Predicting the Severity of Apnea through Machine Learning

Isabella Mixton-Garcia

University High School, 421 North Arcadia Avenue

Tucson, Arizona, 85711

Mentor: Xiaoxiao Sun

Abstract

Prediction modeling has limitless potential in insurance, retail, and healthcare. Predictive models allow corporations and health officials to make decisions based on statistical data. Obstructive sleep apnea (OSA) is a severe disease that affects 5% to 14% of adults. This disease has caused adverse effects to its constituents, and most people suffering from severe OSA die. This project presents a prediction model using random forest (RF) to accurately predict patients with none and severe OSA to make it easier to diagnose. Previous works have addressed models for prediction of OSA severity, but no significant study has been conducted using random forest model. This process was done by collecting demographic and polysomnography (PSG) data from 781 participants. Data analysis was conducted by only considering none and severe OSA and removing those with not applicable variables leading to only 339 participants being considered for model. The data was initiated in the RF model through the program RStudio. The results equated to 74.34% accuracy based on the confusion matrix for testing data. Making RF prediction model a valid model to predict OSA, this can shorten the diagnosis timeframe from a week to just one day, as the model can make its predictions through a questionnaire and overnight PSG test. This quick analysis allows sleep specialists to diagnose and treat the patient quicker.

Table of Contents

Acknowledgments

Thank you, Dr. Xiaoxiao Sun for teaching me the interesting topic of machine learning and helping to initiate the random forest model on RStudio. Thank you, STAR Labs, Margarat Wilch, and facilitators, for providing materials for research, revising research papers, and connecting me to a mentor.

Introduction

Sleep apnea is a severe disease where breathing repeatedly stops and starts. The most common form is obstructive sleep apnea (OSA) which occurs when throat muscles relax and decrease airflow by 30%-50% for more than 10 seconds (Strollo & Rogers, 1996). Central sleep apnea occurs when the brain does not signal the throat muscles to allow breathing, and complex sleep apnea syndrome includes both obstructive and central sleep apnea. This disease affects 5%–14% of adults aged 30–70 years. (Peppard PE et al.,2013) Apnea has adverse consequences on people's daily lives. Patients with OSA significantly complain more about work because of tiredness and can concentrate less on new tasks; furthermore, occupational accidents have a 50% increase in men with OSA and a sixfold increase in women suffering from OSA, (AlGhanim et al., 2007). Overall, this disease disrupts sleep and leads to lower blood oxygen levels, causing health complications later in life. Previous research has documented 810,000 collisions and 1400 fatalities attributed to OSA, costing 15.9 billion dollars in 2000 (AlGhanim et al., 2007).

Polysomnography (PSG) is a method for diagnosing OSA and the degree of sleep fragmentation. PSG records brain waves, blood oxygen level, heart rate, breathing, respiratory rates (RR), eye, and leg movements during hours of monitored sleep (Rodrigues et al., 2020). However, polysomnography is an expensive and time-consuming process, available at only a few sleep centers with highly trained personnel. The average time to receive medical therapy after

PSG diagnosis is 11.6 months (Rotenberg et al., 2010). During a screening for OSA, health officials have to test the severity with one of the three methods: apnea/hypopnea index (AHI), oxygen desaturation index (ODI), or time of hypoxia (T90). AHI is defined as the total number of apneas and hypopneas per hour of total sleep time (Rowley et al., 2000). Apnea happens when you stop breathing for 10 seconds or longer during sleep, and hypopnea is a reduction in nasal pressure amplitude by greater than 30% lasting for at least 10 seconds. The severities are listed as follows: mild apnea ($5 \leq AHI < 15$), moderate ($15 \leq AHI < 30$), and severe ($AHI \geq 30$). Arnardottir and others (2015) analyzed the impact of body mass index (BMI), sex, and age in predicting the AHI. The study also evaluated the relationship between sleep-disordered breathing and sleepiness by using Epworth Sleepiness Scale (ESS) and psychomotor vigilance test (PVT) in relation to AHI. Tsai and others (2021), compared age, sex, and the whole-body profile, consisting of body mass index (BMI), neck circumference, and waist circumference. It was found that BMI is the most influential parameter in OSA severity screening models.

