HR2SleepEvent: Detecting Apneas and Hypopneas During Sleep by Measuring Heart Rate Changes and Assessing Out-of-Sample Performance

Brynn Voigt[1,2], Samantha Unger[1,2], Shaghayegh Chavoshian[1,3], Daniel Franklin[1,2] and Azadeh Yadollahi[1,3]

[1] Institute of Biomedical Engineering, University of Toronto, Toronto, Canada
[2] Ted Rogers Centre for Heart Research, Peter Munk Cardiac Centre – University Health Network, Toronto, Canada
[3] KITE, Toronto Rehabilitation Institute – University Health Network, Toronto, Canada

Abstract—**Sleep Apnea is a breathing condition characterized by episodes of reduced airflow during sleep, with the airway partially (hypopnea) or fully (apnea) obstructed. Conventional methods of diagnosis include overnight sleep studies, which are resource and time intensive. Previous studies have shown that heart rate and heart rate variability are associated with sleep apnea severity. However, the performance of these features in detecting respiratory events (apnea or hypopnea), and particularly in unknown datasets, was not examined. We trained a set of conventional machine learning models to label segments of electrocardiography data based on whether they contained apneas or hypopneas. Tuning hyperparameters using leave-one-subject-out cross-validation, logistic regression was found to have strong performance across area under the receiver-operating curve, accuracy, specificity, sensitivity and F1 score metrics, with scores of 0.736±0.102, 72.3±15.8%, 92.0±4.8%, 27.7±16.9%, and 31.7±17.4%, respectively. The application of this model to another dataset, the apnea-ECG dataset, showed an average accuracy of 60.7%. We have also assessed whether there were age- or sex-based differences in model performance. This study thus provides a workflow for comparing machine learning models for apnea detection and highlights how models may not perform as strongly on other datasets.**

Keywords—**Sleep Apnea, Electrocardiography, Machine Learning, Heart Rate Variability**

I. Introduction

A. Sleep Apnea

Sleep Apnea is a breathing condition resulting in reduced airflow during sleep, with an estimated 30% of Canadians at intermediate or high risk [1]. The reduced airflow can fall into two main categories, apneas and hypopneas depending on the level of airway obstruction [2]. Detecting sleep apnea starts with preliminary screening including sleep history, questionnaires (Epworth Sleepiness Scale, STOP-BANG), and physical examinations to look for indicators of sleep apnea, [2]. If preliminary screening indicates possible sleep apnea, the next step is either a home sleep apnea test, or a polysomnography (PSG), which is considered the gold standard for diagnosis [2]. PSGs, the gold standard, involve monitored in-lab sleep sessions that record signals such as electrocardiogram (ECG), electromyogram, and oxygen saturation [3]. PSGs are considered accurate but are time and resource intensive and require operator expertise [2]. Challenges with conventional sleep apnea diagnostic methods, including costs, availability, and extensive time requirements [4], have motivated development of computer assisted diagnosis [5].

B. Predicted Indicators of Sleep Apnea

To facilitate computer assisted diagnosis, key features of ECG signals are extracted and used in the classification model. For this investigation, heart rate (HR) and heart rate variability (HRV) were extracted from an ECG. HRV is a measurement of the changes in the intervals between heartbeats, and there are a variety of metrics analyzing this variation in both time and frequency [6]. Time based metrics of HRV represent the overall variability in HR, whereas frequency-based metrics provide information about the power distribution of the signal [7]. HR and HRV metrics were selected for the classification model as people with sleep apnea have been shown to have higher resting HR, and observable differences in HRV components [8]. 5-minute segments are recommended as windows to capture the variability in the signal [7].

C. Related Works

Many previous works [5], [9]–[14] have used the apnea-ECG database [15] for computer assisted diagnosis. With this dataset, HRV metrics with time and frequency domain features have been used to achieve accuracy above 90% [9], but many methods have included manual annotation in their pipeline [16], [17]. Beyond HRV, deep learning models such as convolutional neural networks [11] and long short-term memory models [5], [13] have been investigated. Despite strong performance, the applicability of these models to other datasets is underexplored and is crucial for mitigating the gap between model performance "locally" and at other sites, motivating our approaches in this study. This project aims to develop a model to detect apneas and hypopneas using only ECG signals and assess performance by validating the model across other datasets.

II. Methods

A. Dataset

The dataset used to train the model is available on PhysioNet [18] and contains 25 overnight PSGs of individuals with suspected sleep apnea or primary snoring [19]. For each participant, respiratory events during sleep were annotated by a sleep technologist and a file of event types (apneas/hypopneas) and duration is provided. Of the 25 subjects, 21 are male and 4 are female, average age is 50 ± 10 years (range: 28-68), and average body mass index (BMI) is 31.6 ± 4.0 kg/m^2 (range: 25.1 - 42.5 kg/m^2) [19]. Only the ECG signals were analyzed.

