Proceedings Paper:

Romero, H., Ma, N., Brown, G. orcid.org/0000-0001-8565-5476 et al. (1 more author) (2024) SLUMBR: SLeep statUs estiMation from aBdominal Respiratory effort. In: 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 46th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, 15-19 Jul 2024, Orlando, Florida. Institute of Electrical and Electronics Engineers (IEEE) ISBN 9798350371505

https://doi.org/10.1109/EMBC53108.2024.10782490

SLUMBR: SLeep statUs estiMation from aBdominal Respiratory effort

Hector E. Romero[†‡], Ning Ma[†], Guy J. Brown[†] and Sam Johnson[‡]
[†]Department of Computer Science, University of Sheffield, Sheffield S1 4DP, United Kingdom
[‡]Passion for Life Healthcare (UK) Limited, Chester CH1 2NP, United Kingdom
Email: {h.e.romero.ramirez, n.ma, g.j.brown}@sheffield.ac.uk, {hector.romero, sam.johnson}@passionforlife.com

***Abstract*—Accurately monitoring sleep for extended periods remains a challenge due to the cumbersome nature of conventional gold-standard techniques. We propose a novel deep learning method to estimate sleep status from an easily acquired abdominal respiratory effort signal. Our end-to-end convolutional neural network, developed on 476 hours of manually annotated polysomnography recordings from 53 participants, achieves an area under the curve of 0.90, and a more balanced performance across sensitivity and specificity than previous studies: 0.85 and 0.82, respectively. This method eliminates the need for obtrusive equipment and manual processing, paving the way for more accessible sleep monitoring solutions.**

I. INTRODUCTION

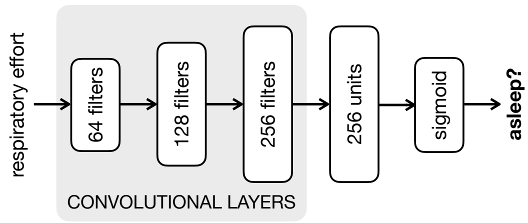
Sleep-disordered breathing (SDB) is a group of highly prevalent and comorbid conditions caused by the collapse of the upper airway during sleep, such as snoring and obstructive sleep apnoea (OSA). In recent years, there has been a growing interest in methods for screening and monitoring SDB progression in the home. These methods can accurately predict breathing parameters and overcome the high cost and obtrusiveness of polysomnography (PSG), the clinical standard for SDB diagnosis. However, most at-home SDB monitoring methods do not assess sleep quality, which is negatively impacted by SDB. Poor sleep quality can lead to excessive sleepiness, daytime dysfunction, increased risk of accidents, and long-term cognitive impairment [1].

Overnight PSG involves manual annotation of sleep stages, from which sleep status (i.e., sleep-wake classification) is derived, using electroencephalography (EEG), electromyography (EMG), and electrooculography (EOG). Four factors measured from sleep status are then used to objectively evaluate sleep quality: sleep latency, wakefulness after sleep onset, sleep duration, and sleep efficiency [2]. Sleep latency is the time it takes to fall asleep, and sleep efficiency is the percentage of time spent asleep in bed. Using EEG, EMG and EOG for long-term assessment of sleep quality is nevertheless impractical, as it requires obtrusive equipment and manual processing. An alternative may be to automatically estimate sleep status from more easily recorded physiological signals.

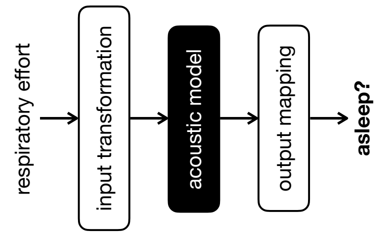
Respiratory patterns differ between wakefulness and sleep. During wakefulness, respiratory rate and depth are relatively regular, and muscles are tonically active. In non-rapid eye movement (NREM) sleep, respiratory rate and depth decrease, and muscle tone diminishes slightly [3]. During rapid eye movement (REM) sleep, respiration becomes faster and more erratic in comparison to wakefulness, and skeletal muscles, including the upper airway muscles and intercostals, become atonic. This increases upper airway resistance and reduces rib cage movement [4], leaving only the diaphragm and extraocular muscles active [5].

