Dynamic Sparse Network in Sleep Apnea Time Series Classification

Wannes Vanwinsen (s1943251), Ruxandra Simioniuc (s3072347), Iris den Hertog (s2177005), Pascal Binnekamp (s1806882), Akram Sabik (s3095681)
University Of Twente
February 5, 2023

Abstract

Recent advances show that machine learning models can automatically detect obstructive sleep apnea episodes based on single-lead ECG data. This method is much more efficient and accessible as opposed to the traditionally used PSG tests which are time-consuming and expensive. While high classification accuracy has been achieved, there might be an opportunity in reducing the computational complexity of the architectures. In this paper, we propose using an existing Dynamic Sparse Network for time series classification to solve the same detection problem by reducing the computational complexity and maintaining state-of-the-art accuracy. Different kernel sizes and different levels of sparsities were tested. Furthermore, the DSN model was compared to its own dense version and two other dense networks. The performance of the models was tested by measuring the accuracy. The computational complexity was measuring the FLOPS and number of parameters. The optimal parameters were chosen to be a kernel size of 30, and a sparsity of 95%. With these parameters, the DSN model had an average accuracy of 91.82% $\pm 0.84\%$ which was higher than the two dense networks and had significantly lower computational costs. The DSN can be deemed adequate for detecting sleep apnea events, especially in the case of low computational resources.

1 Introduction

Obstructive Sleep Apnea (OSA) is the intermittent cessation of breathing while sleeping. It is estimated that nearly 936 million people aged between 30 and 69 suffer from mild to severe OSA [2]. OSA can lead to a variety of health implications such as high blood pressure, heart failure and strokes, which increase mortality rates [5]. Thus, it is of high relevance to detect and treat OSA. Polysomnography (PSG) tests are commonly used to detect OSA. Several electrophysiological signals are recorded during the patient's sleep, including nerve signals (electroencephalography - EEG), eye tracking (electrooculography - EOG), pulse rate and beat (electrocardiography - ECG), chest and stomach activity, jaw muscle tone, ankle motions (electromyography - EMG), and oxygen saturation (SpO2) [2]. To assess sleep apnea, clinicians must manually read the PSG data, giving them a big workload. While very accurate, such tests are time-consuming, expensive and require various types of specialised equipment.

Over the past decades, efforts have been made to simplify the detection of OSA with ECG alone [17]. This is motivated by the fact that ECG signals are widely available, cost-effective, non-invasive, and easy to acquire, making them a suitable alternative for apnea detection. Furthermore, it might be possible to perform this test at home, removing the need for overnight hospital stays and human supervision. As the changes caused by sleep apnea are especially subtle on the ECG, there is a need for machine learning models to detect apneustic episodes [17]. Various supervised learning methods have been proposed to automatically detect these from ECG signals [11]. High classification accuracy has been achieved through the use of deep neural network architectures such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [3]. The downside of using such networks is that they often come with a large computational cost due to their fully connected architectures.

A recently published model found that sparse network architectures can perform similarly if not superior to dense networks with lower computational cost in time series [20]. Currently, the available literature does not reflect the use of sparsity in network architectures in the context of OSA detection. Thus, the following research question will be addressed: **To what extent can a dynamic sparse network for time series classification yield state-of-the-art accuracy in classifying obstructive sleep apnea from single-lead one-minute ECG segments, while being less computationally expensive?**

2 Related work

2.1 Sleep Apnea detection based on ECG

Recent literature proposes the use of different CNN and RNN architectures for the detection of OSA from ECG signals [11, 3, 4]. The proposed methods in current literature have shown that OSA can already be detected at high accuracy [17], residual NNs obtaining 99% and support vector machines achieving even 100%.

In 2019, Erdenebayar et al. [6] achieved test accuracies of up to 99% on a test set of 17 subjects. ECG records were segmented into 10-second intervals. The used architectures were regular DNN, CNN as well as more complex architectures such as LSTM and GRU-based RNN's. However, these are all dense networks. Fatimah et al. [8] demonstrated that it is possible to detect OSA from ECG segments by using a Fourier decomposition method for feature extraction and Support Vector Machine with a Gaussian kernel for classification. The reported accuracy during evaluation reaches up to 92%. Ayatollahi et al. used transfer learning for the adaption of DCNNs to classify OSA based on ECG signals and reached accuracies up to 93.33 %. The authors mention that the models can be improved by decreasing the number of parameters to achieve lower complexity [3].

