2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

SomnNET: An SpO2 Based Deep Learning Network for Sleep Apnea Detection in Smartwatches

Arlene John[1], Koushik Kumar Nundy[2], Barry Cardiff[3] and Deepu John[3]

Abstract— **The abnormal pause or rate reduction in breathing is known as the sleep-apnea hypopnea syndrome and affects the quality of sleep of an individual. A novel method for the detection of sleep apnea events (pause in breathing) from peripheral oxygen saturation (SpO2) signals obtained from wearable devices is discussed in this paper. The paper details an apnea detection algorithm of a very high resolution on a per-second basis for which a 1-dimensional convolutional neural network- which we termed SomnNET- is developed. This network exhibits an accuracy of 97.08% and outperforms several lower resolution state-of-the-art apnea detection methods. The feasibility of model pruning and binarization to reduce the computational complexity is explored. The pruned network with 80% sparsity exhibited an accuracy of 89.75%, and the binarized network exhibited an accuracy of 68.22%. The performance of the proposed networks is compared against several state-of-the-art algorithms.**

Keywords—**Sleep apnea detection, Peripheral Oxygen Saturation, Convolutional neural networks**

I. INTRODUCTION

An increase in healthcare and diagnosis costs, along with an aging populace, has placed tremendous pressure on our healthcare systems. Automatic diagnosis/ detection of diseases through artificial intelligence (AI) techniques in wearable devices are seen as a possible solution to tackle this issue [1], [2]. In many cases, these AI algorithms are integrated into the wearable device itself to reduce channel bandwidth usage [3], [4]. The abnormal reduction or pause in breathing during sleeping, associated with a reduction in blood oxygen levels is known as the sleep apnea-hypopnea syndrome [5]. A complete pause in breathing is termed as apnea, and a temporary reduction in respiration rate indicated by a drop in oxygen saturation for a minimum of 10 seconds is termed as hypopnea [6]. The diagnosis of sleep-related disorders is traditionally done through overnight polysomnography under the supervision of a clinician. Different sensors are attached to the patient's body to obtain signals for later analysis by sleep experts during polysomnography for a final diagnosis [6]. Recording polysomnograms are costly and are not comfortable for patients. Therefore, the development of an automatic sleep-apnea detection method that is easily available, non-intrusive, and readily accessible is of paramount importance.

Peripheral oxygen saturation (SpO2) is an estimation of the oxygen saturation level in the blood usually measured with a pulse oximeter device, which is non-invasive and is usually found in smartwatches. Methods for analyzing obstructive sleep apnea events from pulse oximeter data were reviewed in [7]. Various methods for sleep apnea detection based on deep learning were reviewed in [8]. In literature, most sleep-apnea detection algorithms using deep learning methods exhibit a resolution of 1 minute ie., inferences are obtained on a per-minute basis [6], [9], [10]. Moreover, SpO2 based sleep apnea detection algorithms are often accompanied by other signals to improve inferences [5], [6]. The highest resolution of sleep apnea detection from SpO2 signals (in combination with other signals) was explored in [5] with a resolution of 1 second and with a performance accuracy of 79.61%, which leads to the novelty of this work:

1. Development of a per-second sleep apnea detection algorithm from single-lead peripheral oxygen saturation data which is suitable for deployment in smartwatches and with a resolution higher than most state-of-the-art methods.
2. Development of a 1-dimensional convolutional neural network (1D-CNN) for this task, which we termed SomnNET (for *somnum* net), which reduces the need for feature extraction stages and the task of identifying features useful for apnea detection.
3. Complexity analysis of the developed network and complexity optimization using network pruning methods and binarization[1].

II. METHODOLOGY

A. Method Outline

A method for sleep apnea detection on a per-second basis is proposed in this article. The data preparation for per-second apnea detection using peripheral oxygen saturation signals is carried out as discussed in [11]. A sample of the training data contains a single signal window of 11 seconds with the 2nd second corresponding to the event being detected (an apnea or non-apnea event) as discussed in [11]. Windows that contained SpO2 values less than 50% were considered artifacts and were dropped from the dataset.