Machine learning, an application of artificial intelligence (AI), is the study and development of systems that can learn from and make AI-based diagnoses from the data and has the potential to detect patients with severe OSA based on 2-dimensional images (Tsuiki et al., 2021). Using predictive models, which could contain classification models or regression models, can predict outcomes based on historical data. Classification models are probabilistic models that specify the conditional probability distributions of the output variables given the inputs. Regression models provide a function that describes the relationship between one or more independent variables and a response/dependent variable (Mencar et al., 2019). The classification model is advantageous because it reduces the effect of outliers, while the regression model effectively finds relations between variables more accurately.

An algorithm in machine learning is a procedure that is run on data to create a machine learning model. Some prediction algorithms are support vector machine (SVM), random forest (RF), AdaBoost-SVM, and CN2 rule induction. In comparison between support vector machine (SVM) and random forest (RF) classifiers, random forest performed better as the classifier shown by the F_1, F_1 a method for classification accuracy, ranked from 0 - 100, where a higher number indicates a greater prediction performance. RF has F_1=43.6, and SVM has F_1=43.6 and 41.0. This shows RF has greater precision and is comparable to SVM in recall (Mencar et al., 2019). However, SVM and Linear Regression are most effective in predicting the AHI value. This is the reason RF was chosen as the main predictive algorithm. RF is an ensemble learning method based on decision trees, where the trees arrive at an answer by asking a series of true/false questions about characteristics of the outcome of interest. Each tree in a random forest randomly samples subsets of the training data in a process known as bootstrap aggregating (bagging). The model is fit to these sub samples and the predictions are aggregated from all the decision trees. This allows RF to be trained on different features used to make robust and accurate decisions. This work assists in building the foundation for predictive models that estimate the severity of OSA, leading to an improved model using RF.

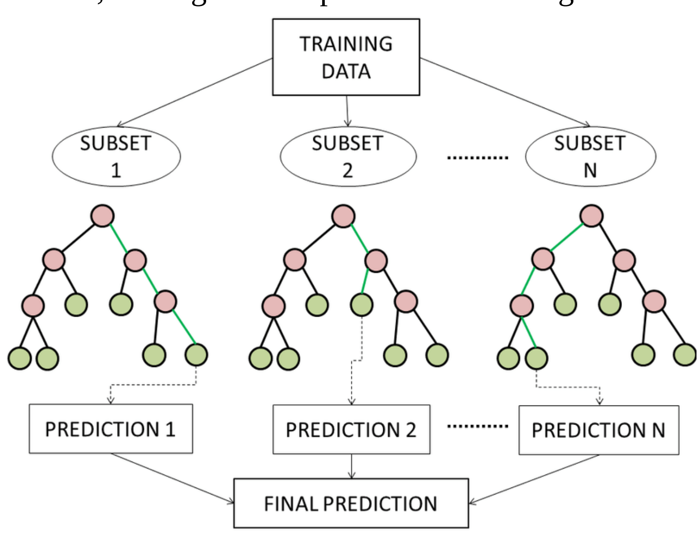

Figure 1: Example of RF workflow (Laudato et al., 2020)

Hypothesis

The use of random forest from polysomnography parameters can predict the severity of obstructive sleep apnea accurately.

Methods

Descriptive analysis

A demographic analysis of the 781 participants was conducted using polysomnography data to diagnose them with OSA officially. Patients with sleep conditions besides OSA were not considered for evaluation. Patients were divided into categories: those with non-OSA AHI<5, mild apnea 5<AHI<15, moderate 15<AHI<30, and severe AHI>30. A chi-squared test checked the calibrated data to see if there was any difference in the data distribution between AHI with categorical variables to get the p-value. Categorical variables are gender, diabetes, hypertension, cardiovascular disease, and smoking. The target variable, OSA severity, is evaluated in association with the categorical variable. A t-test was used to test any difference in the data distribution between AHI with continuous variables. Continuous variables are age, BMI, and waist circumference. The target variable was tested in association with the continuous variable to get the p-value.

