmoid unit for classification. The remaining parameters are identical to the parameters of the single-modality classifiers that will be introduced next. In this way, the network is optimised on both data modalities at the same time, and does not require any additional processing of the output, as opposed to late decision fusion. This network has 2.4 million parameters.

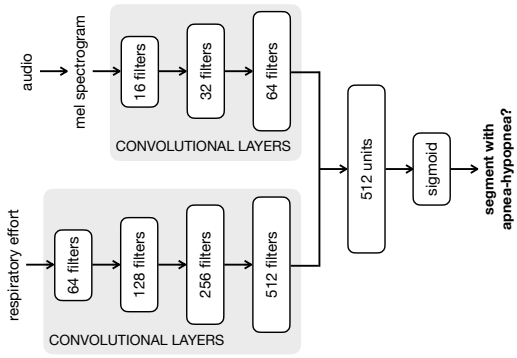


Figure 3: *System architecture for latent space fusion*

3.3. Late decision fusion

For late decision-based fusion, a classifier is trained independently for each data modality. Then, individual classifier outputs are integrated by computing the sum or product of probability to produce the final output [16]. The sum of probability can be seen as the arithmetic mean of the outputs, while the product of probability can be regarded as the geometric mean of the outputs [17].

An overview of the proposed late fusion system is given in Fig. 4. The breathing sounds classifier (upper half of Fig. 4) is based on the system proposed in our previous study [10], which is a CNN that takes as input the mel spectrogram for a 30-second segment, and outputs the probability of the segment containing apnea-hypopnea events. The respiratory effort classifier (lower half of Fig. 4) is a CNN based on architectures that have proven successful for the analysis of 1-dimensional physiological signals such as ECG [18–20]. Similar to its acoustic counterpart, it takes as input the raw effort signal for a 30-s segment, and outputs the probability of the segment containing apnea-hypopnea events. The network consists of 4 1-dimensional convolutional layers of 64, 128, 256, and 512 filters with a kernel size of 1 $\times$ 8. A 1 $\times$ 4 max-pooling layer, and a batch normalisation layer follow each convolutional layer. Other parameters are the same as those of the breathing sounds classifier. The late fusion system has 3.7 million parameters in total.

All systems converged within 50 epochs using TensorFlow [21], a learning rate of 0.001, a batch size of 64, the Adam optimiser, and binary cross-entropy as the loss function.

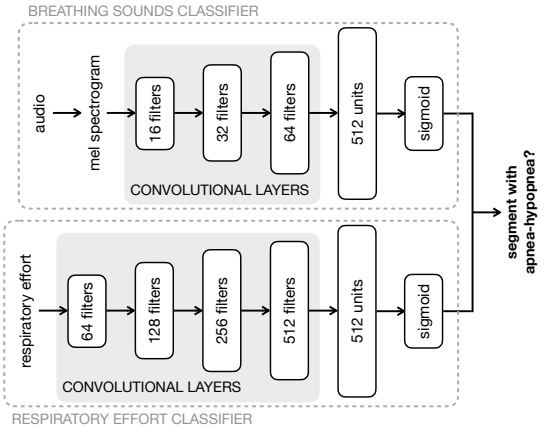


Figure 4: *System architecture for late decision fusion*

3.4. AHI estimation

OSA screening is typically based on the apnea-hypopnea index (AHI) – the average number of OSA events per hour over a night. Our proposed segment-based OSA classification systems do not directly estimate AHI. Since a typical apnea or hypopnea lasts around 30 seconds and segments overlap, adjacent 30-second segments predicted as containing OSA events can be assumed to belong to the same event. Consecutive segments that have the same predicted labels are merged into single events. Then, the AHI is computed as the quotient of the number of events and the duration of the recording in hours.

4. Evaluation

Experiments were conducted on the annotated OSA corpus, described in Section 2, using 10-fold cross-validation – 10 partic-

ipants per fold with the exception of the last one, which had 13. For each cross-validation round, one fold was used for testing; one fold, for validation; and the remaining 8 folds, for training.

The single-modality systems used for the late fusion approach (i.e., the breathing sounds classifier and respiratory effort classifier) were used as the baseline systems, since they evidence the performance of our OSA screening systems when only one data modality is available. Sensitivity (Eq. 1), specificity (Eq. 2), and area under the receiver-operating characteristic curve (AUC) were computed from the pooled results for all system configurations at the generally accepted AHI cut-off points: 5, 15, and 30 events/hour. These metrics indicate the diagnostic or screening capability of a test.

$$\text{sensitivity} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \quad (1)$$

$$\text{specificity} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}} \quad (2)$$

5. Results and discussion

Table 2 lists the sensitivity, specificity, and AUC of the different OSA screening systems at various recognised AHI cut-off points. Compared to the audio only system, the respiratory effort only system achieved better specificity but had lower sensitivity (< 0.7), failing to identify many true OSA events. From a clinical point of view, false positives would be more tolerable than false negatives when screening for a condition with serious comorbidities like OSA. However, false positives would result in reduced diagnostic availability and increased diagnostic cost, as more people would go through PSG, the gold standard for OSA diagnosis [22]. By combining information from both data modalities, the fusion systems in general attained a better performance than the audio only baseline, which evidences that the integration of complementary data modalities is indeed beneficial for the OSA screening task.