[1]Arlene John is with University College Dublin, Ireland, Email: arlene.john@ucdconnect.ie

[2]Koushik Kumar Nundy is with Think Biosolution, Email: kknundy@thinkbiosolution.com

[3]Barry Cardiff and Deepu John are with University College Dublin, Ireland, Email: {barry.cardiff, deepu.john}@ucd.ie

This work was supported in part by the Irish Research Council under the New Foundations Scheme; and in part by the Microelectronic Circuits Centre Ireland under Grant MCCI-2018-03.

[1]Code available at https://github.com/arlenejohn/CNN_Sleep_apnea_SpO2.

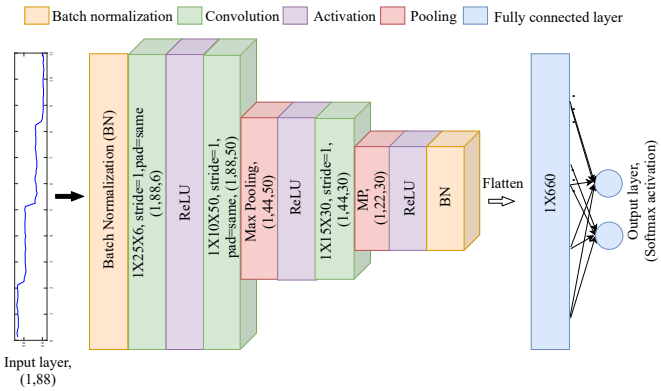

Fig. 1: The architecture of SomnNET which we propose for per-second sleep apnea detection.

B. Dataset

In this article, the UCD St. Vincent's University Hospital's sleep apnea database containing polysomnogram records from 25 patients with a duration of 6-8 hours and with annotations for every second is used [12]. We use the peripheral oxygen saturation (SpO2) signals which were acquired at a sampling rate of 8 Hz for this network. Fig. 2 shows an example of signal windows where the patient is apneic and is non-apneic (record ucddb002), and it can be observed how the SpO2 levels are lower in the window depicting the apneic event. Training, validation, and test set splits are carried out epoch-wise, in the ratio of 8:1:1. Oversampling of the minority apneic class was carried out to balance the training and validation sets. Patient records without any apnea events (ucddb008, ucddb011, ucddb013, and ucddb018) were discarded.

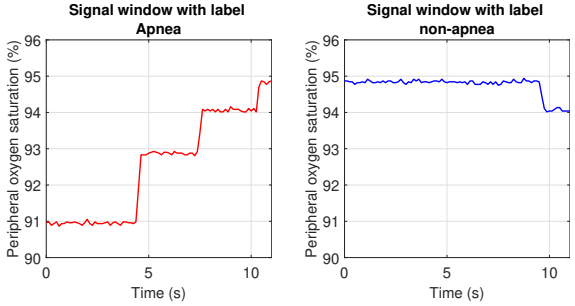

Fig. 2: (Left) SpO2 signal window with label apneic and (Right) SpO2 signal window with label non-apneic.

C. 1D-CNN

SomnNET has 88 (8 Hz sampling frequency) input nodes, and a batch-normalization stage follows the input layer. The CNN architecture uses three convolution layers with filter lengths 25, 10, and 15, with 6, 50, and 30 of them in each layer respectively. The strides of all the convolutional layers were maintained as 1, and zero padding was used. The convolutional neural network architecture is as described in Fig. 1. The network parameters and training parameters are:

1) Dropout in fully connected layer- probability $= 0.25$,
2) Activation function in output layer- Softmax,
3) Activation functions- ReLU,
4) Optimizer- ADAM,
5) Loss function- Binary cross-entropy, and
6) Regularizer- L2 regularizer in output layer.

The network generated has a total of 27,182 parameters, and for consistency throughout the paper, we refer to this network as Model 1.

D. Pruning

The systematic removal of parameters from an existing network is a popular approach for reducing resource requirements at prediction time and is referred to as pruning [13]. Here, magnitude-based weight pruning is used, which gradually zero out model weights during the training process. This enables compression of the model and is suitable for deployment in resource-constrained devices like smartwatches. We use SomnNET and attempt to sparsify the network with sparsity varying from 10% to 80%, and the pruned network is referred to as Model 2.