B. Signal Preprocessing

To extract HR and HRV metrics (selection in Table 1) from the ECG signal, recordings were processed using the NeuroKit2 Python toolbox [20]. The signals were cleaned and denoised, and peaks and features were extracted using the algorithm suggested by NeuroKit2. The ECG signals were windowed into 1-minute non-overlapping segments for the entire recording to match the provided labels in the dataset. HR was calculated based on these 1-minute intervals, but for HRV a window of 5 minutes was used to follow established guidelines to account for signal variability [7]. There was a total of 10,274 windows used for training with 84 features included for each window.

Table 1 Selection of HRV features used for classification

Feature		Description
Time	HRV Mean NN	Mean NN interval in signal
	SDNN	Standard Deviation of NN intervals in signal
Frequency	Low Frequency (LF)	Power Spectrum, Frequency Range 0.04 - 0.15 Hz
	High Frequency (HF)	Power Spectrum, Frequency Range 0.15 - 0.4 Hz
	LF/HF	Ratio LF to HF Power

C. Model Development

Machine learning models were trained using leave-one-subject-out cross-validation due to few subjects in the dataset [21]. Linear and non-linear classifier models from the following list were implemented using the scikit-learn Python library [22]—logistic regression, K-nearest neighbours, linear support vector machine (SVM), decision tree, random forest, neural net, AdaBoost, naive bayes and quadratic discriminant analysis (QDA). Hyperparameters for models were tuned. To assess model performance, accuracy, specificity, sensitivity, and F1 score were calculated from true positive, true negative, false positive, and false negative predictions. Model performance was also evaluated by area under the receiver-operating curve (AUROC) to assess how well the models can discriminate between windows with and without respiratory events (Fig. 1).

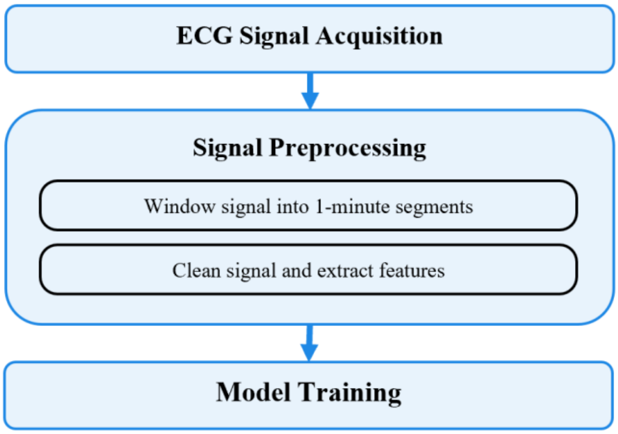
ECG Signal Acquisition
Signal Preprocessing
Window signal into 1-minute segments
Clean signal and extract features
Model Training

Fig. 1 Methodology followed including data acquisition, signal processing, and model development.

Results

D. Model Training

Models were trained sweeping across different hyperparameters to achieve optimal accuracy. For the best model of each type, the accuracy, F1 score, sensitivity, specificity, and AUROC were computed for both the training and validation set for each cross-validation fold.

To perform subsequent analysis, the logistic regression model was selected as the best model for comparison. The decision to choose this model was based on a good ability to discriminate between examples with and without respiratory events, demonstrated by the highest mean AUROC, and performing well across other metrics. This model was then trained on all the subjects in the initial dataset.

E. Out-of-Sample Validation

The logistic regression model was used to predict respiratory events for each subject in the secondary dataset. A summary of performance metrics was generated for each subject, with the average results shown in Table 2 for the cross-validation with the initial dataset and the out-of-sample validation. The model tends to correctly predict windows where no respiratory events occur, but has a high number of false negatives, indicating it tends to predict the absence of a respiratory event even when one is occurring.

Table 2 Performance Metrics for LOSO Cross-Validation of Initial Dataset and Out-of-Sample Validation

Metric	LOSO Performance	Out-of-Sample Performance
Accuracy	72.3 ± 15.8%	60.7 ± 26.9%
Specificity	92.0 ± 4.8%	87.7 ± 0.1%
Sensitivity	27.7 ± 16.9%	13.8 ± 17.0%
F1 Score	31.7 ± 17.4%	15.0 ± 14.1%
AUROC	0.736 ± 0.102	0.569 ± 0.175

To address any specific differences in model performance due to age, sex, or BMI, a statistical analysis was conducted using JMP, a statistical software program [23]. For each category, the subjects were divided into one of two groups, age greater than or less than 50, sex of male or female, and BMI over or under 35. Using a one-way ANOVA, a statistically significant difference was found in validation accuracy between the two age groups ($p< 0.0001$), as well as by sex ($p= 0.012$). The power of the age-based test was 99%, however the residuals were not normally distributed so the conditions to use ANOVA were not met.

