Machine Learning Approaches to Predict Blood Pressure Level and Variability Using Polysomnography Data

---

A Thesis

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment
of the requirements for the degree

Master of Science

by

Gerun Lu

August 2019

APPROVAL SHEET

This Thesis
is submitted in partial fulfillment of the requirements
for the degree of
Master of Science

Author Signature: Glenn Lu

This Thesis has been read and approved by the examining committee:

Advisor: Hyojung Kang

Committee Member: Laura Barnes

Committee Member: Jennifer Lobo

Committee Member: _____

Committee Member: _____

Committee Member: _____

Accepted for the School of Engineering and Applied Science:

Craig H. Benson, School of Engineering and Applied Science

August 2019

Abstract

Blood pressure level has been studied as a medium between sleep apnea and cardiovascular disease because previous studies indicate that sleep apnea is associated with high blood pressure (BP) and high visit-to-visit blood pressure variability (BPV), both of which are risk factors of certain cardiovascular diseases. However, a limited number of studies have been conducted to predict high BP and visit-to-visit BPV using in-lab sleep study data, and the conclusions obtained in those studies are conflicting. This study's objective is to develop a predictive model for BP and BPV using machine learning methods based on in-lab sleep study data collected from the Sleep Disorders Center (Sleep Laboratory) at the University of Virginia (UVA). After pre-processing and combining the sleep data with other available patient data, including demographic and clinical information, a multi-step feature selection procedure was applied, resulting in eighteen variables for the next step. Furthermore, various machine learning models were developed, and their performances were compared. A multiple imputation method for dealing with missing data and feature reduction methods was studied during the process of developing models. The results indicate that this in-lab sleep study data can be employed to build a promising classification model for the high-BP group (systolic BP $\geqslant$130 mm Hg), although it is unable to predict high visit-to-visit BPV. The model's feature importance indicates that sleep-related features and blood oxygen saturation (SpO2)-related features are associated with high BP.

Acknowledgements

In the past two years at UVA, I have learned a considerable amount of new material and have become the person I am today. I shall take this opportunity to thank the individuals and organizations that have supported me throughout my graduate experience.

I would like to firstly thank my advisor, Dr. Hyojung Kang, for her invaluable guidance and encouragement through every stage of my academic career at UVA. I would have been unable to finish my thesis without her professional advice, detailed comments, and constant encouragement. She has taught me a great deal about the research process, information modeling, and health systems, but more importantly, she has taught me to be confident.

I would like to thank Dr. Laura Barnes and Dr. Jennifer Lobo of the Master's committee, as both have provided me with insightful comments and useful suggestions throughout the research process. From taking their courses on modeling and health systems at UVA to working and discussing the research project, they have offered me constant support.

Additionally, I would like to thank Dr. Min Woong Sohn, Dr. Younghoon Kwon, and Dr. Jeongok Logan for providing me with professional feedback and valuable conversations every time I presented in our group meeting. These individuals strengthened my faith in finishing the research project. I wish to thank Soojung Ahn, Xiaoyue Liu, and Elizabeth Harrison, my fellow graduate students and teammates, for their contributions to this study.

I would like to thank the Center for Engineering in Medicine for their financial support granted through a Seed Grants award. I would also like to thank the University of Virginia Healthcare System for their permission to use the data.

Moreover, I would like to extend my deepest gratitude to all my friends and faculty members at UVA. I feel lucky to have had so many supportive colleagues and mentors surrounding me and making UVA truly feel like home.

Table of Contents

List of Tables

List of Figures

Machine Learning Approaches to Predict Blood Pressure Level and Variability Using Polysomnography Data

1. Introduction

Sleep disorders have become a major health concern in the United States. It is estimated that 22 million Americans suffer from moderate to severe obstructive sleep apnea (OSA), which is widely recognized as a significant sleep problem. Many individuals have developed the habit of monitoring their sleep as wearables have become prevalent, although the data obtained from the wearables cannot replace a doctor's diagnostic process because wearables lack sensitivity and professionalism. Patients are often required to describe their sleep quality through questionnaires, but these self-reports are misleading when patients cannot answer questions accurately.