After analysing the available literature in the field of classifying sleep apnea, the majority of the research has been focused on improving the accuracy of the models, rather than reducing the number of parameters or computational complexity. This is more than sensible from a medical perspective as incorrect diagnosis can have dire consequences.

However, with the increasing demand for real-time, low-cost, and portable solutions for sleep apnea monitoring, reducing the computational complexity of the models becomes increasingly important. This is where sparse neural networks can play a significant role, as they have shown to provide high accuracy while still being computationally efficient.

2.2 Sparse Networks

Sparse neural networks aim to reduce the number of parameters in the network, by removing connections with low magnitude weights, which results in lower memory and computation costs. This makes the network more efficient and can allow it to be deployed on devices with limited resources (such as embedded systems) [12]. Moreover, the sparsity does not significantly impact the accuracy of the network in most cases, making it a trade-off between efficiency and performance [12, 15]. The reduction of the number of parameters in the network also helps prevent over fitting, as well as making the network easier to interpret and understand.

There are two main approaches to training a sparse neural network: pruning and training a sparse network from scratch. Pruning involves converting a trained dense network into a sparse one by removing connections with low-magnitude weights. This can be done after the dense network has been trained, but the resulting sparse network can only be as big as the largest trainable dense network.

On the other hand, training a sparse network from scratch involves finding and training a sparse network directly, without first training a dense network. This approach was hypothesized by Evci et al. in their paper "Rigging the lottery: Making all tickets winners." [7]. They argued that if a sparse network can be found through pruning, it is possible to train that sparse network from scratch and that this approach has the potential to be more effective than pruning.

The use of sparsity in supervised learning methods is not reflected in the current literature on OSA detection. Deep learning approaches either apply fully connected layers or other computationally heavy network architectures such as LSTM and GRU [3, 11]. While only slight improvement can be made in terms of classification accuracy, there is still a lot of room for improvement in terms of computational efficiency.

2.2.1 Dynamic Sparse Network

In 2022, Xiao et al. [20] presented a novel approach to apply dynamic sparsity in time series classification (TSC). Their proposed dynamic sparse network (DSN) model can achieve state-of-art performance on TSC data sets with less than 50% computational costs compared with recent methods.

This model contains CNN layers with large kernels but is dynamically sparse. To cover different receptive fields

(RF), this model can automatically learn sparse connections. The DSN is trained by dynamic sparse training (DST). As shown in figure 1, the CNN kernels in each layer are split into sparse groups. Each group can be explored under a constraint region during DST, ensuring that the model is able to capture variable RF. As the kernels are always sparse, the computational costs can be reduced. The architecture of the model is shown in figure 2

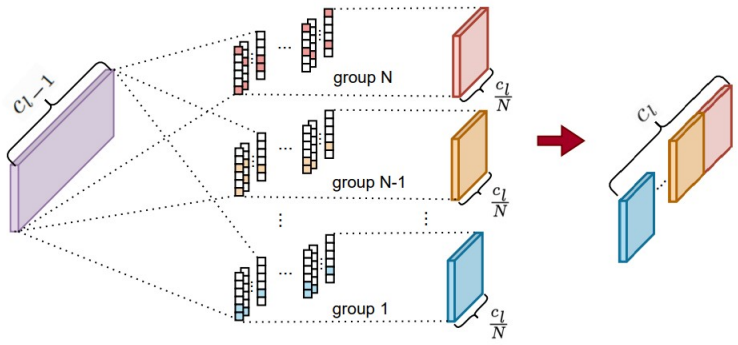


Figure 1: Dynamic sparse CNN layer. The kernels are split into groups with sparse connections [20]