The model's tendency to classify a window as not an apnea on the validation dataset, even when one occurs, results in a higher number of false negatives. False negatives could lead to under-detection of respiratory events.

III. Discussion

Models achieved similar accuracy to conventional machine learning models in a prior work [13]. Despite similarities with prior work, F1 scores from our models are consistently lower than those found in literature. This is likely due to how skewed the dataset is, having mainly windows without apneas. Thus, despite accuracy score, our models have a high false negative rate, leading to low sensitivity and a low F1 score. Future work may explore larger datasets for with more balanced group sizes or to use sampling strategies [24].

The AUROC values, in Table 2, demonstrate that our models can discriminate between windows with and without apneas, as these values exceed 0.5. This implies the threshold for labeling an example as an apnea could be adjusted to improve detection of apneas at expense of more false positives.

The statistically significant ($p<0.0001$) difference in performance between model performance on subjects of different ages also warrants discussion. It is possible that the models perform worse on older adults since they have more apneas in our dataset and our model is conservative in labeling a window as containing an apnea. As our model underpredicts apneas, this could explain poorer performance on groups with higher incidence of these events.

A. Limitations and Next Steps

Due to the small sample size and lack of intersectional diversity within the datasets, our ability to compare model performance in different subgroups was limited. For example, while the apnea-ECG dataset does contain 7 female participants, 6 participants with a BMI above 35 kg/m^2, and 13 participants above age 50, none of the female participants belong to these at-risk groups based on age and BMI. Intersectionality is thus important to be considered in participant demographics. We urge future researchers to consider this when selecting participants for studies regarding sleep apnea and encourage more datasets to be made openly available.

Future work should also perform investigations on the impact of window size on results. The model developed in this paper considered 1-minute non-overlapping windows, and as such there is a possibility of only the last second of a window containing the respiratory event, making it unlikely for the model to correctly identify the event. Changing the window size, or considering a combination of overlapping time windows, could create a more representative model that would better account for the time to see physiological changes.

B. Significance

With the present low diagnosis rate of sleep apnea [25], an automated method to detect respiratory events could serve as an additional screening tool. A cited barrier to diagnosis has been poor coordination of health services with long wait times for assessment and a delay in receiving results [26]. Automating detection with models like those developed in this report could support at-home screening tools and reduce time needed interpreting the studies. Left untreated, sleep apnea is associated with daytime sleepiness and cardiovascular disease [4]—an automated method could help improve treatment rates, improving health and quality of life.

IV. Conclusion

The linear regression model presented displays promising initial results towards automated classification of respiratory events based solely on ECG signal. The results of testing and out-of-sample validation indicate future work should focus on minimizing false negatives. Given the negative health outcomes associated with sleep apnea, further attention on automated apnea classification is valuable for diagnosis.

Acknowledgment

S.U. and S.C. gratefully acknowledge support from the NSERC CREATE Translating Cardiovascular Remote Diagnostic and Monitoring for Health Equity (CaRDM Eq) training program (Grant #575430-2023).

Conflict of Interest

The authors had no conflict of interest.

References

[1] S. C. Government of Canada, "Sleep Apnea in Canada, 2016 and 2017." Accessed: Oct. 15, 2023. [Online]. Available: https://www150.statcan.gc.ca/n1/pub/82-625-x/2018001/article/54979-eng.htm

[2] J. V. Rundo, "Obstructive sleep apnea basics," CCJM, vol. 86, no. 9 suppl 1, pp. 2–9, Sep. 2019, doi: 10.3949/ccjm.86.s1.02.

[3] B. Jafari and V. Mohsenin, "Polysomnography," Clinics in Chest Medicine, vol. 31, no. 2, pp. 287–297, Jun. 2010, doi: 10.1016/j.ccm.2010.02.005.

[4] D. J. Gottlieb and N. M. Punjabi, "Diagnosis and Management of Obstructive Sleep Apnea: A Review," JAMA, vol. 323, no. 14, pp. 1389–1400, Apr. 2020, doi: 10.1001/jama.2020.3514.

[5] P. Panindre, V. Gandhi, and S. Kumar, "Artificial Intelligence-based Remote Diagnosis of Sleep Apnea using Instantaneous Heart Rates," in 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Jan. 2021, pp. 169–174. doi: 10.1109/Confluence51648.2021.9377149.

[6] F. Shaffer and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms," Front Public Health, vol. 5, p. 258, Sep. 2017, doi: 10.3389/fpubh.2017.00258.

