Lab 3 - Reproducibility / Data Science course 2023

A Reproducibility Study of Apnea Detection

Ernest Onifade[1], Harry Setiawan Hamjaya[1]Mariama Oliveira[1]
[1]Åbo Akademi

24 October 2023

Reproducibility Summary

It is becoming more common for people to experience sleep apnea due to aging and an unhealthy diet. Currently, the usual way to diagnose apnea is through testing in a hospital, which is not always convenient or easily accessible. With the increasing demand, hospitals are struggling to keep up with the workload. Therefore, it would be beneficial for everyone if individuals could monitor their sleep at home.

Scope of Reproducibility — Addressing missing data in noisy ECG signals and developing a lightweight neural network model that could classify apnea accurately.

Methodology — We used the code provided on the GitHub repository of the original study to implement the SE-MSCNN model. Afterward, we compared its performance to other baseline machine learning models. The original code required minimal modifications. The experiment was conducted on Colab, using its default configuration, and took around 3.5 hours to complete.

Results — Overall in comparison to the author's model, our model had an accuracy of 97.14%, sensitivity of 95.65%, and a correlation of 0.968, while that of the author was 100%, 100%, and 0.979 respectively, but had similar specificity (100%) which is quite comparable.

What was easy — A clear machine learning pipeline, easy data access, and perfect code documentation are the main factors of the success of the reproducibility task.

What was difficult — The primary difficulty in replicating the paper lies in comprehending the research document, which is notably challenging due to its focus on a biological subject, specifically apnea.

Communication with original authors — As of writing this document, we have not been able to establish contact with the author. This is because there have been no significant issues that require additional clarification, as most of the paper and the code are self-explanatory. Furthermore, we were unable to find the author's contact details.

Correspondence should be addressed to Xianhui Chen (xhchen@nyu.edu)
The authors have declared that no competing interests exists.
Code is available at https://github.com/Bettycxh/Toward-Sleep-Apnea-Detection-with-Lightweight-Multi-scaled-Fusion-Network.
Data is available at https://physionet.org/content/apnea-ecg/1.0.0/.

1 Introduction

Sleep apnea, a common sleep disorder, happens when a person's upper airway gets blocked during sleep, often because of the tongue or excess fatty tissue. This can lead to breathing problems for at least 10 seconds per minute, which, in turn, reduces oxygen and increases carbon dioxide levels in the blood. People with sleep apnea face several problems, like feeling tired during the day, memory issues, and an increased risk of health conditions such as stroke, diabetes, and heart disease. In severe cases, it can even be fatal.

The number of people with sleep apnea has been increasing, mainly because of aging and changes in our diets. The usual way to diagnose it is with a sleep test in a hospital, but this test is inconvenient, and there aren't enough resources in hospitals to meet the growing demand. So, finding a way for people to monitor their sleep at home could make things easier for both patients and hospitals.

2 Scope of reproducibility

This study aims to reproduce the original work of the paper by Chen et al [1]. The main claims of the work are listed below:

- The study addresses missing data in noisy ECG signals, nearby segments are used to fill gaps.
- The study develops a simple neural network that combines information from different scales.
- The model created is lightweight with only 41,162 parameters, equivalent to 0.157 megabytes, making it small enough to be used in wearable devices for real-time detection.
- The proposed model (Single lead ECG - Multi Scaled Convolutional Neural Network) is a highly accurate method for detecting sleep apnea, outperforming other methods in most measures.

In the previous experiments, they evaluated the performance of the model per segment and per recording. However, in our reproducibility study, we focus on evaluating the model per recording.

3 Methodology

Since the code found on the study GitHub was well-organized, reusing the code was almost straightforward. The only alteration we made was converting all the Python files (.py) to a single Jupyter notebook (.ipynb). We also had to relocate the files and change the paths used in the code, but no other changes were needed. The code of the reproducibility study can be found at GitHub [1].

The main libraries employed to run the experiment include pickle, BioSPPy, wfdb, SciPy, NumPy, scikit-learn, and TensorFlow. In addition, we used the basic configuration of Colab, 12GB RAM, and the TPU of the system.