Polysomnography (PSG), also known as a sleep study, is required by physicians if they believe a patient is experiencing these sleep-related problems. During a sleep study, brainwaves, eye movement, breathing, blood oxygen level, heart rate, and muscle activity are monitored, and a final sleep report is thereby generated. This report is more informative and persuasive than is a sleep questionnaire.

Researchers anticipate that PSG data can interpret more about cardiovascular diseases than can self-reported questionnaires. Sleep-related problems typically do not develop independently, and hypertension is one complication that has been confirmed as being associated with sleep-disordered breathing (Harding, 2000). Systolic blood pressure (SBP) is the amount of pressure in the arteries during a heart muscle contraction that indicates hypertension and the risk of cardiovascular disease (Vasan et al., 2002). In addition to blood pressure (BP) itself, visit-to-visit SBP variability and maximum SBP are strong predictors of stroke and coronary events (Rothwell et al., 2010). As a result, the BP level can be studied as a medium between sleep apnea and other diseases.

Because sleep data obtained by different sleep labs and the available methods of collecting and representing BP levels vary significantly, the conclusions reached in

previous studies are conflicting. In addition, previous studies mainly used univariate analyses and logistic regression analyses when studying the association between sleep data and hypertension status.

This study's objective is to apply machine learning techniques to sleep data and interpret BP levels. The first objective is to build a prediction model that predicts BP levels by comparing different machine learning algorithms. We employed both hypertension status and visit-to-visit BPV as outcomes, while sleep data were established as predictors. Another objective is to apply different feature reduction methods and observe their impacts on the prediction results. The final objective is to compare the prediction results using the complete dataset and the multiple-imputed datasets.

2. Literature Review

Previous studies that utilized PSG data to investigate the association between hypertension and sleep apnea employed mixed outcome variables and thus generated differing results. Appleton et al. (2016) conducted analyses of hypertension and AHI in undiagnosed sleep apnea patients, wherein they defined hypertension as SBP >= 140 mm Hg, diastolic BP >= 90 mm Hg, and/or medication use, thus determining prevalent hypertension as being significantly associated with AHI during REM sleep (Appleton et al., 2016). Dean et al. (2013) performed a systematic assessment to determine an association between indices of sleep breathing disturbances, sleep fragmentation, sleep duration, hypoxemia, limb movements, and sleep stage distribution with both SBP and diastolic blood pressure (DBP). The BP values were collected once and adjusted according to medication use with an imputation method. The linear model the authors built demonstrates that the aroused AHI and limb movements were consistently associated with elevations in BP (Dean et al., 2013). Blood pressure variability (BPV) has also been studied in patients with OSA. Marrone and Bonsignore (2018) focused on very short-term BPV in OSA patients and determined that high BPV is common in patients with OSA. Further, OSA treatment slightly reduced their participants' mean 24-hour BP levels and nocturnal beat-by-beat BPV (Marrone & Bonsignore, 2018).

Methods for detecting the association between BP and sleep data are currently limited to observational studies and basic models, such as univariate and multivariate analyses. Shiina et al. (2016) performed a multiple logistic regression on AHI and other related variables to evaluate their association with visit-to-visit BPV and reported a significant association between OSA and visit-to-visit BPV. Dean et al. (2013) used a cross-validated linear regression model to assess the association between sleep data and BP. Tanigawa et al. (2004) conducted a cross-sectional survey-based study among Asian populations and employed both a multiple linear regression and a logistic regression to examine the relationship between sleep-disordered breathing and BP levels.

An overview of machine learning approaches to medical diagnoses revealed that

machine learning is well suited for analyzing medical data, especially medical diagnosis works. A sound machine learning system should perform sufficiently and be able to appropriately deal with missing and/or noisy data when diagnosing a patient (Kononenko, 2001). Machine learning is also widely used in sleep-related fields. Khalighi et al. (2013) applied support vector machines in their study for sleep–wake detection and multiclass sleep staging, while Hassan (2016) studied the following nine well-known classifiers to determine the most efficient classification model for detecting automated sleep apnea: naive Bayes, kNN, neural network, AdaBoost, bagging, random forest, extreme learning machine, discriminant analysis, and restricted Boltzmann machine. Xiao et al. (2013) employed random forest as a new classifier for sleep stage classification, while Araujo et al. (2018) applied machine learning methods, including random forest and XGBoost, to predict which patients would be likely to cease their CPAP treatment with these models.

