Obstructive Sleep Apnea identification based on VGGish networks

Salvatore Serrano, Luca Patanè and Marco Scarpa
University of Messina
Department of Engineering
Messina, Italy
Email: {sserrano, lpatane, mscarpa}@unime.it

KEYWORDS

Obstructive Sleep Apnea, convolutional neural network, classification, spectrogram image analysis.

ABSTRACT

Sleep disorders are continuously growing in the population and can have a significant negative impact on everyday life. Economic and non-invasive systems able to support the diagnosis procedure will be more and more adopted in the next years. The aim of this work is to investigate the classification performance of a convolutional neural network, based on a VGG structure, to identify obstructive sleep apnea events. A recently developed dataset containing audio signals recorded from high-quality contact microphones placed on the trachea of the subjects under study has been adopted to perform transfer learning over a pre-trained VGGish network. Spectrogram images have been extracted from the audio signals to serve as inputs for the classification process. The importance of the time window selection has been also investigated and comparisons with other recent methods proposed in the literature are reported.

INTRODUCTION

Humans spend a third of their life sleeping, so sleep plays an important role in staying healthy [1]. Unfortunately, the quality of life of several people is affected by a hidden sleep pathology denoted as Obstructive Sleep Apnea-Hypopnea Syndrome (OSAHS) [2][3].

OSAHS is generated by the narrowing of the upper airway, at multiple levels. During awake time the increased muscle tone prevents the upper airways from collapsing. Reversely, during sleep, the combination of the extraluminal pressure exerted from surrounding soft tissue structures and negative intraluminal pressure of the upper airway during inspiration, this can result in upper airway collapse. Obese subjects can be affected by a further reduction of the upper airway caliber causing more severe clinical consequences.

Subjects affected by sleep apnea can decrease their sleep quality, resulting in drowsiness during the day, poor memory, an increased risk of accidents owing to extreme sleepiness and, in general, low productivity. In more serious cases, in adult subjects, OSAHS can cause hypertension, coronary heart disease, stroke, arrhythmia, and other diseases while, in infants, it can cause behavioral disorders and even sudden death [4]. Adults affected by this pathology have an increased probability to cause traffic accidents, they usually are affected by mood swings and depression and, accordingly, OSAHS has also a not negligible financial and social impact [5]. Nowadays, approximately $6\% - 13\%$ of the world population suffer from this disease [6] but 80% of apnea patients remain undiagnosed [7]. The apnea/ hypopnea index (AHI) is used to determine the severity of OSAHS. The index counts the number of apnea and hypopnea per hour of sleep. A single apnea is defined by a drop in the peak respiratory airflow by $\geq 90\%$ from the baseline and the duration of the event lasts at least 10 seconds [8]. OSAHS will be classified as mild if AHI is in the range $[5-15]$, moderate if AHI is in the range $[16-30]$ and severe if AHI is greater than 30. In order to evaluate the AHI, patients must undergo a clinical examination named Polysomnography (PSG), which is usually conducted in hospitals. PSG records several bio-signals such as respiratory, heart-beat, movement, snoring, oxygen saturation, pharyngeal movement, and others. To obtain all these signals the patients have to wear a device capable of synchronously recording signals from several sensors, usually linked to the recorder device in a wired way. All these cables, sensors and devices cause discomfort for the patients, typically in a non-homecare unfriendly experience [9]. As a consequence, not detected wrong OSAHS diagnosis frequently occurs and a high number of OSAHS-affected patients are untreated [10]. Accordingly, it's important to promote the development of equipment and or technologies able to support the diagnosis of OSAHS in a more comfortable way [11]. In this work, we investigated the classification performance of a deep convolutional neural network (CNN) based on a Visual Geometry Group (VGG) architecture, whose goal is to identify Obstructive Sleep Apnea Syndrome (OSAS) events by means of features extracted from a unique audio signal, recorded during sleep time.

The remaining of the paper is organized as follows: Section "RELATED WORKS" introduces related works; the proposed methodology for apnea event classification is described in Section "PROPOSED METHODOLOGY", the analysis of the simulation results is provided in Section "RESULTS" and finally, the conclusions are drawn in Section "CONCLUSIONS".