E. Binarized CNN

Courbariaux *et al.* proposed an algorithm to develop neural networks with binary weights and activations [14]. Binarized kernel elements and weights substantially improve power efficiency by reducing memory size and accesses. In this paper, the kernel parameters in the convolution layers and the weights in the fully connected layers of SomnNET are binarized. Since the activation functions used are ReLU, we do not attempt to binarize the activation functions. This network has 27,094 parameters as the bias terms in Model 1 are eliminated. The binarized SomnNET is referred to as Model 3.

III. Results

A. Performance Analysis

A method of training over the full dataset and simultaneous validation on the validation set for each epoch was used for generating SomnNET. A validation callback was carried out to track the set of weights that exhibited the highest validation accuracy during training, and these weights were chosen as the final network weights. Model 1 exhibited an accuracy of 97.08%, a specificity of 97.42%, and a sensitivity of 84.65% on the test set. The performance is found to outperform other state-of-the-art methods, which could be attributed to a suitable 1D CNN network that was developed along with the regularization methods used, and the validation callback ensuring that the final weights are not based on the training data, thereby preventing overfitting. The performance parameters are detailed in Table I.

The performance of the pruned networks when the sparsity of SomnNET is at 10% to 80% is observed. In the pruned networks, the complexity is reduced by increasing the sparsity to the desired levels in the convolutional layers and the fully connected layer. The performance parameters of the pruned networks for different sparsity levels are shown in

Fig. 3. The performance drops with an increase in sparsity levels as is expected, and therefore an optimal sparsity level can be achieved by deciding a trade-off between power/resource consumption and performance. However, it can be observed that accuracy and specificity increase when the network is sparsified by 10%, but it is accompanied by a significant drop in sensitivity. Even though accuracy increases, the weights that make the network sensitive are pruned away when sparsity is increased to 10% here. We also observe that sensitivity increases when sparsity is increased to 70% and 80%. It may appear that even though overall performance drops with a drop in accuracy, the weights that make the network highly specific are pruned away with an increase in sparsity, making the network more sensitive. The pruned SomnNET exhibited an accuracy of 89.75%, a specificity of 90.19%, and a sensitivity of 73.39% on the test set consisting of data from all the patients at 80% network sparsity.

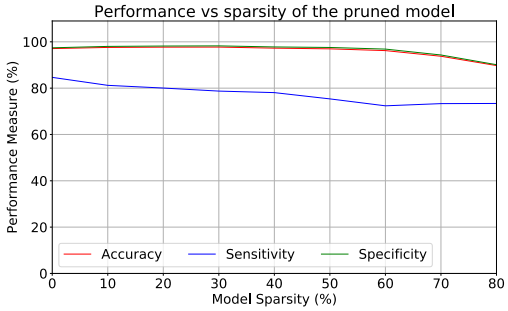

Fig. 3: The performance parameters (Accuracy, Sensitivity, and Specificity) of the pruned network on the test set when the sparsity of Model 2 is increased from 10% to 80%.

The binarized SomnNET model was trained in a similar manner as Model 1. This network exhibited an accuracy of 68.22%, a specificity of 67.94%, and a sensitivity of 78.44% on the test set. The performance of the binarized network is poor when compared to Model 1 and Model 2. When compared to Model 2 (at 80%) in terms of sensitivity alone, Model 3 performs better at the cost of accuracy. An investigation into various combinations of binarized layers and non-binarized layers is required to achieve the desired performance levels at low computational complexity.

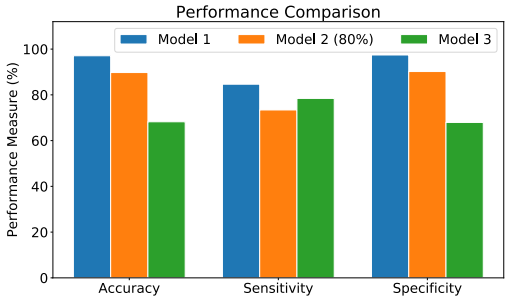

Fig. 4: Bar plots of Accuracy, Specificity, and Sensitivity of the three networks Model 1, Model 2, and Model 3 for comparison on the test set.