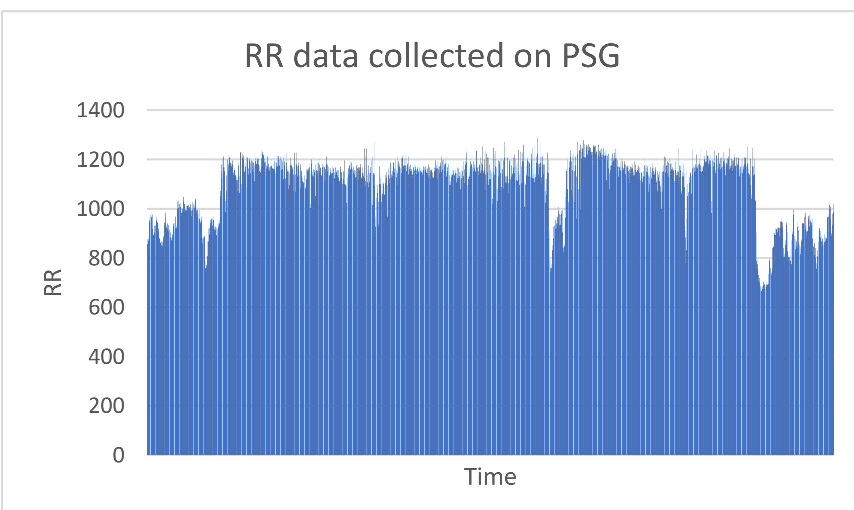

Figure 2: RR, breaths per minutes, data taken over time (every 9 seconds) by PSG; Subject ID: 200059; OSA: Severe; Date collected: 1/1/1985.

Furthermore, the data collected by the PSG included 31 variables, including the mean and standard deviation of RR (Figure 2), QT, QTc, and QTbc. QT interval is measurement used to assess some of the electrical properties of the heart. It is the time taken from when the cardiac ventricles start to contract to when they finish relaxing. QTc and QTbc are variations of QT. These variables were included in the final prediction model.

Name	**Description**	**P-value in association with AHI**
Body Mass Index (BMI)	Person's weight divided by the square of height. Correlates with waist circumference and hypertension	$p<0.01$
Age	A continuous variable ranging from 39 years to 90 years	$p>0.05$
Gender	Categorial variable with female and male.	$p<0.01$
Diabetes	Disease that occurs when your blood glucose, also called blood sugar, is too high. Categories include yes or no.	$p>0.05$
Waist Circumference	Measurement taken around the abdomen at the level of the umbilicus	$p<0.01$
Hypertension	Blood pressure reading of 130/80 mm Hg or higher. Categories include yes or no.	$p<0.01$
Cardiovascular Disease (CVDs)	A group of disorders of the heart and blood vessels like heart attack, stroke, etc. Categories include yes or no.	$p>0.05$
Smoking	Inhalation of the smoke of burning tobacco encased in between cigarettes, pipes, and cigars. Categories include never, or former, or current.	$p<0.05$

Table 1: Different variables used in prediction with description and p-value association with AHI.

BMI vs. AHI is shown to be statistically significant ($p<0.01$); this is supported by many studies stating BMI is an indicator of AHI severity. Age vs. AHI is shown to be statistically insignificant ($p>0.05$), showing it is not a good indicator for predicting AHI. Gender vs. AHI was statistically significant ($p<0.01$), this data states the gender is an indicator for AHI. However, gender should not be significant to AHI. This may be caused by the unfair distribution within people with no apnea, including females (n=120) and males (n=47). This trend continues with females being concentrated in lower AHI until mild apnea (AHI<15); however, a shift occurs at the moderate and severe apnea where the distribution shifts towards males. If data contained more participants, then the data could be less skewed. Diabetes vs. AHI is shown to be statistically insignificant ($p>0.05$); this does not match the previous analysis of BMI. Waist Circumference vs. AHI is statistically significant ($p<0.01$); however, this does match with the previous analysis on BMI but not diabetes, since waist circumference is positively correlated to BMI and diabetes. Even though previous studies have shown diabetes vs. waist circumference to be related, comparisons between diabetes and waist circumference were shown to be statistically insignificant ($p>0.05$). It is possible that separation between males and females would make the data significant to BMI. Hypertension vs. AHI is statistically significant ($p<0.01$); however, similarly to waist circumference, studies have shown hypertension correlates with diabetes. This is also shown in a comparison within the data between hypertension and diabetes, which showed it to be statistically significant ($p<0.01$). Since this relates to diabetes, it further proves there are issues with the data as inconsistencies continue to appear. Cardiovascular disease vs. AHI is statistically insignificant ($p>0.05$), therefore, it’s not a good indicator for AHI. Smoking vs. AHI is shown to be statistically significant ($p<0.05$). This is interesting and most likely caused by people who never smoked(n=376) and former smokers(n=346) taking the majority of the data while current smokers contain fewer participants (n=59) this causes the data to not be representative of people who smoke.