[7] T. Pham, Z. J. Lau, S. H. A. Chen, and D. Makowski, "Heart Rate Variability in Psychology: A Review of HRV Indices and an Analysis Tutorial," Sensors, vol. 21, no. 12, Art. no. 12, Jan. 2021, doi: 10.3390/s21123998.

[8] S. Ucak, H. U. Dissanayake, K. Sutherland, P. de Chazal, and P. A. Cistulli, "Heart rate variability and obstructive sleep apnea: Current perspectives and novel technologies," Journal of Sleep Research, vol. 30, no. 4, p. e13274, 2021, doi: 10.1111/jsr.13274.

[9] T. Penzel, J. McNames, P. de Chazal, B. Raymond, A. Murray, and G. Moody, "Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings," Biological Engineering, vol. 40, 2002.

[10] 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," Neurocomputing, vol. 294, pp. 94–101, Jun. 2018, doi: 10.1016/j.neucom.2018.03.011.

[11] S. M. I. Niroshana, X. Zhu, K. Nakamura, and W. Chen, "A fused-image-based approach to detect obstructive sleep apnea using a single-lead ECG and a 2D convolutional neural network," PLOS ONE, vol. 16, no. 4, p. e0250618, Apr. 2021, doi: 10.1371/journal.pone.0250618.

[12] A. Sheta et al., "Diagnosis of Obstructive Sleep Apnea from ECG Signals Using Machine Learning and Deep Learning Classifiers," Applied Sciences, vol. 11, no. 14, Art. no. 14, Jan. 2021, doi: 10.3390/app11146622.

[13] M. Bahrami and M. Forouzanfar, "Sleep Apnea Detection From Single-Lead ECG: A Comprehensive Analysis of Machine Learning and Deep Learning Algorithms," IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1–11, 2022, doi: 10.1109/TIM.2022.3151947.

[14] J. Hayano et al., "Quantitative detection of sleep apnea with wearable watch device," PLoS One, vol. 15, no. 11, p. e0237279, Nov. 2020, doi: 10.1371/journal.pone.0237279.

[15] T. Penzel, G. B. Moody, R. G. Mark, A. L. Goldberger, and J. H. Peter, "The apnea-ECG database," in Computers in Cardiology 2000. Vol.27 (Cat. 00CH37163), Sep. 2000, pp. 255–258. doi: 10.1109/CIC.2000.898505.

[16] B. Raymond, R. M. Cayton, R. A. Bates, and M. Chappell, "Screening for obstructive sleep apnoea based on the electrocardiogram-the computers in cardiology challenge," in Computers in Cardiology 2000. Vol.27 (Cat. 00CH37163), Sep. 2000, pp. 267–270. doi: 10.1109/CIC.2000.898508.

[17] J. N. McNames and A. M. Fraser, "Obstructive sleep apnea classification based on spectrogram patterns in the electrocardiogram," in Computers in Cardiology 2000. Vol.27 (Cat. 00CH37163), Sep. 2000, pp. 749–752. doi: 10.1109/CIC.2000.898633.

[18] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals," Circulation, vol. 101, no. 23, Jun. 2000, doi: 10.1161/01.CIR.101.23.e215.

[19] W. McNicholas, L. Doherty, S. Ryan, J. Garvey, P. Boyle, and E. Chua, "St. Vincent's University Hospital / University College Dublin Sleep Apnea Database." physionet.org, 2004. doi: 10.13026/C26C7D.

[20] D. Makowski et al., "NeuroKit2: A Python toolbox for neurophysiological signal processing," Behav Res, vol. 53, no. 4, pp. 1689–1696, Aug. 2021, doi: 10.3758/s13428-020-01516-y.

[21] D. M. Hawkins, S. C. Basak, and D. Mills, "Assessing Model Fit by Cross-Validation," J. Chem. Inf. Comput. Sci., vol. 43, no. 2, pp. 579–586, Mar. 2003, doi: 10.1021/ci025626i.

[22] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, no. 85, pp. 2825–2830, 2011.

[23] "JMP Pro." Cary, NC: SAS Institute Inc., 2022.

[24] N. V. Chawla, "Data Mining for Imbalanced Datasets: An Overview," in Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach, Eds., Boston, MA: Springer US, 2005, pp. 853–867. doi: 10.1007/0-387-25465-X_40.

[25] L. Simpson et al., "High prevalence of undiagnosed obstructive sleep apnoea in the general population and methods for screening for representative controls," Sleep Breath, vol. 17, no. 3, pp. 967–973, Sep. 2013, doi: 10.1007/s11325-012-0785-0.

[26] L. Ye, W. Li, and D. G. Willis, "Facilitators and barriers to getting obstructive sleep apnea diagnosed: perspectives from patients and their partners," J Clin Sleep Med, vol. 18, no. 3, pp. 835–841, Mar. 2022, doi: 10.5664/jcsm.9738