[1] https://github.com/mariamaOlive/reproduction_apnea_detection

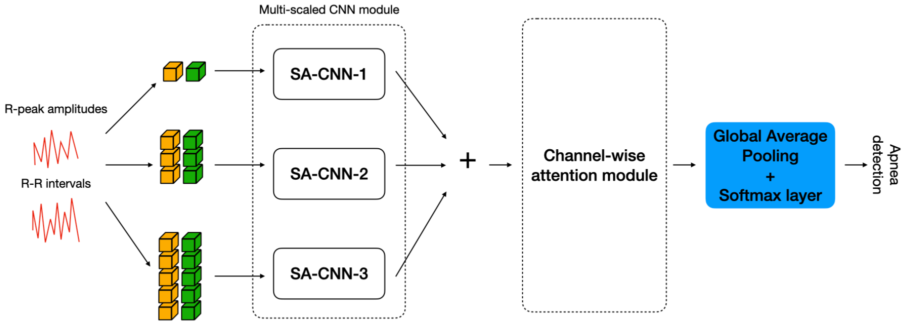


Figure 1. SE-MSCNN model architecture. It is comprised mainly of three modules (1) a multi-scaled convolutional neural network (CNN), (2) a channel-wise attention module and (3) a global average pooling layer and a Softmax layer.

3.1 Model descriptions

The experiment proposed a model named SE-MSCNN, which comprised of two modules: a multi-scaled convolutional neural network (CNN) module and a channel-wise attention module (refer to Figure 1). The first module consists of three CNNs, each responsible for receiving a segment of R peak amplitudes and R-R intervals in different sizes (1, 3, and 5 minutes). The second module is a channel-wise attention component that was responsible for aggregating the multiple scaled features extracted by the first model. After the second module, a global average pooling layer and a Softmax layer classify the segments as normal or apnea.
Besides the SE-MSCNN model, other traditional ML models were used as baselines, including SVM, KNN, Logistic Regression, and MLP.

3.2 Datasets

The dataset of the experiment was provided by the PhysioNet Apnea-ECG, Philipps University [2]. It is a dataset utilized to measure mainly sleep quality. It consists of 70 overnight recordings of ECGs sampled at 100 Hz. The age of the participants ranges from 27 to 63 years old, and the duration of the recordings ranges from 401 to 578 minutes. The recordings are divided into 35 for training and 35 for test. The total duration of the recordings is 34039 minutes, of which 21010 minutes (61.72%) are normal and 13029 minutes (38.28%) show sleep apnea occurrences.
Since EGC can be easily polluted by noise, the recorded signals underwent several transformations before extracting the features. A low-pass filter was utilized to remove any high-frequency noise with a cut-frequency (<16 Hz), while a high-pass filter with a cut-frequency (>8 Hz) was used to eliminate low-frequency noise. A moving average window with 80 ms was also utilized to accurately position the QRS complexes and compute the R peak amplitudes and R-R intervals.

3.3 Hyperparameters

Default hyperparameters were used for the model, while others were initialized in the experiment.

- The initial learning rate was assigned a value of 0.001. This value was later reduced by a factor of 0.1, every 10 iterative epochs, after the number of iterative epochs exceeded 70;

[2] https://physionet.org/content/apnea-ecg/1.0.0/

- The maximum number of iterative epochs was set to 100.
- The batch size was fixed at 128.
- For the one-dimensional convolutional filter, the He_normal initializer was chosen.
- The l2 regularization was included in the error loss function to reduce overfitting.

These hyperparameters were optimized using a training and validation set of 70 and 30, respectively.