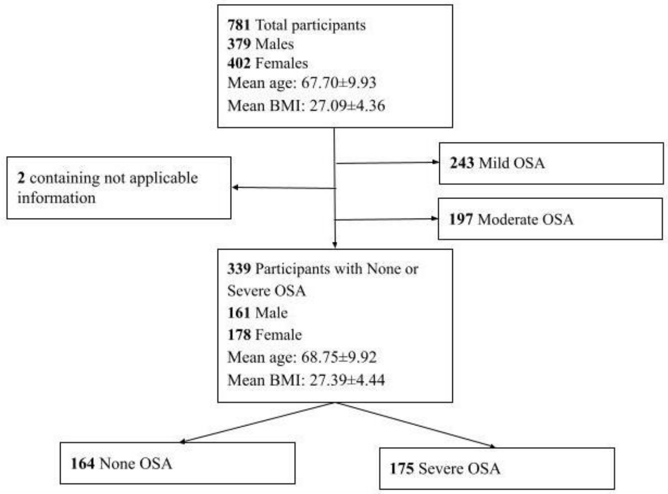

Figure 3: Flowchart to visually show filtering process.

Finally, the study filtered the participants. First, mild, and moderate OSA are not considered, because an important aspect of the machine learning model is predicting apnea in severe cases, since those people need to receive diagnoses and treatment quickly. Therefore, severe and no OSA will only be considered in the final model. Then, the data containing no data in certain machine parameters would not be considered in the final model.

Prediction model

RStudio, an integrated development environment for the programming language R that does statistical computing and graphics, was used to implement random forest to predict the target variable, OSA severity as AHI. Once the machine learning predicted the level of AHI, the overall accuracy was reported.

Alternative approaches

When the prediction accuracy was less than 80% a different method was looked at to predict the severity of OSA, the deep learning (DL) method, a subfield of machine learning that uses artificial neural network. DL uses multiple layers of interconnected nodes, each building upon the previous layer to refine and optimize the prediction. This method was created to mimic the human brain with a sequence containing data inputs, weights, and biases. Furthermore, with DL used to compare the study's results with existing results, long short-term memory (LSTM) method was used. LSTM is a method that is part of deep learning framework based on an artificial recurrent neural network (RNN). This network uses advanced RNN to store information indefinitely.

Results

The confusion matrix (Table 2) is a visualization of the performance of random forest. Each row represents the actual case, and the columns represent the predicted case. Overall, the TP and TN (n=252) show the correctly classified severity of OSA. While the FN and FP (n=87) show the incorrectly categorized severity of OSA. The prediction accuracy is calculated by adding TP and TN then dividing it by total participants in model.

	None (PP)	Severe (PN)	Class error
None (P)	115 (TP)	49 (FN)	0.2987805
Severe (N)	38 (FP)	137 (TN)	0.2171429

Table 2: Confusion matrix of prediction model. PP=predicted positive, PN=predicted negative, P=actual positive, N=actual negative, TP=true positive, FN=false negative, FP=false positive, TN=true negative

After running the random forest prediction model in R, the prediction accuracy was 74.34%. The goal of the model was to produce a prediction accuracy of greater to or equal to 80%.

Discussion

Significance

Even though it is not the anticipated 80%, it's still a good prediction model. We tried a different model using DL and LSTM network to see if it would produce greater prediction accuracy. However, DL proved to insufficiently predict OSA severity compared to RF model due to the lack of enough training data. The OSA severity prediction model can significantly decrease a patient's time spent in diagnosis and the patient's ability to receive treatment. As more and more people are able to receive treatment quicker, there will be a decrease in OSA-affected fatalities. Furthermore, this can shorten the diagnosis timeframe from a couple of weeks to just one day, as the model can make its predictions through a questionnaire and overnight PSG test. This quick analysis allows sleep specialists to diagnose and treat the patient quicker. Even though it has less than 80% accuracy, the model includes both females and males in the study. Therefore, this model allows for predictions to apply to males and females, which is not included in most other models.