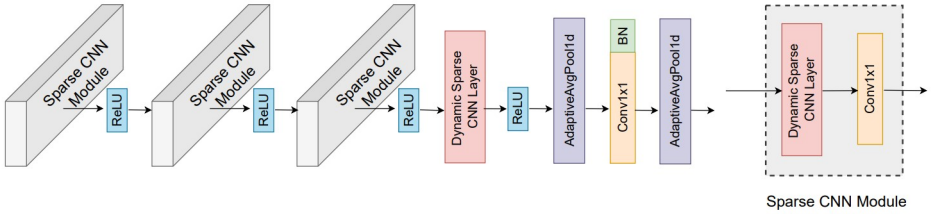


Figure 2: Architecture of the DSN model [20]

3 Dataset

This project is based on the PhysioNet challenge "Detecting and Quantifying Apnea Based on the ECG: The PhysioNet/Computing in Cardiology Challenge 2000" [9] and focuses on their second challenge: quantitative assessment of apnea. The aim is to classify each minute of ECG recordings as "Sleep Apnea" or "No Sleep Apnea". The challenge uses the Apnea-ECG Database[16], which consists of 35 labelled ECG records of sleeping patients. The first label of every subject corresponds to the first minute of ECG data [1]. The recordings range in duration from 7 to 10 hours each, giving a total of approximately 17000 minutes. The recordings include a discretized ECG signal, and a set of apnea annotations, indicating the presence or absence of sleep apnea in that minute. The ECG signals have been recorded at a sampling rate of 100 Hz with 16-bit resolution and 200 A/D units per mV.

Fig.3 shows two samples of 10 seconds extracted from two one-minute ECG signals belonging to each category along with their respective histogram. Nonetheless, a slight difference is visible with the naked eye for these 10-second intervals, but in order to correctly diagnose a patient with OSA, the number of minutes of sleep that are affected needs to surpass a certain threshold. Therefore, it is not only tedious but also extremely difficult, even for a professional, to successfully and accurately diagnose OSA based solely on ECG signals.

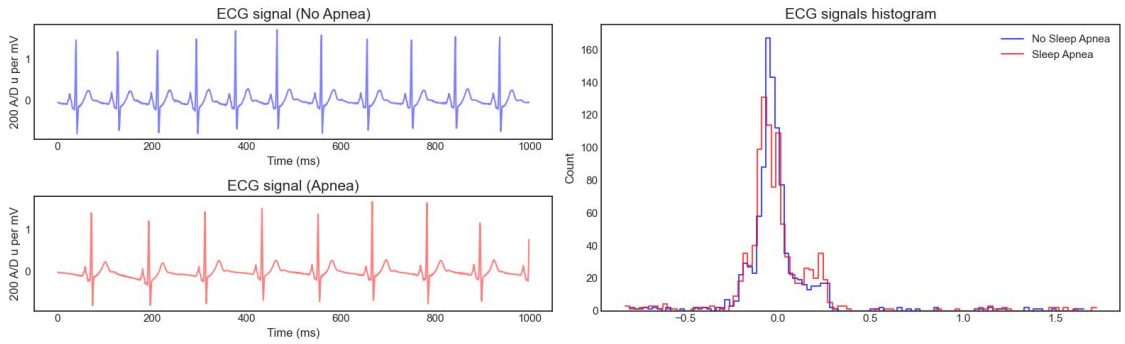


Figure 3: Left: 10 seconds ECG Signals with (upper graph) and without sleep apnea (lower graph); Right: Histogram of specific amplitudes in the Left ECG signals

4 Methodology

This section describes the methodology used to test the performance and computational complexity of the dynamic sparse neural network (DSN). First, data preparation is performed. Afterwards, the performance and computational cost of the DSN for apnea detection based solely on ECG are tested. This is compared to two dense networks.

4.1 Data preparation

The data preparation of the Apnea-ECG dataset consists of organizing the data and balancing classes. In its raw form, the data is available in *.apn* and *.dat* files, which contain the binary annotations for each minute and the actual ECG signal respectively. The labelled ECG data from all 35 subjects first needs to be transformed in order to fit the input required by the used models, thus the data is converted into an array format. The data of all recordings are then concatenated into one big array. Every row contains a label, followed by the corresponding minute of ECG data with 6000 data points. This results in a $n * 6001$ shaped array, with n being the total number of labelled minutes.

Classes are balanced to avoid a skewed model favouring the most frequently occurring class. Since there is an excess of minutes where sleep apnea is absent, 4000 samples are removed from this respective class are removed.