Communications of the ECMS, Volume 37, Issue 1,
Proceedings, ©ECMS Enrico Vicario, Romeo Bandinelli,
Virginia Fani, Michele Mastroianni (Editors) 2023
ISBN: 978-3-937436-80-7/978-3-937436-79-1 (CD) ISSN 2522-2414

RELATED WORKS

Recently, several researchers focused on sound-based OSAHS identification. In one of the first works tackling this issue, the authors proposed a classifier of snore sounds based on spectrogram image analysis performed by CNNs [12]. They use the Munich-Passau Snore Sound Corpus as a dataset; it contains 828 snore samples from four classes which reflect the place of obstruction causing the snore: Velum, Oropharyngeal, Tongue, and Epiglottis. Although the snore can be a marker for OSAHS, the dataset was not annotated accordingly to the occurrence of apnea episodes.

In [13], the authors proposed a system to identify OSAHS based on the recording of video and audio of a patient. Specifically, from the audio signal, the authors extract three different kinds of features: a mixed set of features with acoustic and prosodic parameters; features obtained by applying WPT to the spectrogram of the signal and features obtained by applying Nonnegative Matrix Factorization (NMF) to the spectrogram of the signal. The overall audio signal recorded for each patient ($16kHz$, 16 bit per sample) is divided into 10 s length segments with an overlap of 5 s. From each segment of audio sub-segments with a time length of 5 s and a step interval of 0.5 s were extracted. The spectrogram of each sub-segment was obtained by means of a Short Time Fourier Transform (STFT) evaluated with 25 ms STFT-window and 12.5 ms increments. For their experiments and performance evaluation, authors used a proprietary dataset made up of the recordings of 4 patients on two different nights each lasting about 480 minutes. Recordings of video and audio were performed synchronized with the signals of a PSG and, subsequently, they were manually marked according to 4 different classes: i) central apnea, ii) obstructive or mixed apnea, iii) hypopnea, and iv) all the other events that are available from the ground truth labels. They used as classifiers both Support Vector Machines (SVM) and Neural Networks (NN). Using only the audio component, the best performance in terms of accuracy (99.17%) was obtained using a SVM as a classifier and NMF features. An inverse 5-fold cross-validation (CV) (using 1-fold for training and the remaining 4-folds for testing) was used to obtain the above-mentioned result.

An OSAHS recognition algorithm based on CNN was proposed in [14]. The authors use Mel Frequency Cepstral Coefficients (MFFCs) as audio features. MFFCs were extracted using a time window of 40 ms. Three different architectures of CNNs were considered: VGGNet [15], Inception (GoogLeNet) and [16], ResNet [17]. Performance results, in terms of accuracy on the same dataset used in [12], were compared with a baseline algorithm based on a Gaussian Mixture Model (GMM) classifier. Both the proposed CNN architectures and baselines show very low performance in OSAHS classification (always lower than 50%) and, moreover, it appears not clear how the snore-based dataset in [12] was annotated according to apnea events.

Authors of [18] and [19] propose a wearable system to recognize human contexts such as breathing, heartbeat pattern, and swallowing using audio sensors. Specifically, the proposed device contains two different microphones: the first one is an open-air mic and the second one is a contact mic; both mics are housed on the same body and record audio signals synchronously. A specific dataset was built to perform an evaluation of the best positioning of the device and to evaluate the classification accuracy of five possible audio events corresponding to breathing, swallowing, movement, oral sound and others. The dataset contains the recordings of seven healthy people which mimic typical movements and sounds of a sleeping person. From the audio signal (sampled at $44100Hz$) a vector of 28 MFCCs-based parameters was extracted together with a vector of 14 parameters built of 10 FFT peaks and 4 statistical and time-domain features from windows of 1, 2 and $3s$ with a step of $0.25s$. Vectors were used to train models based on SVM and Random Forest (RF). A time frame of 4 minutes of the overall recording length of 5 minutes for each actor was used to train the models and the remaining 1 minutes was used for tests. Performance results and comparisons were presented both using data from single microphones and using data aggregation. Although the work shows the potential of the proposed wearable device and its better positioning on the patient body in order to increase the recording quality, classification results performed on a small mimed dataset and without a suitable OSAHS annotation, constitute two major drawbacks in order to prove the goodness of this approach for the OSAHS classification topic.