Exploiting these physiological differences, some studies have estimated sleep status from body and respiratory movements. Scott et al. [6] developed a signal-processing-based sleep tracking algorithm using the accelerometer signals from an actigraphy device worn on the index finger. Their algorithm was tested on data from 25 participants, and achieved a sensitivity of 0.91, a specificity of 0.59 and an accuracy of 0.85 when compared to ground-truth measurements from PSG. Chinoy et al. [7] compared the performance of consumer sleep-tracking devices with PSG on data from 34 users. Similar to the previous study, high sensitivity and low specificity were reported. Fitbit Alta HR obtained 0.95 sensitivity, 0.54 specificity and 0.90 accuracy. A sensitivity of 0.99, a specificity of 0.19 and an accuracy of 0.88 was attained by Garmin Vivosmart 3. One study is notable for the use of deep learning to track sleep at a short distance from the user. Dixon et al. [8] trained a convolutional neural network (CNN) to classify signals from the frequency-modulated continuous wave radar of Google Nest Hub. Their approach achieved 0.96 sensitivity, 0.55 specificity and 0.87 accuracy in comparison to PSG when tested on 33 subjects. All of the above studies demonstrated high sensitivity (>0.90) but low specificity (<0.60), indicating difficulty in distinguishing wakefulness from sleep. Balancing these metrics remains a key challenge in this task.

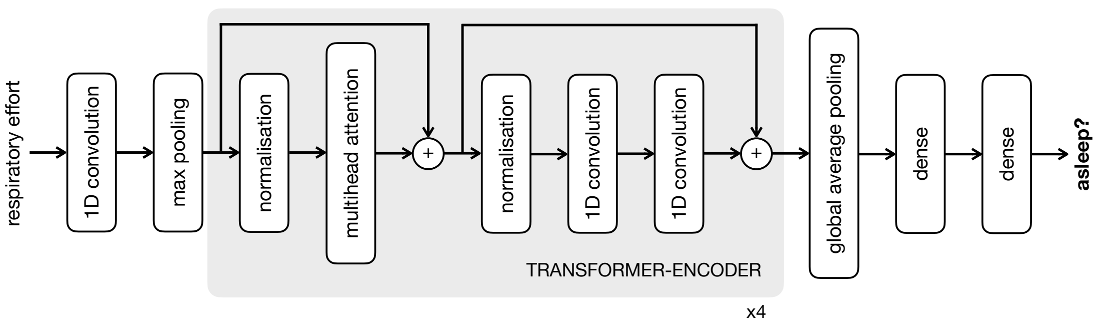
Here, we apply deep learning techniques to estimate sleep status from abdominal respiratory *effort* with the aim of investigating the feasibility of using abdominal respiratory *movement* as a proxy for assessing sleep quality. Unlike EEG, EMG and EOG, which require specialised equipment, abdominal respiratory movement – as a surrogate for abdominal respiratory effort – can be measured using simple and affordable sensors, such as an accelerometer placed on the abdomen. This makes respiratory movement monitoring suitable for home-based sleep studies potentially enabling long-term assessment of sleep quality. Given the lack of a respiratory movement dataset with sleep status annotations, this proof-of-concept study employs respiratory effort, since it is routinely recorded during PSG and, consequently, readily available in open-access PSG databases.



(a) Convolutional neural network (CNN)



(b) Reprogramming an acoustic model



(c) Transformer-encoder

Fig. 1: Deep neural network architectures for sleep status estimation from abdominal respiratory effort.

II. Data

We used the Montreal Archive of Sleep Studies (MASS) [9], an open-access database of PSG recordings from 200 participants (97 men and 103 women, aged 18–76 years). We specifically used the first subset of MASS, as it is the only one that includes sleep stage annotations. It consists of 476 hours of data from 53 subjects (34 men and 19 women, aged 55–76 years). The PSG recordings included 17- or 19-electrode EEG, 2-channel EOG, 5-channel bipolar EMG, 1-channel electrocardiography (ECG), 1-channel respiratory thermistance, 1-channel airflow, 1-channel oximetry, and 1- or 2-channel respiratory effort (abdominal and thoracic). Sleep stages were manually scored according to the American Academy of Sleep Medicine (AASM) guidelines [10], with wake, REM, N1, N2, and N3 stages annotated in non-overlapping 30-second epochs. Stages N1 to N3 are NREM sleep.