4.2 Testing performance and computational complexity of the DSN

In order to test the DSN for apnea detection based solely on ECG data, two comparisons are performed:

1. The first comparison aims to measure the influence of the DSN's hyperparameters on its own performance. The chosen parameters are the sparsity of the network and the kernel size. All other parameters are fixed.
2. The second comparison aims to measure the effects of the sparsity of the DSN by comparing it to two state-of-the-art dense networks and a dense version of itself.

Both comparisons are done based on the performance and computational complexity. The performance is quantified by computing the mean accuracy on 10 different test sets and the computational complexity by calculating the FLOP's and number of parameters used for the inference of a single sample.

In order to train and test the performance of the models, the balanced array is split into a train and test set with a ratio of 80/20. The experiments' results are averaged over 10 different splits per test to prevent bias that could be caused by a specific distribution over a train/test split. The same splits are used for all different tests to allow for a fair comparison.

4.2.1 Hyperparameters of the dynamic sparse neural network

For both the kernel size and sparsity hyperparameters of the DSN, different values are compared. For each value, the model is trained and tested on the 10 different train/test splits for 40 epochs.

Kernel Size
Kernel sizes between 6-48 are evaluated. The sparsity was set to a constant value of 80% for this experiment.

Sparsity
Sparsity is varied between a moderate sparsity of 60% as well as extreme sparsity of between 98% up A constant kernel size is used, which is chosen based on the results of the kernel size test and will therefore be mentioned in the results section.

4.2.2 Sparse vs dense network comparison

For the sparse vs dense network comparison, the sparse DSN is compared with three dense networks: an Omni-scale CNN (OS-CNN), a residual network (ResNet) and a dense version of the DSN. These state-of-the-art networks are chosen to obtain an image of how the DSN performs compared to the current state-of-the-art dense TSC networks.

The Omni-scale CNN is a dense neural network architecture designed for time-series classification, proposed by Tang et al. [19]. The innovation of this architecture lies in its universal rule for setting kernel sizes, which allows for an optimal receptive field (RF) sizes to be explored and utilized for improved performance. The authors claim that the OS-CNN provides comparable results to traditional convolutional neural networks without searching for the optimal RF size. The OS-CNN is also used as a comparison in the article by Xiao et al.[20], in which the DSN was presented. On the datasets that were used for testing in the paper, the OS-CNN was not always outperformed by the DSN, which makes it a relevant dense model to consider for the sleep apnea problem.

The ResNet model used was presented by Kachuee et al.[14] who created a ResNet model to classify five different arrhythmias based on ECG data. The application of ECG data makes the model suitable to use in this research. A ResNet model uses skip connections to deal with the vanishing gradient problem. This enables a network with more layers, suitable for more complex TSC tasks [13]. Similarly to the OS-CNN, ResNet was also one of the dense models used to compare and evaluate DSN's results.

It is difficult to set an objective measure for the 'best' kernel size and sparsity since this depends on the accuracy and computational complexity, which often times are competing. Therefore it is decided to choose the 'best' kernel size and sparsity by finding a trade-off between high accuracy and low computational cost. As a result, the kernel size that was found to best represent the trade-off between the two features is **30** for both the sparse and dense DSN, and the sparsity is set to 0 and 0.95 for the dense and sparse DSN respectively.

In order to achieve a fair comparison, the OS-CNN and ResNet are slightly adapted such that their architectures are as similar as possible to the architecture of the DSN. As a result, both models have 47 filters per layer and a stride of 1. The ResNet has the same number of blocks as the DSN: four blocks. For the OS-CNN, an architecture with 2 blocks is used. The same 10 train/test splits were used to evaluate all models.

5 Results

5.1 Hyperparameters of the dynamic sparse neural network

5.1.1 Evaluation of kernel size

Figure 4 depicts the effect of kernel size on predictive performance as well as the computational complexity of the model. The computational complexity grows linearly with the kernel size. The accuracy stabilizes for a kernel size of around 27-33. It was decided to keep the kernel size fixed at 30 during subsequent experiments.