In 2021, an open and freely available dataset, comprising 212 polysomnograms along with synchronized high-quality tracheal and ambient microphone recordings was released [20]. The whole dataset was manually annotated by medical experts according to “respiratory” episodes, which include among others all apnea-related episodes of a specific type: “Obstructive Apnea”, “Central Apnea”, “Mixed Apnea” and “Hypopnea”. Accordingly, this dataset represents an important milestone to evaluate and compare the performance of OSAHS classifiers, and, specifically, OSAHS classifiers based on audio signals recordings.

In this paper, we exploited this dataset in order to propose an OSAHS classification approach based on CNNs. Specifically, we used a pre-trained network based on a VGGish model which is often adopted in sound classification tasks. We evaluated performance comparing results obtained by our approach versus already proposed techniques in literature, which are based on different features and classifiers. Consequently, the main novelties of this work consist in:

- the use of a pre-trained CNN based on a VGGish model adopting as inputs the mel-spectrograms extracted from audio signals to perform OSAHS classification;
- the use of a freely available dataset, recently developed, specifically and independently marked for OSAHS, for performing training, validation and test and

performance comparisons of the considered classification approaches;

- the impact of the time window length selected for the input data on the classification performance.

PROPOSED METHODOLOGY

Deep neural networks for audio processing

In the last years the drastic progress in terms of computational power, mostly related to the introduction of massively multi-core GPUs, pushed towards the use of deep learning techniques to develop classification and regression networks applied in many different domains [21]. Machine learning approaches were widely applied both in the fields of signal processing and telecommunications networks [22][23][24][25][26][27]. Sound classification and recognition are one of the investigated research fields including applications to music classification [28][29][30][31], speech recognition [32], prediction maintenance and others.

In this work, among the different solutions available in the literature, a convolutional network based on the VGG family introduced by Google was considered. it was introduced in 2014 as an evolution of Alexnet. The peculiar characteristics consist of the presence of the ReLU activation function and the use of small receptive fields in the convolutional filters (i.e., 3×3). This model was originally adopted for image processing and subsequently used by Google in 2017 on audio signals. The training was performed on a large YouTube dataset [33].

The pre-trained network contains 24 layers among which nine layers present learnable weights: six convolutional and three fully connected layers. The input provided to the VGGish consists of a series of mel-spectrograms obtained by decomposing the audio signals in a series of overlapped time frames. Due to the availability of small datasets related to the application in exam, it is not possible to train a CNN from the scratch, instead, a transfer learning approach can be adopted. The pre-trained VGG is then considered. In particular, this network provides in output a 128-element feature vector for each input pattern that can be extracted from the last fully connected layer. Introducing a final fully connected layer followed by softmax and classification layers on top of the network, it is possible to develop a classifier to distinguish the presence of apnea events from the audio signals. Following the transfer learning methodology, the network needs to be fine-tuned on the available dataset, specific to the case of study considered. This process was performed by increasing the learning rate factor for the newly added learnable layers by a factor of ten, to guarantee that the learning process is faster in the new layers than in the transferred ones that are only fine-tuned.

The block scheme of the proposed architecture is depicted in Fig. 1.

The first block identifies the dataset used to train, test and validate the proposed system. It is built starting from the sound data and, specifically, collecting and splitting the audio excerpts into training, validation and testing subset. The pre-processing block permits to obtain the inputs of the subsequent classification system. It splits the audio excerpts in overlapping frames of a specific time-length and evaluates the mel spectrogram for each frame; moreover, the audio excerpts are arranged in overlapping windows of a specific time-length and the Mel-based spectrogram is evaluated in each window. Each spectrogram contains the time-frequency representation of a short-time audio window and it is stored as a matrix of specific dimension. According to the length of the audio excerpts and to the time step with which each window is extracted, a specific number of matrices is obtained. The last block consists of two sub-blocks: the first one contains the original layer of the pre-trained VGGish CNN; the second one contains the output layer introduced in order to classify the input matrices according to the OSAHS outcomes. Both the sub-blocks are used in training and validation/testing steps in a different way. During the training step the parameters of the VGGish pretrained network are tuned in order to identify features in the matrices which higlight the differences between those obtained from sounds captured when the patients are affected by obstuctive apnea and those obtained from sounds captured during no apnea conditions. At the same way, at this step, the parameters of the classification layer are tuned to maximize an index of performance of the classification goodness of the feature vectors coming from the last layer of the VGGish CNN according to the two output classes. In validation/testing step these two sub-blocks are simply used in order to obtain the outcome of the classification using the tuned parameters in the training step.