III. System description

This study investigates different deep learning architectures and state-of-the-art approaches to estimate sleep status from the abdominal respiratory effort signal by leveraging the physiological differences in respiratory patterns between wakefulness and sleep. Three end-to-end methods are considered: CNN, transformer, and reprogramming of acoustic models.

A. Convolutional neural network (CNN)

Since the abdominal respiratory effort signals in MASS are sampled at either 128 or 256 Hz, they are first resampled to 128 Hz if needed. Following standard sleep staging practice, the signals are divided into non-overlapping 30-second segments, and then normalised to zero mean and unit variance. Every segment is labeled as either wake or sleep according to the manual scoring of sleep stages. Stages REM, N1, N2 and N3 are collapsed into the sleep label. The 30-second segments are provided as input to a CNN system similar to one proposed in our previous study [11]. As shown in Fig. 1a, it consists of three 1-dimensional convolutional layers of 64, 128, and 256 filters with a kernel size of 1×8. A batch normalisation layer, and a 1×8 max-pooling layer follow each convolutional layer. A dropout rate of 0.5, and ReLU activation are used by all convolutional layers. The output of the convolutional layers is flattened and passed to a dense layer of 256 ReLU activation units. Finally, a dense layer of 1 sigmoid unit carries out the classification. The network has 0.7 million parameters. As most of the segments are labeled as sleep, class weights and bias initialisation are implemented to deal with the unbalanced amount of data for each class during training. Class weights adjust the loss function, prompting the model to focus more on samples from the under-represented class. Similarly, initialising the final layer's bias to reflect the data distribution mitigates bias towards the majority class. The CNN converged within 50 epochs using the Adam optimiser, binary cross-entropy as the loss function, and a learning rate of 0.001.

B. Transformer

The transformer architecture is regarded as the state of the art in sequence modelling [12]. It is based on an attention mechanism, which allows the model to learn long-range dependencies in sequential data [13], such as the abdominal respiratory effort signal. Transformers are also more efficient to train than recurrent neural networks (RNNs), since they can be parallelised more easily [14].

An overview of the proposed transformer architecture for sleep status estimation is provided in Fig. 1c. This is based on a transformer model developed by Keras that has proven successful for time series classification tasks [15]. It consists of one 1-dimensional convolutional layer with 16 filters and ReLU activation, a max-pooling layer, and 4 identical transformer-encoder blocks. Each block has an attention layer with 4 heads and a key dimension of 256. The output of the multihead attention layer is then projected by a pair of 1-dimensional convolutional layers with a kernel size of 1×1. The first one has 4 filters with ReLU activation, and the second one, 1 filter with linear activation. There are 2 residual connections in each block. Lastly, the projection is passed through a global average pooling layer, and 2 dense layers perform the classification. The dense layers consist of 128 ReLU units and 1 sigmoid unit, respectively. This network has 0.3 million parameters, and its input is the same as that of the CNN system. The transformer converged within 100 epochs using a learning rate of 0.0001, the Adam optimiser, and binary cross-entropy as the loss function.

C. Reprogramming acoustic models

Although publicly available foundation models exist for audio (e.g., OpenAI Whisper [16]), text (e.g., Google PaLM 2 [17]), and images (e.g., Meta AI DINOv2 [18]), there are none for physiological signals or, more generally, for time series data. To overcome this issue, Yang et al. [19] propose *reprogramming* acoustic models for time series classification. Their approach is motivated by the observations that modern acoustic models are a mature technology trained on very large datasets, and audio is a univariate temporal signal. Therefore it is likely that an acoustic model can be reprogrammed as a robust feature extractor for time series tasks. Specifically, model reprogramming consists of three components: (1) a trainable universal input transformation function, (2) a large-scale pre-trained acoustic model, and (3) an output mapping function. Unlike transfer learning, the acoustic model is not updated during training. In the study by Yang et al., this approach outperformed or matched the performance of state-of-the-art methods on 20 of 30 time series tasks, such as ECG classification.