It is of concern that sleep data, together with demographic and other clinic variables, result in a highly dimensional dataset. The feature reduction approaches have been discussed in papers that involve PSG reports. Dean et al. (2013) employed the dataset collected by MESA, a multicenter prospective study involving participants from six sites, and eventually reduced more than two thousand dataset features to eight with a literature variable recommendation and a cluster analysis. Zinchuk et al. (2018) conducted an OSA evaluation on more than two thousand veterans from three veteran affairs centers using a two-step feature reduction method. Principal component-based clustering was firstly applied, while they secondly retained >75% variance with each domain. The variable amount was reduced from 65 to 17.

3. Data and Method

3.1 Data Pre-Processing

To predict BP levels and compare different machine learning techniques, we included a complete dataset with 2,742 observations and 18 variables. Multiple-imputed datasets with 3,002 observations were additionally created for the analyses.

The original sleep studies obtained from the Sleep Disorder Center at UVA comprise 7,494 sleep studies, of which we firstly excluded duplicate and home studies. To maintain our data's consistency and effectiveness, we required that our sleep data follow the same rule and remain consistent with other information. As a result, solely 1B-rule and in-lab studies from October 2010 to September 2015 were kept; of these patients, we included those who had at least three BP readings following the sleep study documented in their records.

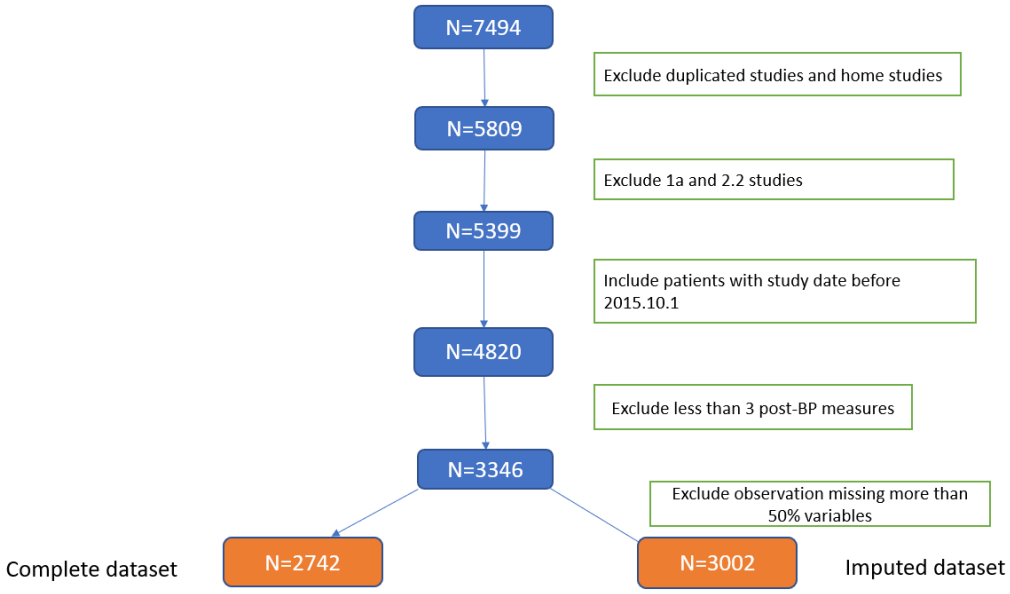

Figure 3-1. *Filter Process of the Sleep Studies*

We selected variables for our dataset by summarizing 37 sleep variables in 4 categories as AHI-related, blood oxygen saturation (SpO2)-related, sleep time-related, and sleep change-related. Demographic and clinical information (including BP) was retrieved from the Clinical Data Repository for patients who had participated in sleep

studies. Before applying feature selection techniques of machine learning, we went through a two-step feature selection process. Firstly, 37 sleep-related variables in 4 different categories and 8 selected demographic or clinical variables were checked for missing and overlapping data, and we also combined some suggestions from doctors regarding the medical importance of certain variables, such as the main sleep position and essential sleep stages.