To evaluate the performance of the proposed architecture, we analyzed the confusion matrices considering positive samples corresponding to apnea events and negative others. Following this definition, we can identify, comparing the true and the predicted classes, the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) outcomes that are used to calculate the performance indexes as follow:

- **Precision**: the ability to properly identify positive samples $P = \frac{TP}{TP+TN}$;
- **Recall**: the fraction of positive samples correctly classified $R = \frac{TP}{TP+FN}$;
- **Specificity**: the fraction of negative samples correctly classified $S = \frac{TN}{TN+FP}$;
- **Accuracy**: the fraction of predictions correctly classified $A = \frac{TP+FP}{TP+FP+FN+TN}$;
- **F1 score**: the harmonic mean between precision and recall $F1 = 2 \times \frac{P \times R}{P+R}$.

OSAHS Dataset

As described in [20], data were collected from 212 individuals, who visited the Sleep Study Unit of the Sismanoglio – Amalia Fleming General Hospital of Athens for SAS diagnosis. Audio signals were acquired and stored by means of a dual-channel portable multitrack recorder (Tascam DR-680 MK II) and synchronized with PSG. One of the channels was connected to a con-

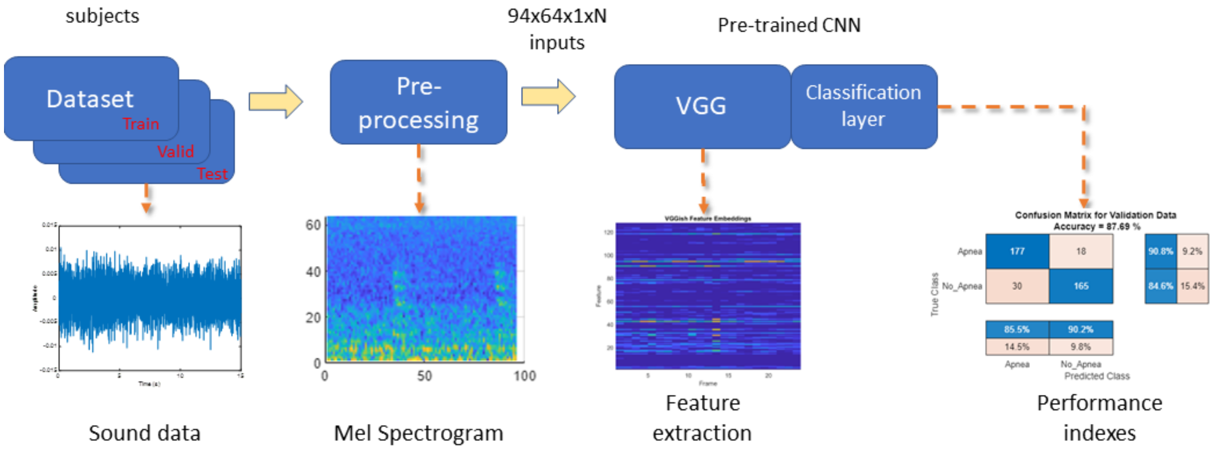

Fig. 1. Block scheme of the proposed method

tact microphone (Clockaudio CTH100) placed on the trachea of the patient. The second channel was connected to an ultra-linear measurement condenser microphone (Behringer ECM8000) placed approximately 1 m above the patient's bed, over the head position. Both sound signals are sampled at 48 kHz and originally stored using 24-bit per sample. In order to store audio signals synchronized with other PSG signals in European Data Format (EDF), the number of bits per sample was reduced to 16. PSG data consists of 16 channel including Electroencephalogram, Electroculogram, Leg movement signal, Electrocardiogram (ECG), RR interval in the ECG, pulse rate extracted by the ECG, thoracic volume changes, abdomen volume changes, nasal/oral flow pressure, the position of the body, oxygen level (oxygen saturation) of the blood. PSG study is performed by the health specialists of the Sleep Study Unit of the Sismanoglio – Amalia Fleming General Hospital of Athens. For each patient, sleep stages and apnea events are scored by two specialists: a certified technician performs first-level scoring and a 30-year-experienced and certified doctor performs final scoring, with verification of the true positive annotated events and addition of missed events.