Here, we explore the potential of Google VGGish for estimating sleep status using the framework developed by Yang et al. [19]. VGGish is a 73.2 million-parameter acoustic model for audio event recognition, trained on about 6,000 hours of manually annotated data [20]. As depicted in Fig. 1b, the abdominal respiratory effort signal is provided to the input transformation layer, which learns a universal function that transforms the physiological signal into an acoustic-like signal as expected by the pre-trained model. Then VGGish takes as input the transformed effort signal, and outputs a 128-dimensional vector that is mapped to two classes: wake and sleep. 64 of the original 'classes' are assigned to the new wake class, and the remaining 64, to the new sleep class. All systems were developed using TensorFlow [21].

D. Viterbi decoding

To exploit information about the temporal sequence of sleep and wake states, transition probabilities between wake and sleep states were computed from the first subset of MASS. They are detailed in Table I. For instance, the probability of going from wake to sleep is 0.11. Transition probabilities are applied to the sequence of predictions made by the classifiers using the Viterbi algorithm [22] to find the most likely sequence of states. In this way, the final output of the sleep estimation systems is based on both local (i.e., the prediction for each 30-second segment) and global (i.e., transition probabilities) information.

TABLE I: Transition probabilities between wake/sleep states.

		Next state	
		Wake	**Sleep**
Current state	**Wake**	0.89	0.11
	Sleep	0.05	0.95

IV. Evaluation

Experiments were performed on the first subset of MASS (see Section II). A 10-fold cross-validation approach was used – 5 participants per fold with the exception of the final fold, which had 8. In each round of cross-validation, 8 folds were used for training; one fold, for validation; and the remaining fold, for testing. Epoch-to-epoch accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the curve (AUC) were calculated from the pooled results for all the systems proposed, using the labels from the PSG recordings as the ground truth.

V. Results and discussion

Table II compares the epoch-to-epoch accuracy, sensitivity, specificity, PPV, NPV, and AUC of the developed sleep status estimation systems to the performance reported by previous studies. Our CNN and transformer systems performed very similarly, with sensitivity and specificity above 0.82. This demonstrates that sleep status can be estimated from the abdominal respiratory effort signal even though it is not routinely used for that purpose, and suggests that the actions we took to deal with highly unbalanced data during training were effective. Contrary to our expectations, the transformer did not outperform the CNN. This is likely due to the fact that transformers generally require more training data than CNNs [23], [24]. However, it is also possible that capturing local spatial patterns is more relevant for the task at hand than capturing global dependencies. That is, for this task the characteristics of individual breaths may be more distinctive than the overall breathing pattern in a 30-second segment.

The CNN and transformer systems outperformed the reprogrammed VGGish, which had a lower epoch-to-epoch accuracy (0.76) and specificity (0.71). Nonetheless, the VGGish-based system highlights the potential of reprogramming pre-trained acoustic models for time series tasks that lack a readily available foundation model. This approach is likely to benefit