Future Plans

In a future study, the parameter of demographics could be increased to include Epworth Sleepiness Scale (ESS), PaO_2 (mmHg), and $PaCO_2$ (mmHg). This expansion could allow greater precision and greater accuracy to the random forest model or deep learning method. Furthermore, instead of just looking at severe and no OSA, a model could be programmed to recognize mild and moderate OSA for a more precise model that helps doctors produce a tailored treatment for each patient. Furthermore, usable participants can be expanded to over 1000 to create a more accurate model.

Limitations

The overall population was 781 participants, but the data usable for the model decreased to 339. This could be the reason for the lower accuracy. Furthermore, the data from the participants was collected in the late 1990s. This could have caused some of the machine data to not be correct, leading to a more skewed prediction.

References

AlGhanim, N., Comondore, V. R., Fleetham, J., Marra, C. A., & Ayas, N. T. (2007). The Economic Impact of Obstructive Sleep Apnea. *Lung*, *1*, 7–12. https://doi.org/10.1007/s00408-007-9055-5

Arnardottir, E. S., Bjornsdottir, E., Olafsdottir, K. A., Benediktsdottir, B., & Gislason, T. (2015). Obstructive sleep apnoea in the general population: highly prevalent but minimal symptoms. *European Respiratory Journal*, *1*, 194–202. https://doi.org/10.1183/13993003.01148-2015

Lantz, B. (2019). *Machine Learning with R - Third Edition*. Packt Publishing.

Laudato, G., Oliveto, R., Scalabrino, S., Colavita, A. R., De Vito, L., Picariello, F., & Tudosa, I. (2020). Identification of R-peak occurrences in compressed ECG signals. 2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA). https://doi.org/10.1109/memea49120.2020.9137207

Mencar, C., Gallo, C., Mantero, M., Tarsia, P., Carpagnano, G. E., Foschino Barbaro, M. P., & Lacedonia, D. (2019). Application of machine learning to predict obstructive sleep apnea syndrome severity. *Health Informatics Journal*, *1*, 298–317. https://doi.org/10.1177/1460458218824725

Peppard, P. E., Young, T., Barnet, J. H., Palta, M., Hagen, E. W., & Hla, K. M. (2013). Increased Prevalence of Sleep-Disordered Breathing in Adults. *American Journal of Epidemiology*, *9*, 1006–1014. https://doi.org/10.1093/aje/kws342

Rodrigues, J. F., Pepin, J.-L., Goeuriot, L., & Amer-Yahia, S. (2020). An Extensive Investigation of Machine Learning Techniques for Sleep Apnea Screening. *Proceedings of the 29th*

ACM International Conference on Information & Knowledge Management. https://doi.org/10.1145/3340531.3412686

Rotenberg, B. W., George, C. F., Sullivan, K. M., & Wong, E. (2010). Wait Times for Sleep Apnea Care in Ontario: A Multidisciplinary Assessment. *Canadian Respiratory Journal*, *4*, 170–174. https://doi.org/10.1155/2010/420275

Rowley, J. A., Aboussouan, L. S., & Badr, M. S. (2000). The Use of Clinical Prediction Formulas in the Evaluation of Obstructive Sleep Apnea. *Sleep*, *7*, 929–938. https://doi.org/10.1093/sleep/23.7.929

Strollo, P. J., & Rogers, R. M. (1996). Obstructive Sleep Apnea. *New England Journal of Medicine*, *2*, 99–104. https://doi.org/10.1056/nejm199601113340207

Tsai, C.-Y., Liu, W.-T., Lin, Y.-T., Lin, S.-Y., Houghton, R., Hsu, W.-H., Wu, D., Lee, H.-C., Wu, C.-J., Li, L. Y. J., Hsu, S.-M., Lo, C.-C., Lo, K., Chen, Y.-R., Lin, F.-C., & Majumdar, A. (2021). Machine learning approaches for screening the risk of obstructive sleep apnea in the Taiwan population based on body profile. *Informatics for Health and Social Care*, 1–16. https://doi.org/10.1080/17538157.2021.2007930

Tsuiki, S., Nagaoka, T., Fukuda, T., Sakamoto, Y., Almeida, F. R., Nakayama, H., Inoue, Y., & Enno, H. (2021). Machine learning for image-based detection of patients with obstructive sleep apnea: an exploratory study. *Sleep and breathing*, *4*, 2297–2305. https://doi.org/10.1007/s11325-021-02301-7