Specifically, in our experiments, we only used the audio signals recorded from the high-quality contact microphone placed on the trachea of the first 25 patients. "Respiratory" type annotations were used as a reference and, in particular, they were grouped in a unique class named "Apnea" including "Obstructive Apnea", "Central Apnea" and "Mixed Apnea", "Hypopnea" events were excluded due to their nature of respiration frequency reduction. We extracted 15 s of audio starting from the position addressed by each "apnea" annotation from each recording of the considered patients. We obtained globally 430 excerpts of audio. In order to build a balanced dataset we collected for each excerpt annotated as "apnea" a corresponding piece of audio (extracted from the same recording), lasting 15 s, which, in reverse, does not contain annotations belonging to the "apnea" class. Globally, the considered dataset contains 860 non-overlapped excerpts of audio. Accordingly to the format of input required by VGGish, we extracted mel-spectrograms from each excerpt of audio each representing 1 s of audio and with a step of 0.25 ms. Each spectrogram was evaluated as a matrix of 96x64 values: 96 is the number of 25 ms frames in each mel-spectrogram and 64 represents the number of mel bands spanning from 125 Hz to 7.5 kHz. A performance comparison of our approach was conducted versus [13]. In current research literature, analyzed in section "RELATED WORKS", and to the best of our knowledge, [13] seems to be the most promising approach about OSAHS classification. Accordingly, from the same excerpts of audio, we extracted features obtained applying NMF to the spectrogram of the signal, as described in [13], since the authors have shown that this set of parameters guarantees the best results among those proposed.

RESULTS

A 5-fold cross-validation approach was used to train and test the performance of the proposed approach. After a preliminary optimization of the hyperparameters, we finally trained the VGGish networks using the Adam optimizer, a mini-batch size equal to 512 and a maximum number of epochs equal to 5. To determine the output class for the sequence of time windows behaving to the same audio block, every single prediction was combined using a majority-rule decision. Fig. 2 shows the confusion matrix chart with column (classwise precision) and row (class-wise recall) summaries obtained over the 5 folds. The obtained performance is balanced in terms of precision and recall and exceeds the 95% of accuracy.

In order to compare and evaluate performance results, we trained two different binary SVM classifiers as a baseline. The former (SVM1) was trained and tested using a unique vector obtained as proposed in [13]. The latter (SVM2) was trained using vectors obtained by each sub-segment lasting 5 s with 0.5 s increments as inputs and, subsequently, we applied a majority-rule decision to classify the entire excerpts. We show the confusion matrix chart for both baseline approaches in figs. 3-4.

Table I compares all performance indexes for the con-

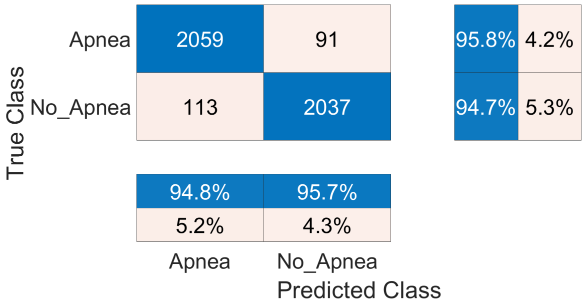

Fig. 2. A confusion matrix chart was obtained for the proposed VGGish-based approach.

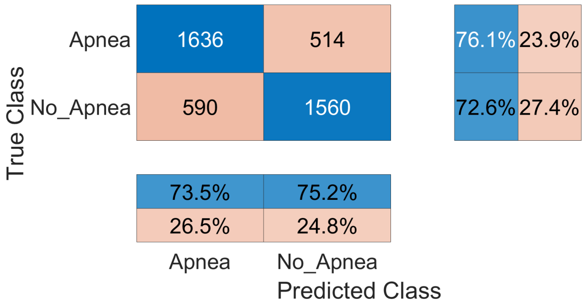

Fig. 3. A confusion matrix chart obtained adopting the SVM1 baseline approaches

sidered approaches. The proposed VGGish classifier outperforms SVMs for every considered parameter and, moreover, the modified version of the SVM approach (row SVM2) gives better results than the original one proposed in [13] (row SVM1). This result indicates that the classification of single time frames is preferable instead of a unique feature vector even in the presence of simple consensus methods.