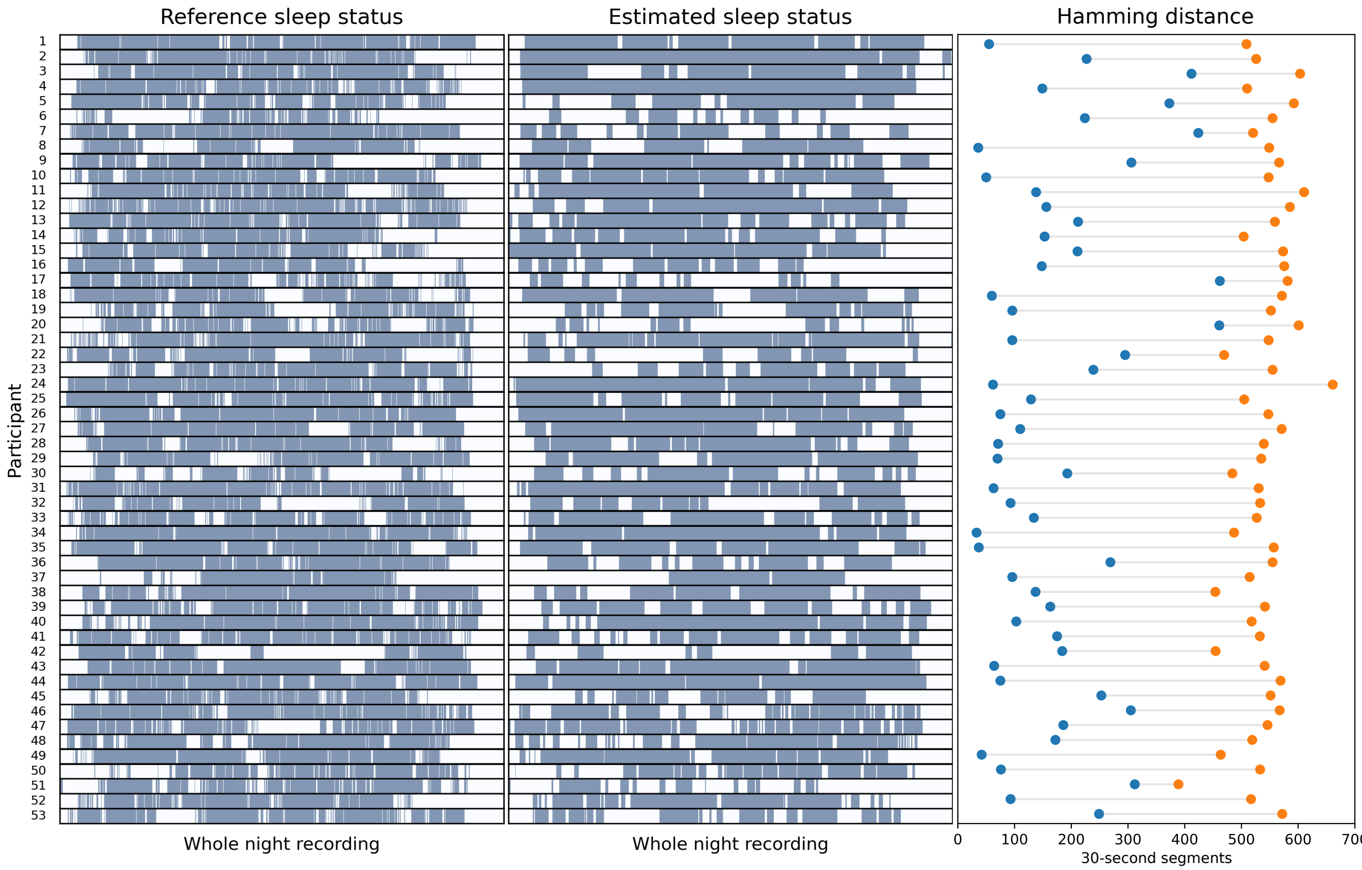


Fig. 2: Whole night reference sleep status, estimate from abdominal respiratory effort using the proposed CNN system, and Hamming distance for each participant. In the first two panels, white regions correspond to wakefulness, and blue regions, to sleep. In the third panel, blue dots indicate the Hamming distance (in segments or epochs) between the estimate and the reference, and orange dots, the expected random performance. Participants' IDs are shown on the y-axis.

TABLE II: Performance of the proposed sleep status estimation systems and related studies with respect to PSG.

	Acc.	Sen.	Spe.	PPV	NPV	AUC
CNN	0.84	0.85	0.82	0.91	0.72	0.90
Transformer	0.83	0.83	0.82	0.91	0.69	0.89
VGGish	0.76	0.82	0.71	0.75	0.78	0.70
Scott et al. [6]	0.85	0.91	0.59	–	–	–
Fitbit Alta [7]	0.90	0.95	0.54	0.94	0.58	–
G. Vivosmart [7]	0.88	0.99	0.19	0.89	0.74	–
Dixon et al. [8]	0.87	0.96	0.55	0.88	0.86	–

from more robust acoustic models such as OpenAI Whisper. However, deploying our CNN or transformer system on mobile devices is more feasible than deploying a reprogrammed foundation model, given its size and the compute power required.