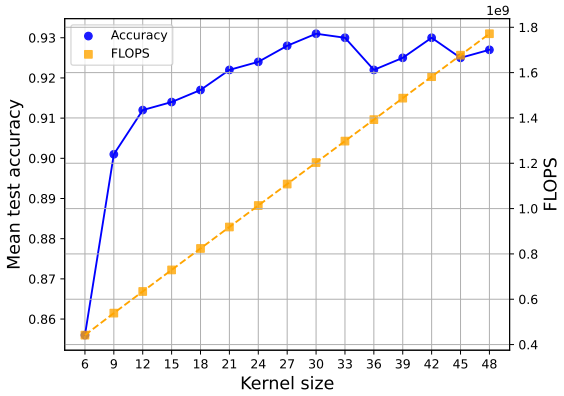


Figure 4: Mean accuracy and computational complexity (FLOPs)
for various kernel sizes (sparsity = 80%)

5.1.2 Evaluation of sparsity

The results for the different sparsities are shown in Figure 5. The computational complexity decreases linearly with increasing sparsity. The accuracy shows an overall decrease for increasing sparsity, with an accuracy of 92.03% for sparsity of 0.6 and an accuracy of 90.92% for extreme sparsity of 0.98. It was decided to use a DSN with a sparsity of 0.95 for the dense vs sparse comparison.

Figure 6 (see appendix A) gives a more detailed overview of the accuracy and error rates for both classes. The overall error rate of both apnea and non-apnea samples increases slightly when increasing the sparsity of the model. Furthermore, the model is less capable of correctly identifying apnea for sparser kernels.

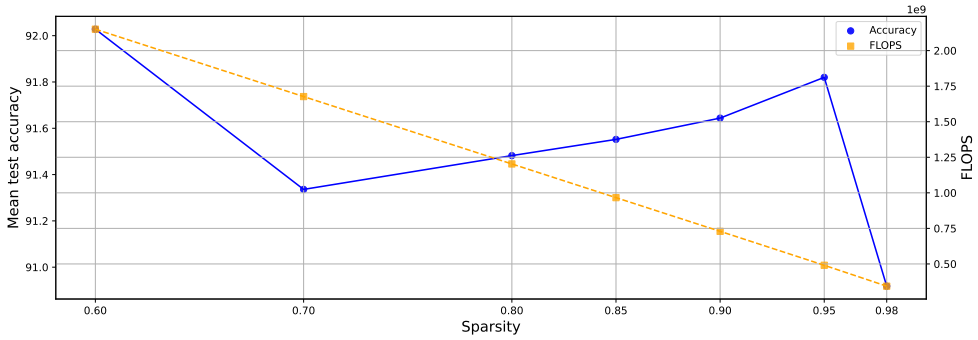


Figure 5: Mean accuracy and computational complexity (FLOPs)
for various sparsity's (kernel size = 30)

5.2 Sparse vs dense network comparison

	Mean accuracy	FLOPs (M)	# of parameters (K)
Dense DSN (S=0%)	92.18% ($\pm$0.75%)	7361.956	621.955
Sparse DSN (S=95%)	91.82% ($\pm$0.84%)	490.679	41.17
OS-CNN	88.84% ($\pm$1.03%)	2767.88	235.74
ResNet	81.48% ($\pm$1.11%)	104	578.751

Table 1: Results of the sparse vs dense comparison.

Table 1 shows the results of the comparison of the sparse DSN with 3 dense networks: the DSN with sparsity set to 0, OS-CNN and ResNet. The sparse DSN has a higher mean accuracy (91.82%) compared to the OS-CNN (88.84%) and ResNet (81.48%) and a slightly lower mean accuracy compared to the dense DSN (92.18%). The sparse DSN uses less FLOPS compared to the dense DSN and OS-CNN and more than the ResNet. Regarding the number of parameters, the sparse DSN has the lowest number of parameters of all four models.

6 Discussion and critical reflection

6.1 Interpretation of the results

6.1.1 Hyperparameters of the dynamic sparse neural network

Evaluation of kernel size
Accuracy on the test set appears to stabilize around 92%-93% for a kernel sizes of 21 and higher. Increasing the kernel size from this point only increases the computational cost while predictive performance remains the same. A kernel size of 30 seems to be a good trade-off between performance and computational complexity. It was decided to keep the kernel size fixed at 30 during the sparsity experiments with the DSN.
A limitation of the experiment is that the sparsity was fixed at 80% due to time restrictions. Ideally, the kernel size would be evaluated across different kernel sparsity's.