We also analyzed the classification performance varying the size of the excerpts in a range of $[5-15]$ s. As can be noticed by analyzing Fig. 5, both VGGish and SVM2 show an increase of the F1 parameter until the excerpt sizes reach values around 10 s. Exceeding this threshold the F1 parameter maintains almost constant or, in VGGish case, decreases. The range of variations appears very small overall range for both approaches. This result demonstrates that the proposed method shows a limited reduction of performance (lower than 1.5%) if a $5s$ time window is considered whereas the SVM-based approach presents a degradation of about 5%.

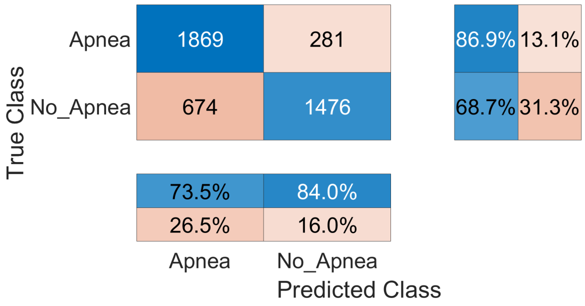

Fig. 4. A confusion matrix chart obtained adopting the SVM2 baseline approaches

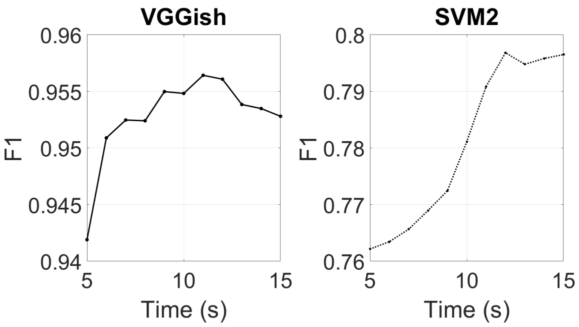

Fig. 5. Trend of the F1 index when the size of the excerpts used to perform classification changes within the range of $[5-15]$ s.

CONCLUSIONS AND FUTURE WORKS

The obstructive sleep apnea identification through audio signal processing was investigated by comparing the state-of-the-art methods based on the SVM classifier with the proposed CNN-based solutions developed using a transfer learning strategy on a VGGish network pre-trained using a large general-purpose audio dataset. The proposed network, receiving in input the audio spectrogram of recorded audio signals from a recently collected dataset, extracts 128 features that are used to classify apnea events. Several performance indexes were reported to compare the VGGish network with other solutions showing that the classification of single time frames with a majority-based consensus method is preferable if compared to the processing of a unique aggregated feature vector. Moreover, the degradation of performance when changing the analyzed time window was investigated showing that the VGGish solution is more resilient at shorter time intervals, up to 5 s than the other approaches compared.

As future work we planned: 1) to investigate performance of others CNN-based classifier on the same dataset (like the pre-trained YAMNet neural network); 2) to extend the dataset used both for training and testing the classifiers adding new patients; 3) to train and test classifiers which are capable to operate with the sounds captured by the environmental microphone of the dataset. If the performance of this latter system will be acceptable, the development of a prototype of a simple system capable to support a preliminary diagnosis of OSAHS can start. This kind of system can, for example, run as an application on patient smartphone, capturing environmental sound during the night, and it can permit the preliminary diagnosis of OSAHS without the need of complex and annoying PSG recording systems.

References

[1] M. K. Pavlova and V. Latreille. Sleep disorders. *The American Journal of Medicine*, 132(3):292–299, 2019.

[2] M. Armstrong et al. The effect of surgery upon the quality of life in snoring patients and their partners: a between-

TABLE I: Comparison between the proposed model and the state-of-the-art solutions.

	P (%)	R (%)	S (%)	A (%)	F1 (%)
VGGish	95.77	94.80	94.74	95.26	95.28
SVM1	76.09	73.50	72.56	74.33	74.77
SVM2	86.93	73.50	68.65	77.79	79.65

subjects case-controlled trial. *Clinical Otolaryngology & Allied Sciences*, 24(6):510–522, 1999.

[3] R. Gall et al. Quality of life in mild obstructive sleep apnea. *Sleep*, 16(suppl_8):S59–S61, 1993.

[4] K. Zhu et al. Vision-based heart and respiratory rate monitoring during sleep–a validation study for the population at risk of sleep apnea. *IEEE Journal of Translational Engineering in Health and Medicine*, 7:1–8, 2019.