3.4 Experimental Setup

The model was trained and tested on the aforementioned dataset and hyperparameters. Afterwards, its metrics were obtained to compare with other baselines.
In order to verify the model performance, the original study opted for two stages of apnea detection. The first stage involved detecting apnea in 1-minute segments, while the second stage involved detecting apnea per recording. Both stages were binary classification tasks. However, in our reproducibility study, we chose only the detection per recording. The metrics used to evaluate the classification per segment were accuracy (Acc), sensitivity (Sens), specificity (Spec), and area under the receiver operating characteristic curve (AUC). In addition to these metrics, the per-recording performance was also evaluated based on the Apnea-Hypopnea Index (AHI), which measures sleep quality. The Pearson correlation of the AHI of the prediction and ground truth was also calculated.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Sen = \frac{TP}{TP + FN} \tag{2}$$

$$Spec = \frac{TN}{TN + FP} \tag{3}$$

$$AHI = \frac{60}{T} \times N \tag{4}$$

3.5 Computational requirements

The experiment was conducted using Google Colab. Therefore, the configuration was completely dependent on the system provided by Google. As a result, we used the basic configuration of Colab, 12GB RAM, and the TPU of the system. The total number of hours to run the experiment was about 3.5 hours.

4 Results

Overall, though our model's performance metrics are comparable to the author's, the author's model slightly outperforms ours regarding model accuracy, sensitivity, and correlation. Nevertheless, our model's performance supports the author's claim that the SC-MSCNN method achieves the best classification performance in assessing an individual's sleep quality. Furthermore, it corroborates the author's SC-MSCNN model consistency in making predictions on overnight ECG recordings between actual AHI and predicted AHI with a correlation of 0.968. Similarly, aside from the autoencoder method by Li et al. [2], which had comparable performance to the author's own SC-MSCNN model, our model outperformed other state-of-the-art SA detection methods.

Finally, our model, like the author's, is lightweight, occupying approximately 0.2MB of space, and thus supports the claim to potentially embed it into a wearable device for monitoring sleep quality.

Method	Acc (%)	Sens (%)	Spec (%)	AUC (%)	Corr (%)
SVM	88.6	100	66.7	97.9	85.2
LR	88.6	100	66.7	98.2	84.1
KNN	82.9	100	50	98.6	84.5
MLP	85.8	95.7	66.7	95	81.4
SE-MSCNN (Our Model)	**97.2**	**95.7**	**100**	**99.6**	**96.8**

Table 1. Evaluation metrics between our model and other methods

4.1 Our Model's Performance Relative to Author's

In comparison to the author's model, our model had an accuracy of 97.14%, sensitivity of 95.65%, and a correlation of 0.968, while that of the author was 100%, 100%, and 0.979 respectively. In terms of the specificity, our model and that of the author's were similar at 100%.

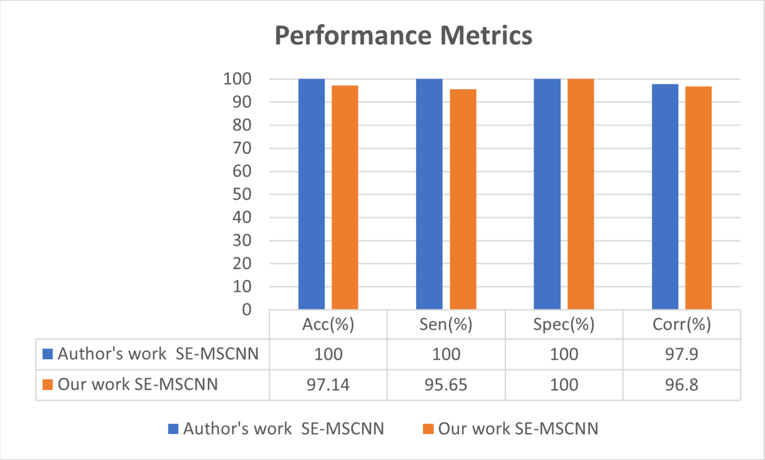


Figure 2. Performance metrics between our model and author's

4.2 Our Model's Performance Relative to Other Methods

In relation to the other methods (SVM, LR, KNN, MLP), our model SC-MSCNN outperformed them by a margin of about 10% in accuracy, 40% in specificity, 2-3% in AUC, and about 10% in correlation. However, surprisingly, they had a higher sensitivity than our model by about 5%.