Evaluation of sparsity
The results for the sparsity show an overall decrease in accuracy with increasing sparsity, see figure 5 and 6. This decrease in accuracy was expected to be consistent over the tested sparsities, but the tests show an increase in accuracy with increasing sparsity from 0.7 to 0.95. Nevertheless, the differences in accuracy are relatively small (all in a range of 90.92% and 92.03%). The inconsistency can be explained by

1. A low number of train/test splits used during evaluation. Increasing this might result in a more consistent decrease in accuracy with increasing sparsity.
2. The pruning and re-growing of connections is dictated by some degree of randomness under different sparsity ratio's.

Furthermore, it is observed that the error rate of the classification increases together with the sparsity of the kernel. An explanation for this is that smaller, sparser receptive fields are less capable of capturing the small deviations between apnea and non-apnea samples than denser versions.

6.1.2 Sparse vs dense network comparison

Table 1 clearly shows that the DSN outperforms the other two dense networks in terms of accuracy and is also more consistent in its results, with a standard deviation of less than 1%. The DSN with 0% sparsity achieved a slightly better accuracy compared to the sparse DSN but at a much higher computational complexity, having over 15 times more parameters and FLOPs than its sparse counterpart. This highlights the effectiveness of the dynamic sparsity mechanism in improving computational efficiency while maintaining performance.

Regarding the computational complexity, it stands out that the ResNet requires the lowest FLOPS, but does have the highest number of parameters after the dense DSN. Since the ResNet is a dense network, it was hypothesized that it would require high FLOPS compared to the sparse DSN. This deviation might be caused by the different libraries used to calculate the FLOP's for different model implementations.

Overall, it can be concluded that the sparse DSN has the best balance between performance and computational complexity. If an improved computational complexity outweighs a small decrease in accuracy depends on the application. For apnea detection from ECG, it can be argued that a slightly lower accuracy is worth the improved computational complexity. For this particular disease, the diagnosis is set based on the number of apnea events per hour, called the apnea-hypopnea index (AHI)[10]. The scale for this index is comprised of four distinct grades: none or minimal SA (less than 5 events), mild SA (between 5 and 15 events per hour), moderate (from 15 to 30) and severe (more than 30 events)[10]. Thus, one misclassified apnea event influences the diagnosis minimally. Therefore a 1% higher or lower accuracy does not have high consequences for the diagnosis and consequential treatment and health of the

patient. In the context of different medical applications, this 1% could be much more significant, so each case should be analysed closely.

6.2 Future recommendations/improvements

Regarding the use of the dataset, better pre-processing steps could help achieve better performance. Some ECG signals in the dataset were noisy but still used for training the models which can negatively impact the classification task. Furthermore, 4000 minutes of ECG data were discarded in order to balance out classes. However, training and testing models with more data can help reach better results. Having balanced classes without reducing the size of the dataset could have been achieved with weighted metrics, weighted loss functions, or data augmentation.

Some minor improvements could enhance the training of the models. For example, the use of a leave-one-out split method could have prevented possible overfitting due to less difference between training and test data. This method is particularly relevant when the dataset is split between specific patients. During the sparsity test, the kernel size was only tested for an arbitrarily chosen sparsity of 80%. Different sparsity values could affect the obtained results for the kernel size experiment. Regarding the OS-CNN, the number of blocks was not equalized. Using the same number of blocks for the OS-CNN as for the other models would give a fairer comparison between the four models. Finally, it was noticed that reducing the depth of the DSN from 4 to 3 led to a slight decrease in performance on the test set (-2%) and also to a reduction of the number of FLOPS by a third, which indicates that this parameter is likely to be relevant to the research question and could be researched further.

Regarding other interesting methods that were not implemented in this paper, a technique that could have made the model easier to interpret is Grad-Cam [18]. This technique allows knowing which parts of the input data were most responsible for the final classification decision by taking advantage of the loss function's gradient. This might even enable the identification of sleep apnea on ECG signals using the naked eye, although it remains a difficult task.