The overall best performance was achieved by the *late fusion – sum of probability* system (sensitivity: 0.93, specificity: 0.92, AUC: 0.95) when screening for severe OSA (i.e., AHI $\geq$ 30 events/hour), outperforming the baseline systems. The sensitivity was especially improved, from 0.78 for the audio only system and 0.69 for the respiratory effort only system to 0.93. This is likely due to the fact that OSA events are in general under-represented in the training data and systems have a tendency to under-report OSA events. At the AHI cut-off of 30, 39 nights are considered to be severe OSA whereas 118 nights are mild or healthy. Both late fusion systems were able to exploit the two data modalities, achieving an improvement in sensitivity while maintaining good specificity. At other AHI cut-off points, the late fusion systems also retained a similar sensitivity to the audio only system while having an improved specificity. The integration of these data modalities resulted in a better balance between sensitivity and specificity as well.

Although the *latent space fusion* system was outperformed by the late fusion approaches, it showed similar improvement over the single-modality systems. Furthermore, the former has the advantage of being contained in a single neural network model that takes as input both data modalities rather than consisting of two separate models. This facilitates its deployment in low-resource mobile devices. The *early feature fusion* approach did not perform better than the baseline systems. It was likely caused by the manner respiratory effort features and mel spectrograms were concatenated: effort features were appended next to the higher mel coefficients, and the CNN exploits local patterns. The relationship between effort features and the whole mel spectrogram might not have been fully learned by the network and a different network architecture might be needed.

Table 2: *Performance of OSA screening systems at different AHI cut-off points. In the second row, the split of negative | positive OSA nights (i.e., below vs. above cut-off point) is shown.*

	AHI	**5**	**15**	**30**
	157 nights	13 | 144	80 | 77	118 | 39
Audio only	Sensitivity	0.86	0.81	0.78
	Specificity	0.62	0.84	0.93
	AUC	0.75	0.84	0.92
Respiratory effort only	Sensitivity	0.64	0.57	0.69
	Specificity	0.94	0.97	0.99
	AUC	0.84	0.91	0.94
Early feature fusion	Sensitivity	0.74	0.67	0.67
	Specificity	0.53	0.72	0.88
	AUC	0.74	0.79	0.86
Latent space fusion	Sensitivity	0.81	0.81	0.88
	Specificity	0.82	0.74	0.87
	AUC	0.87	0.86	0.93
Late fusion – sum of probability	Sensitivity	0.86	0.81	**0.93**
	Specificity	0.82	0.77	**0.92**
	AUC	0.87	0.89	**0.95**
Late fusion – product of probability	Sensitivity	0.81	0.81	0.86
	Specificity	0.94	0.82	0.94
	AUC	0.92	0.90	0.94

One of the study's objectives was to establish a baseline for the development of robust OSA screening systems that can run on smartphones along with novel and low-cost hardware for the recording of respiratory effort. Low-cost sensors, such as Bluetooth-enabled accelerometers or gyroscopes instead of the more expensive HSAT equipment, can be paired to a smartphone to record abdominal respiratory movement as a surrogate for respiratory effort. Together with acoustics from a smartphone microphone, the proposed approach offers the prospect of inexpensive, continuous monitoring for SDB at home which facilitates treatment intervention.

6. Conclusions

Robustly screening for OSA in a real home environment during sleep is a challenging task. There are two main problems: 1) ambient sound recordings can be affected by background noise; 2) sensor data could be corrupted or missing due to incorrect fitting or falling off during sleep. This paper has proposed a novel solution by exploiting the temporal pattern of breathing sounds and integrating complementary information from abdominal respiratory effort signals in a multimodal deep learning framework. Evaluated using data collected in over 100 real home scenarios, the best performance was obtained with the late fusion approach, which offers an inexpensive home-based solution for accurate and reliable assessment of SDB. Future work will investigate ways to automatically fall back to single-modality approaches when data is missing, corrupted or affected by noise. We have also identified the analysis of a subject and their bed partner's breathing sounds during sleep as an area for future research.

7. References

[1] T. Young, P. E. Peppard, and D. J. Gottlieb, "Epidemiology of obstructive sleep apnea: a population health perspective," *Am J Respir Crit Care Med*, vol. 165, no. 9, pp. 1217–1239, 2002.

[2] K. K. Motamedi, A. C. McClary, and R. G. Amedee, "Obstructive sleep apnea: a growing problem," *Ochsner Journal*, vol. 9, pp. 149–153, 2009.