4.3 Our Model's Performance Relative to State-of-the-art Methods

Compared to the state-of-the-art methods, our model was similar to Song et al. HMM-SVM [3], Sharma et al. LS-SVM [4], and Wang et al. LeNet-5 [5] in terms of accuracy and specificity. However, our model had a higher correlation than theirs and a higher specificity than Wang et al. LeNet-5 method. Our model way outperformed Alvarez et al. LR [6] in all evaluation metrics by about 10%.

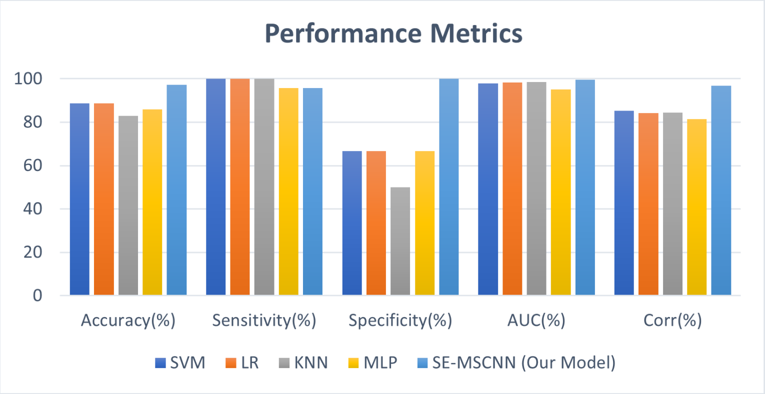
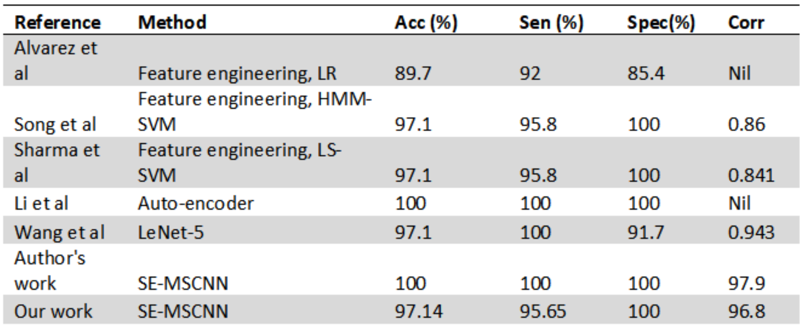


Figure 3. Perfromance metrics across different classification methods

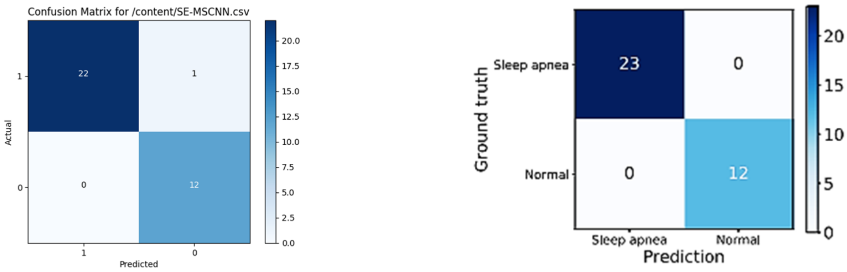


Reference	Method	Acc (%)	Sen (%)	Spec(%)	Corr
Alvarez et al	Feature engineering, LR	89.7	92	85.4	Nil
Song et al	Feature engineering, HMM-SVM	97.1	95.8	100	0.86
Sharma et al	Feature engineering, LS-SVM	97.1	95.8	100	0.841
Li et al	Auto-encoder	100	100	100	Nil
Wang et al	LeNet-5	97.1	100	91.7	0.943
Author's work	SE-MSCNN	100	100	100	97.9
Our work	SE-MSCNN	97.14	95.65	100	96.8

Figure 4. Evaluation metrics between our model and state-of-the-art methods

4.4 Confusion metrics Predictions

To further visualize the performance of our model, the confusion metric shows a total of 22 predictions for the presence of sleep apnea per recording classification, out of which none of the predictions were wrong. It made a total of 13 predictions for the absence of sleep apnea, out of which it got one wrong. This is very similar to the predictions made by the author's SE-MSCNN model asides, which got all predictions for the absence of sleep apnea correct.