7 Conclusions

Sleep apnea detection is not a new subject in the field of machine learning and remarkable results have been already obtained by using various methods. This paper tried to solve the same problem of detection by reducing the computational complexity of the architecture while still maintaining high accuracy values by using a dynamic sparse neural network. While DSN in particular showed that it can match and sometimes outperform its dense counterparts on a couple of datasets, the results we obtained reflect that it is also a good match for TSC in the case of sleep apnea. Since the number of FLOPs and parameters can be reduced drastically and the accuracy value is over 90% in almost all cases, the DSN can be deemed adequate for detecting sleep apnea events, especially in the case of low computational resources.

References

[1] (2016). Apnea-ECG Database Annotations.

[2] Anu Shilvya, J. and Jose, P. S. H. (2022). Review on obstructive sleep apnea detection using ecg signals. page 680 – 684.

[3] Ayatollahi, A., Afrakhteh, S., Soltani, F., and Saleh, E. (2022). Sleep apnea detection from ecg signal using deep cnn-based structures. *Evolving Systems*.

[4] Bahrami, M. and Forouzanfar, M. (2022). Sleep apnea detection from single-lead ecg: A comprehensive analysis of machine learning and deep learning algorithms. *IEEE Transactions on Instrumentation and Measurement*, 71:1–11.

[5] Cowie, M. R. (2017). Sleep apnea: State of the art. *Trends in Cardiovascular Medicine*, 27(4):280–289.

[6] Erdenebayar, U., Kim, Y. J., Park, J.-U., Joo, E. Y., and Lee, K.-J. (2019). Deep learning approaches for automatic detection of sleep apnea events from an electrocardiogram. *Computer Methods and Programs in Biomedicine*, 180:105001.

[7] Evci, U., Gale, T., Menick, J., Castro, P. S., and Elsen, E. (2020). Rigging the lottery: Making all tickets winners. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2943–2952. PMLR.

[8] Fatimah, B., Singh, P., Singhal, A., and Pachori, R. B. (2020). Detection of apnea events from ecg segments using fourier decomposition method. *Biomedical Signal Processing and Control*, 61:102005.

[9] Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. (2000 (June 13)). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220. Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

[10] Goyal, M. and Johnson, J. (2017). Obstructive Sleep Apnea Diagnosis and Management. *Missouri medicine*, 114(2):120–124.

[11] Gupta, K., Bajaj, V., and Ansari, I. A. (2022). Osacn-net: Automated classification of sleep apnea using deep learning model and smoothed gabor spectrograms of ecg signal. *IEEE Transactions on Instrumentation and Measurement*, 71:1–9.

[12] Han, S., Pool, J., Tran, J., and Dally, W. J. (2015). Learning both weights and connections for efficient neural networks. *CoRR*, abs/1506.02626.

[13] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

[14] Kachuee, M., Fazeli, S., and Sarrafzadeh, M. (2018). ECG heartbeat classification: A deep transferable representation. *CoRR*, abs/1805.00794.

[15] Mostafa, H. and Wang, X. (2019). Parameter efficient training of deep convolutional neural networks by dynamic sparse reparameterization.

[16] Penzel, T., Moody, G., Mark, R., Goldberger, A., and Peter, J. (2000). The apnea-ecg database. In *Computers in Cardiology 2000. Vol.27 (Cat. 00CH37163)*, pages 255–258.

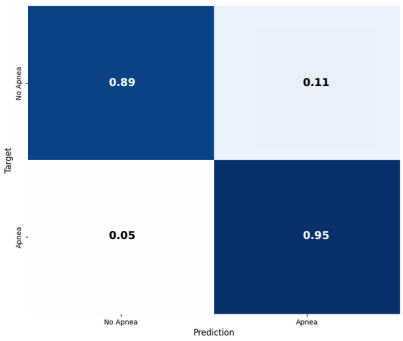
[17] Salari, N., Hosseinian-Far, A., Mohammadi, M., Ghasemi, H., Khazaie, H., Daneshkhah, A., and Ahmadi, A. (2022). Detection of sleep apnea using machine learning algorithms based on ecg signals: A comprehensive systematic review. *Expert Systems with Applications*, 187:115950.

[18] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2019). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359.