[5] S. A. Imtiaz. A systematic review of sensing technologies for wearable sleep staging. *Sensors*, 21(5):1562, 2021.

[6] A. Sabil et al. Comparison of apnea detection using oronasal thermal airflow sensor, nasal pressure transducer, respiratory inductance plethysmography and tracheal sound sensor. *Journal of Clinical Sleep Medicine*, 15(2):285–292, 2019.

[7] I. Fietze et al. Prevalence and association analysis of obstructive sleep apnea with gender and age differences–results of ship-trend. *Journal of sleep research*, 28(5):e12770, 2019.

[8] R. B. Berry et al. The AASM manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine*, 176:2012, 2012.

[9] A. M. Bhutada et al. Obstructive sleep apnea syndrome (OSAS) and swallowing function—a systematic review. *Sleep and Breathing*, 24(3):791–799, 2020.

[10] L. Almazaydeh et al. Obstructive sleep apnea detection using SVM-based classification of ECG signal features. In *2012 annual international conference of the IEEE engineering in medicine and biology society*, pp. 4938–4941. IEEE, 2012.

[11] F. Mendonca et al. A review of obstructive sleep apnea detection approaches. *IEEE journal of biomedical and health informatics*, 23(2):825–837, 2018.

[12] S. Amiriparian et al. Snore sound classification using image-based deep spectrum features. 2017.

[13] C. Yang et al. Sleep apnea detection via depth video and audio feature learning. *IEEE Transactions on Multimedia*, 19(4):822–835, 2016.

[14] Q. Dong et al. Convolutional neural network-based obstructive sleep apnea identification. In *2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS)*, pp. 424–428. IEEE, 2021.

[15] L. Wang et al. Places205-vggnet models for scene recognition. *arXiv preprint arXiv:1508.01667*, 2015.

[16] C. Szegedy et al. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[17] S. Targ et al. Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029*, 2016.

[18] A. A. Maritsa et al. Audio-based wearable multi-context recognition system for apnea detection. In *2021 6th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, volume 6, pp. 266–273. IEEE, 2021.

[19] A. A. Maritsa et al. Apnea and sleeping-state recognition by combination use of open-air/contact microphones. In *INTERACTION 2022*, pp. 87–96. Information Processing Society of Japan (IPSJ), 2022.

[20] G. Korompili et al. PSG-audio, a scored polysomnography dataset with simultaneous audio recordings for sleep apnea studies. *Scientific data*, 8(1):1–13, 2021.

[21] L. Alzubaidi et al. Review of deep learning : concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8:Article number: 53, December 2021.

[22] K. Arulkumaran et al. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6):26–38, 2017.

[23] N. C. Luong et al. Applications of deep reinforcement learning in communications and networking: A survey. *IEEE Communications Surveys & Tutorials*, 21(4):3133–3174, 2019.

[24] M. Bkassiny et al. A survey on machine-learning techniques in cognitive radios. *IEEE Communications Surveys & Tutorials*, 15(3):1136–1159, 2012.

[25] S. Serrano et al. Random sampling for effective spectrum sensing in cognitive radio time slotted environment. *Physical Communication*, 49:101482, 2021.

[26] P. S. Bithas et al. A survey on machine-learning techniques for UAV-based communications. *Sensors*, 19(23):5170, 2019.

[27] C. Grasso et al. H-HOME: A learning framework of federated FANETs to provide edge computing to future delay-constrained IoT systems. *Computer Networks*, 219:109449, 2022.

[28] S. Serrano et al. A new fingerprint definition for effective song recognition. *Pattern Recognition Letters*, 160:135–141, 2022.

[29] M. A. B. Sahbudin et al. IOT based song recognition for FM radio station broadcasting. In *2019 7th International Conference on Information and Communication Technology (ICoICT)*, pp. 1–6. IEEE, 2019.

[30] M. A. B. Sahbudin et al. Mongodb clustering using k-means for real-time song recognition. In *2019 International Conference on Computing, Networking and Communications (ICNC)*, pp. 350–354. IEEE, 2019.

[31] S. Serrano and M. Scarpa. Fast and accurate song recognition: an approach based on multi-index hashing. In *2022 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, pp. 1–6, 2022.

[32] S. Alharbi et al. Automatic speech recognition: Systematic literature review. *IEEE Access*, 9:131858–131876, 2021.

[33] S. Hershey et al. CNN architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.