While our systems and those of related studies cannot be directly compared due to differences in the data used, both were evaluated against overnight PSG, the clinical standard for sleep-disordered breathing diagnosis [25]. Related studies have reported higher sensitivity ($\geq$0.91), and lower specificity ($\leq$0.59), whereas our approach has a more balanced performance across both metrics. One potential reason for this is that their methods are mainly based on body movement rather than respiratory effort. The latter displays more distinctive physiological patterns between wakefulness and sleep, as discussed in Section I. This suggests that sleep quality can be accurately assessed with an approach based on respiratory movement as a surrogate for respiratory effort.

Fig. 2 displays the reference sleep status, the estimate derived from abdominal respiratory effort, and the Hamming distance between the estimate and the reference for each participant in the first subset of MASS. The reference was derived from manual sleep staging using EEG, EOG and EMG signals, whereas the estimate was obtained from the CNN system described here. The Hamming distance between two sequences of binary labels (i.e., sleep and wakefulness) is the number of positions at which their corresponding labels differ [26]. The proposed CNN system achieved better than random performance – approximated as 50% of the sequence length – across all nights in MASS. In most cases, the Hamming distance was also much lower than what would be expected by chance. This indicates that the network effectively learned to differentiate between wakefulness and sleep from the abdominal respiratory effort signal. For the nights with a large Hamming distance – for example, participant 17's data – the CNN underestimated the number of sleep segments. That is, the system produced many false negatives. An inspection of the data revealed no notable differences from other nights. By providing the CNN with more examples of wakefulness, it can be trained to better recognise the subtle patterns that differentiate wakefulness from sleep, potentially leading to a reduction in misclassifications. For participant 8's data, one of the nights with the smallest Hamming distance, a few isolated sleep and wakefulness epochs were missed or smoothed out by our system. However, the overall pattern was correctly predicted, including sleep onset and offset. This suggests that our system performs well on nights with disrupted sleep, and further supports the idea of using a respiratory movement-based method to evaluate sleep quality.

VI. Conclusions

Objective long-term monitoring of sleep quality with gold-standard equipment (i.e., EEG, EOG and EMG) is impractical due to its high cost and obtrusiveness. This study introduced a novel deep learning approach to estimate sleep status by exploiting the physiological differences in respiratory effort between wakefulness and sleep. Evaluated on data from over 50 participants against the clinical standard, our systems showed a better balance among sensitivity and specificity than previous studies. In the future, we will extend the technology developed here to an approach based on respiratory movement as a surrogate for respiratory effort. This has the advantage that the former can be measured using low-cost hardware, such as an accelerometer attached to the abdomen. We will also investigate the feasibility of estimating specific sleep stages (e.g., NREM and REM) from the respiratory effort signal.

Acknowledgment

This work was supported by Innovate UK Knowledge Transfer Partnership (KTP) 012741.

References

[1] V. K. Chattu, M. D. Manzar, S. Kumary, D. Burman, D. W. Spence, and S. R. Pandi-Perumal, "The global problem of insufficient sleep and its serious public health implications," *Healthcare*, vol. 7, no. 1, pp. 1–16, 2018.

[2] K. L. Nelson, J. E. Davis, and C. F. Corbett, "Sleep quality: An evolutionary concept analysis," *Nursing Forum*, vol. 57, no. 1, pp. 144–151, 2022.

[3] A. C. Guyton and J. E. Hall, *Textbook of Medical Physiology*. Elsevier Saunders, 2006, ch. States of Brain Activity: Sleep, Brain Waves, Epilepsy, Psychoses, pp. 739–747.

[4] H. R. Colten and B. M. Altevogt, Eds., *Sleep Disorders and Sleep Deprivation: an Unmet Public Health Problem*. The National Academies Press, 2006, no. 2, ch. Sleep Physiology, pp. 33–53.