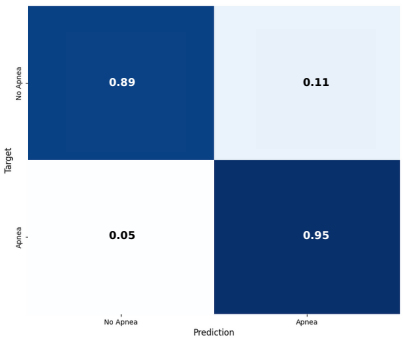
[19] Tang, W., Long, G., Liu, L., Zhou, T., Blumenstein, M., and Jiang, J. (2020). Omni-scale cnns: a simple and effective kernel size configuration for time series classification.

[20] Xiao, Q., Wu, B., Zhang, Y., Liu, S., Pechenizkiy, M., Mocanu, E., and Mocanu, D. C. (2022). Dynamic sparse network for time series classification: Learning what to "see". In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

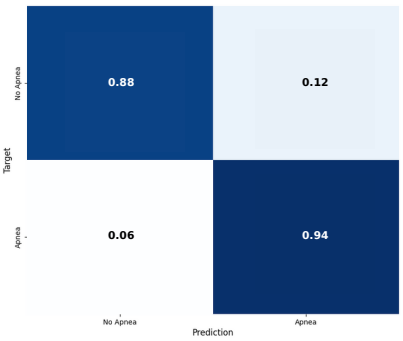
Appendix A



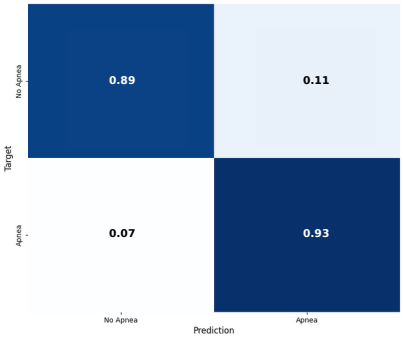
(a) S = 0%



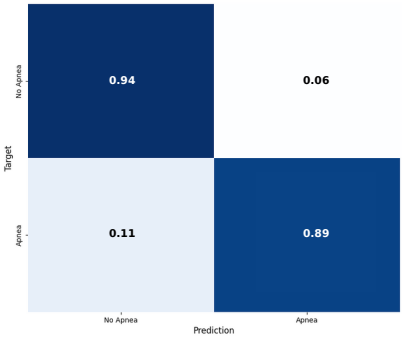
(b) S = 60%



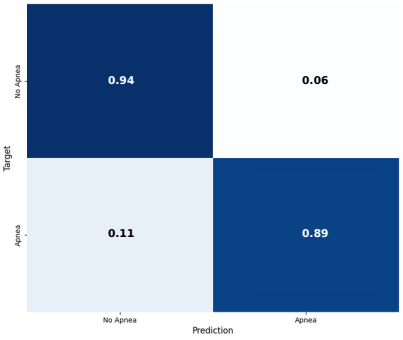
(c) S = 70%



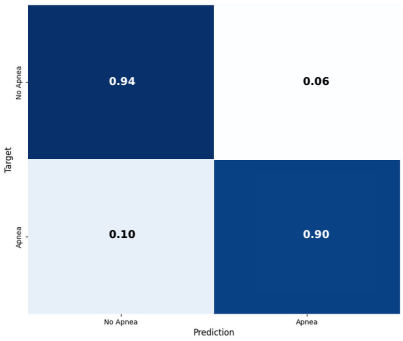
(d) S = 80%



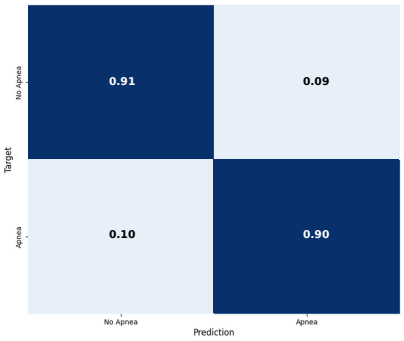
(e) S = 85%



(f) S = 90%



(g) S = 95%



(h) S = 98%

Figure 6: Confusion matrices (averaged) for models trained at different sparsities