TABLE I: Performance of the three networks Model 1, Model 2, and Model 3 in terms of accuracy, sensitivity, and specificity

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)
Model 1	97.08	84.65	97.42
Model 2 (10%)	97.58	81.21	98.02
Model 2 (20%)	97.70	80.04	98.18
Model 2 (30%)	97.73	78.73	98.25
Model 2 (40%)	97.27	78.07	97.79
Model 2 (50%)	96.97	75.36	97.56
Model 2 (60%)	96.21	72.37	96.86
Model 2 (70%)	93.77	73.32	94.32
Model 2 (80%)	89.75	73.39	90.19
Model 3	68.22	78.44	67.94

TABLE II: Computational Complexity in terms of multiplication and addition operations and the energy consumption of the networks during prediction

Network	Params	Mul	Add	Energy (μJ)
Model 1	27182	1270016	1272876	0.4964
Model 2 (10%)	24485	1143428	1146288	0.4470
Model 2 (20%)	21788	818840	821700	0.3205
Model 2 (30%)	19091	889724	892584	0.3481
Model 2 (40%)	16394	762608	765468	0.2985
Model 2 (50%)	13697	636020	638880	0.2492
Model 2 (60%)	11000	508904	511764	0.1996
Model 2 (70%)	8303	382316	385176	0.1502
Model 2 (80%)	10106	255200	258060	0.1006
Model 3	27094	1496	1179946	0.0236

The performance of the various versions of SomnNET (Model 2 at 80%) in terms of accuracy, sensitivity, and specificity are compared in Fig. 4. The performance parameters of all the networks are detailed in Table I. From the table, it can be observed that Model 1 and Model 2 are suitable for per second apnea detection in smartwatches, while Model 3 requires further investigation.

B. Computational Complexity Analysis

The computational complexities of the three networks were calculated in terms of the number of multiplications and additions required for each second [11]. Detection of sleep apnea events with SomnNET (Model 1) requires 2260016 multiplications and 2262876 additions. In the case of Model 2, the gains due to pruning can be approximated by estimating the number of operations with non-zero numbers. When SomnNET is at 80% sparsity, the number of computations required involves 255200 multiplications and 258060 additions respectively. Detection of apnea events with the binarized SomnNET network requires just addition operations at the convolution layers and fully connected layer, since the weights are either +1 or -1, and therefore Model 3 requires 1496 multiplications and 2169946 additions. The total energy consumption during prediction is found to be 0.4964 μJ, 0.1006 μJ, and 0.0236 μJ for Model 1, Model 2, and Model 3 respectively. This is estimated by assuming that the energy required for a 16-bit multiplication accumulation (MAC) operation is 0.39 pJ [15], [16], and for a 16-bit adder is around 20 fJ [17] in 28nm FD-SOI technology. The complexity in terms of multiplications and additions and the corresponding energy consumption are discussed in Table II.

TABLE III: Comparison of state-of-the-art SpO2 record based sleep apnea detection algorithms with the performance of the proposed 1D-CNN based models for apnea detection.

Article	Mostafa *et al.* [9]	Almazayadeh *et al.* [10]	Xie *et al.* [6]	Cen *et al.* [5]	This work		
Dataset	Physionet Apnea ECG database and UCD St. Vincent's database	Physionet Apnea ECG database	UCD St. Vincent's database	UCD St. Vincent's database	UCD St. Vincent's database		
Signal	SpO2	SpO2	SpO2 and electrocardiogram	SpO2, oronasal airflow, ribcage and abdominal movement	SpO2		
Resolution	1 minute	1 minute	1 minute	1 second	1 second		
Method	3 layer deep belief network	Multilayer neural network	Bagging with RepTree	2D CNN	SomnNET		
					Model 1	Model 2 (80%)	Model 3
Accuracy	97.64 %	93.30 %	84.40 %	79.61%	97.08 %	89.75 %	68.22 %
Sensitivity	78.75 %	87.50 %	79.75 %	-	84.65 %	73.39 %	78.44 %

The performance of the proposed 1D-CNN networks is compared with that of state-of-the-art algorithms in Table III. It can be seen that Model 1 and Model 2 exhibit comparable performance in terms of accuracy but higher sensitivity to methods proposed in [9] which is tested on a combined database, and therefore a direct comparison to our method is not possible. Model 1 and Model 2 outperform the method proposed in [10] in terms of accuracy, which is tested on the Physionet Apnea Database. Model 1 and Model 2 outperforms the methods proposed in [6] and [5] in terms of accuracy and sensitivity. Direct comparison with [6] and [5] is possible due to the same dataset being used. It is also noteworthy that the method proposed in this work uses a single sensor source, SpO2, while in [6] and [5] a combination of different signals is used for inference, and yet our method outperforms these two works. This could be because in [6], bagging with repTree is used along with 1-minute resolution, which significantly reduces the number of training samples and due to the hand-engineered features. In the case of [5], a 2D CNN with a combination of different signals is used. The model complexity and larger number of learnable parameters could have caused the model to overfit to the training set, leading to poor performance on the test set.