[3] S. Strausz, T. Kiiskinen, M. Broberg, S. Ruotsalainen, J. Koskela, A. Bachour, A. Palotie, T. Palotie, S. Ripatti, and H. M. Ollila, "Sleep apnea is a risk factor for severe COVID-19," *BMJ Open Respiratory Research*, vol. 8, no. 1, 2021.

[4] A. Castaneda, E. Jauregui-Maldonado, I. Ratnani, J. Varon, and S. Surani, "Correlation between metabolic syndrome and sleep apnea," *World Journal of Diabetes*, vol. 9, no. 4, pp. 66–71, April 2018.

[5] A. Voulgaris, L. Ferini-Strambi, and P. Steiropoulos, "Sleep medicine and COVID-19. Has a new era begun?" *Sleep Medicine*, vol. 73, pp. 170–176, September 2020.

[6] Y. Castillo-Escario, I. Ferrer-Lluis, J. M. Montserrat, and R. Jane, "Entropy Analysis of Acoustic Signals Recorded With a Smartphone for Detecting Apneas and Hypopneas: A Comparison with a Commercial System for Home Sleep Apnea Diagnosis," *IEEE Access*, vol. 7, pp. 128 224–128 241, 2019.

[7] T. Kim, J. W. Kim, and K. Lee, "Detection of sleep-disordered breathing severity using acoustic biomarker and machine learning techniques," *Biomedical Engineering Online*, vol. 17, 2018.

[8] H. Nakano, T. Furukawa, and T. Tanigawa, "Tracheal sound analysis using a deep neural network to detect sleep apnea," *Journal of Clinical Sleep Medicine*, vol. 15, no. 8, pp. 1125–1133, 2019.

[9] R. Tiron, G. Lyon, H. Kilroy, A. Osman, N. Kelly, N. O'Mahony, C. Lopes, S. Coffey, S. McMahon, M. Wren, K. Conway, N. Fox, J. Costello, R. Shouldice, K. Lederer, I. Fietze, and T. Penzel, "Screening for obstructive sleep apnea with novel hybrid acoustic smartphone app technology," *Journal of Thoracic Disease*, vol. 12, no. 8, pp. 4476–4495, 2020.

[10] 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 Journal of Biomedical and Health Informatics*, vol. 26, no. 7, pp. 2941–2950, 2022.

[11] G. Memis and M. Sert, "Multimodal classification of obstructive sleep apnea using feature level fusion," in *2017 IEEE International Conference on Semantic Computing (ICSC)*. IEEE, 2017, pp. 85–88.

[12] A. Prabha, A. Trivedi, A. A. Kumar, and C. S. Kumar, "Automated system for obstructive sleep apnea detection using heart rate variability and respiratory rate variability," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. Udupi, India: IEEE, 2017, pp. 1303–1307.

[13] S. Saha, M. Kabir, N. M. Ghahjaverestan, M. Hafezi, B. G. anf K. Zhu, H. Alshaer, and A. Yadollahi, "Portable diagnosis of sleep apnea with the validation of individual event detection," *Sleep Medicine*, vol. 69, pp. 51–57, 2020.

[14] H. E. Romero, N. Ma, and G. J. Brown, "Snorer diarisation based on deep neural network embeddings," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 876–880.

[15] SOMNOmedics GmbH, "SOMNOtouch RESP," https://somnomedics.de/en/solutions/sleep_diagnostics/polygraphy-devices/somnotouch-resp/, accessed: June 11, 2021.

[16] T. Baltrusaitis, C. Ahuja, and L. P. Morency, "Multimodal machine learning: a survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[17] B. Xie and H. Minn, "Real-time sleep apnea detection by classifier combination," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 3, p. 469–477, 2012.

[18] M. Sallem, A. Ghrissi, A. Saadaoui, and Z. V., "Detection of cardiac arrhythmias from varied length multichannel electrocardiogram recordings using deep convolutional neural networks," in *2020 Computing in Cardiology*, Rimini, Italy, 2020, pp. 1–4.

[19] M. Wu, Y. Lu, W. Yang, and S. Y. Wong, "A study on arrhythmia via ECG signal classification using the convolutional neural network," *Frontiers in Computational Neuroscience*, vol. 14, 2021.

[20] M. Chourasia, A. Thakur, S. Gupta, and A. Singh, "ECG heartbeat classification using CNN," in *2020 IEEE 7th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, Prayagraj, India, 2020, pp. 1–6.

[21] M. Abadi et al., "TensorFlow: a system for large-scale machine learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation*. USENIX, 2016, pp. 265–283.

[22] L. D. Maxim, R. Niebo, and M. J. Utell, "Screening tests: a review with examples," *Inhalation Toxicology*, vol. 26, no. 13, pp. 811–828, 2014.