[5] S. Singh, N. Yadav, and S. Jain, "Sleep and Health: an Introduction," *International Journal of Head and Neck Surgery*, vol. 10, no. 1, pp. 1–3, 2019.

[6] H. Scott, N. Lovato, and L. Lack, "The Development and Accuracy of the THIM Wearable Device for Estimating Sleep and Wakefulness," *Nature and Science of Sleep*, vol. 13, pp. 39–53, 2021.

[7] E. D. Chinoy, J. A. Cuellar, K. E. Huwa, J. T. Jameson, C. H. Watson, S. C. Bessman, D. A. Hirsch, A. D. Cooper, S. P. A. Drummond, and R. R. Markwald, "Performance of seven consumer sleep-tracking devices compared with polysomnography," *Sleep*, vol. 44, no. 5, pp. 1–16, 2021.

[8] M. Dixon, L. Schneider, J. Yu, J. Hsu, A. Pathak, D. Shin, R. S. Lee, M. R. Malhotra, K. Mixter, M. McConnell, J. Taylor, and S. Patel, "Sleep-wake detection with a contactless, bedside radar sleep sensing system," *Google*, 2021.

[9] C. O'Reilly, N. Gosselin, J. Carrier, and T. Nielsen, "Montreal Archive of Sleep Studies: an Open-access Resource for Instrument Benchmarking and Exploratory Research," *Journal of Sleep Research*, vol. 23, no. 6, pp. 628–635, 2014.

[10] American Academy of Sleep Medicine, *AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications (Version 2.6)*, January 2020.

[11] H. E. Romero, N. Ma, G. J. Brown, and S. Johnson, "Obstructive sleep apnea screening with breathing sounds and respiratory effort: a multimodal deep learning approach," in *Proceedings of INTERSPEECH 2023*, Dublin, Ireland, 2023, pp. 5451–5455.

[12] Q. Fournier, G. M. Caron, and D. Aloise, "A Practical Survey on Faster and Lighter Transformers," *ACM Computing Surveys*, vol. 55, no. 14s, pp. 1–14, 2023.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[14] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 119, 2020, pp. 5156–5165.

[15] T. Ntakouris, "Time series classification with a transformer model," https://keras.io/examples/timeseries/timeseries_classification_transformer/, *Keras*, 2021.

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202, 2023, pp. 28 492–28 518.

[17] R. Anil et al., "PaLM 2 Technical Report," *Google*, 2023.

[18] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P. Y. Huang, S. W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning Robust Visual Features without Supervision," *Meta AI*, 2023.

[19] C. H. H. Yang, Y. Y. Tsai, and P. Y. Chen, "Voice2Series: Reprogramming Acoustic Models for Time Series Classification," in *Proceedings of the 38th International Conference on Machine Learning*, 2021, pp. 1–12.

[20] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 131–135.

[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] A. J. Viterbi, "A personal history of the Viterbi algorithm," *IEEE Signal Processing Magazine*, vol. 23, no. 4, pp. 120–142, 2006.

[23] L. Deininger, B. Stimpel, A. Yuce, S. Abbasi-Sureshjani, S. Schönenberger, P. Ocampo, K. Korski, and F. Gaire, "A comparative study between vision transformers and CNNs in digital pathology," in *Proceedings of the IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR) 2022*, New Orleans, United States, 2022, pp. 1–8.

[24] S. Ahmed, I. E. Nielsen, A. Tripathi, S. Siddiqui, R. P. Ramachandran, and G. Rasool, "Transformers in Time-Series Analysis: A Tutorial," *Circuits, Systems, and Signal Processing*, vol. 42, pp. 7433–7466, 2023.

[25] K. Murase and O. D. Lyons, "Impact of the COVID-19 pandemic on diagnosis of sleep apnea: an observational study of a hybrid virtual care clinical pathway," *Sleep and Biological Rhythms*, vol. 21, pp. 309–317, 2023.

[26] R. Hamming, "Classical Error-Correcting Codes: Machines should work. People should think." in *Classical and Quantum Information*, D. C. Marinescu and G. M. Marinescu, Eds. Boston, United States: Academic Press, 2012, no. 4, pp. 345–454.