(a) Our model SE-MSCNN confusion matrix

(b) Author's model confusion matrix

Figure 5. Confusion matrices

5 Discussion

During our study, we faced several challenges with the first paper we chose, ”Matching news articles and Wikipedia tables for news augmentation” [7]. Subsequently, we decided to switch to another paper, ”Toward sleep apnea detection with lightweight multi-scaled fusion network” [1]. The latter paper proved to be simpler to replicate compared to the former, which required a lot of computational power such as CPU and GPU memory for its experiments. With the second paper, we were able to easily redo and reproduce the experiments.

5.1 What was easy

There are several notable good practices that are provided by the author:

- Easy data access. It is easily accessed even by someone who isn’t in the Biology domain since it is publicly available.
- Perfect code documentation. It is easily run after installing several important libraries such as biosppy and wfdb.
- Clear machine learning pipeline. It is well written and described in both the paper and the GitHub.

5.2 What was difficult

There is one significant challenge in reproducing the paper which is understanding the research paper which is quite challenging because it deals with a biology topic, specifically, Apnea.

5.3 Communication with original authors

At the time of writing this document, we haven’t been in touch with the author. This is because there aren’t any major issues requiring further clarification, as the majority of the paper and the code are quite straightforward.

6 Conclusion

In the end, we managed to get similar results to the ones obtained by the original study. Implementing the code was also easy since the GitHub of the study was well organized and provided instructions on how to run the files. However, this does not mean that the Lab 3 did not present any challenges.

Initially, this lab assignment proved to be quite difficult because we chose a paper [7] that presented several setbacks, mainly related to memory resource limitations. We attempted different approaches, including running the code on various environments such as Colab, personal computers, and university computer. We even contacted one of the authors but ultimately gave up after discovering that the proposed architecture in the paper was too large for our available resources. This could have been avoided if the original paper had provided detailed information about the experiment’s system configuration and running time. Moreover, the project’s GitHub repository was not clear enough since it did not specify the necessary libraries and versions.

Overall, the Lab 3 assignment was interesting as it taught us the importance of including detailed experiment steps in research papers and creating a well-documented repository of the experiment code.

References

1. X. Chen, Y. Chen, W. Ma, X. Fan, and Y. Li. “Toward sleep apnea detection with lightweight multi-scaled fusion network.” In: **Knowledge-Based Systems** 247 (2022), p. 108783.
2. K. Li, W. Pan, Y. Li, Q. Jiang, and G. Liu. “A method to detect sleep apnea based on deep neural network and hidden Markov model using single-lead ECG signal.” In: **Neurocomputing** 294 (2018), pp. 94–101.
3. C. Song, K. Liu, X. Zhang, L. Chen, and X. Xian. “An obstructive sleep apnea detection approach using a discriminative hidden Markov model from ECG signals.” In: **IEEE Transactions on Biomedical Engineering** 63.7 (2015), pp. 1532–1542.
4. H. Sharma and K. Sharma. “An algorithm for sleep apnea detection from single-lead ECG using Hermite basis functions.” In: **Computers in biology and medicine** 77 (2016), pp. 116–124.
5. T. Wang, C. Lu, G. Shen, and F. Hong. “Sleep apnea detection from a single-lead ECG signal with automatic feature-extraction through a modified LeNet-5 convolutional neural network.” In: **PeerJ** 7 (2019), e7731.
6. D. Alvarez, R. Hornero, J. V. Marcos, and F. del Campo. “Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis.” In: **IEEE Transactions on Biomedical Engineering** 57.12 (2010), pp. 2816–2824.
7. L. Silva and L. Barbosa. “Matching news articles and wikipedia tables for news augmentation.” In: **Knowledge and Information Systems** 65 (4 Apr. 2023), pp. 1713–1734.