IV. Conclusions

In this article, we analyze the feasibility of sleep apnea detection from SpO2 signals on a per-second basis. The requisite features for apnea event detection are learned by the proposed 1D-CNN network, which is termed as SomnNET. Two strategies to reduce the network size to make it suitable for implementation in smartwatches are analyzed, and the performance of these two methods is studied. The proposed SomnNET network achieved an accuracy of 97.08%, the pruned SomnNET network at 80% sparsity achieved an accuracy of 89.75%, and an accuracy of 68.22% was exhibited by the binarized SomnNET. These networks can work at a high resolution (per-second) for any subject with minimal tuning and is suitable for implementation on smartwatches. In future works, the filters learned by the 1D-CNN layers of SomnNET can be analyzed to explain the feature extraction process. The correlation with human understanding of requisite filter banks for SpO2 based apnea detection with the filter weights learned by the CNN can be analyzed to understand its impact on the model performance.

References

[1] L. P. Malasinghe, N. Ramzan, and K. Dahal, "Remote patient monitoring: a comprehensive study," *J Ambient Intell Human Comput*, vol. 10, pp. 57–76, 2019.

[2] D. L. T. Wong *et al.*, "An integrated wearable wireless vital signs biosensor for continuous inpatient monitoring," *IEEE Sensors Journal*, vol. 20, no. 1, pp. 448–462, 2020.

[3] A. John *et al.*, "An approximate binary classifier for data integrity assessment in IoT sensors," in *2020 27th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, 2020, pp. 1–4.

[4] A. John *et al.*, "Binary classifiers for data integrity detection in wearable IoT edge devices," *IEEE Open Journal of Circuits and Systems*, vol. 1, pp. 88–99, 2020.

[5] L. Cen *et al.*, "Automatic system for obstructive sleep apnea events detection using convolutional neural network," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 3975–3978.

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

[7] P. I. Terrill, "A review of approaches for analysing obstructive sleep apnoea-related patterns in pulse oximetry data," *Respirology*, vol. 25, no. 5, pp. 475–485, 2020.

[8] S. S. Mostafa *et al.*, "A systematic review of detecting sleep apnea using deep learning," *Sensors*, vol. 19, no. 22, 2019.

[9] S. S. Mostafa *et al.*, "SpO2 based sleep apnea detection using deep learning," in *2017 IEEE 21st International Conference on Intelligent Engineering Systems (INES)*, 2017, pp. 091–096.

[10] L. Almazaydeh *et al.*, "A neural network system for detection of obstructive sleep apnea through spo2 signal features," *International Journal of Advanced Computer Science and Applications*, vol. 3, no. 5, 2012.

[11] A. John *et al.*, "A 1D-CNN based deep learning technique for sleep apnea detection in iot sensors," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021, pp. 1–5.

[12] "St. Vincents University Hospital/ University College Dublin sleep apnea database," 2011. [Online]. Available: http://physionet.org/physiobank/database/ucddb/

[13] D. Blalock *et al.*, "What is the state of neural network pruning?" 2020.

[14] M. Courbariaux *et al.*, "Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1," 2016.

[15] H. Reyserhove, N. Reynders, and W. Dehaene, "Ultra-low voltage datapath blocks in 28nm UTBB FD-SOI," in *2014 IEEE Asian Solid-State Circuits Conference (A-SSCC)*, Nov 2014, pp. 49–52.

[16] R. Taco *et al.*, "An 88-fJ/40-MHz [0.4 V]– 0.61-pJ/1-GHz [0.9 V] Dual-Mode Logic 8 $\times$ 8 bit Multiplier Accumulator With a Self-Adjustment Mechanism in 28-nm FD-SOI," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 2, pp. 560–568, Feb 2019.

[17] R. Taco *et al.*, "Evaluation of Dual Mode Logic in 28nm FD-SOI technology," in *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2017, pp. 1–4.