Table 3-1. *Description of Selected Variables for Developing Models*

Category	Variables	Description
Demographic	sex	Female, male
	age	age
Clinical	bmi	Body mass index
Sleep time	sleep_time	Total sleep time
	supine_time	Total supine sleep time
	sleep_onset_time	Time to first sleep onset
	first_rem_time	Time to first REM sleep
	n3_sleep	Stage N3 sleep (%)
	rem_sleep	Stage REM sleep (%)
Sleep change	plm_arousal	Periodic limb movements per hour of sleep associated with arousal
	spontaneous_arousals	Spontaneous arousal per hour of sleep
	awakenings	Awakenings during sleep
	stage_changes	Total sleep stage changes
AHI	ahi_apnea	Apnea per hour
	ahi_hypopneas	Hypopneas per hour
	ahi_supine	AHI during supine time
SpO2	spo2_mean	Mean oxyhemoglobin saturation during sleep
	spo2_min_rem_nrem	Minimum oxyhemoglobin saturation during sleep

We maintained at least one variable in each category based on missing data, overlapping data, and doctor recommendations. Then, we checked correlations among these variables and removed one variable of each highly correlated pair (see Appendix A). The original 45 variables were reduced to 18, while at least 1 remained for each sleep category (Table 3-1). We subsequently filtered the original dataset based on the selected variables and obtained a complete dataset of 2,742 observations without any missing variables (see Figure 3-1). The dataset was then split into 80% training data and 20% test data. This study's comprehensive evaluation was performed on this test dataset.

At the same time, we created a new dataset using a multiple imputation method. In addition to 2,193 observations in the training dataset, 260 observations achieved less than 50% missing variables and were thus added to the original training dataset. We then filled in the missing data using multiple imputation by chain equations (MICE) and generated multiple complete datasets; this process was completed with the R "mice" package (Azur, Stuart, Frangakis, Leaf, & Washington, 2011). For this imputation method, missing data are modeled according to other variables and their distribution. For example, binary variables are modeled with logistic regression, while continuous variables are modeled with linear regression. Missing data are updated iteratively, and the process is repeated until convergence of models (Azur, Stuart, Frangakis, Leaf, & Washington, 2011). We compared the results using both the complete and imputed datasets.

To detect high BP and high visit-to-visit BPV, we compared different thresholds and calculation methods and then developed standards to divide patients into a high-BP group and a low-BP group as well as a high-BPV group and a low-BPV group. Before calculating the outcome variables, we conducted the following process to adjust BP measures for accuracy: (1) the period time was restricted, and BP measures within one year following the sleep study date were exclusively kept; (2) outliers were removed, and SBP between 50 and 300 as well as DBP between 30 and 180 were exclusively considered to be valid; (3) daytime and outpatient control was implemented, wherein all inpatient measures between an inpatient's admittance and discharge dates were dropped, thus allowing that all BPs be measured during the daytime; and (4) the number

of measures were controlled in that the average of all BP measures taken in one day was calculated as the only record for that day.

To decide the BP level, we took both the mean and median of one patient's total measures, implemented 130, 140, 150, and 160 mm Hg as cutoffs to distinguish the high-BP and low-BP groups, and compared the results of logistic regression models to determine the most efficient cutoff value. We treated hypertension medication use in two different ways: by either adjusting BP classification to reclassify patients who took hypertension medication and were initially in the low-BP or by employing hypertension medication as a predictor in our models. The results of these adjustments were compared using machine learning models.

Visit-to-visit BPV was not calculated if the number of BP measures was less than three; if the number was greater than three, then solely three measures were randomly chosen for BPV calculation. If the coefficient variation of the selected measures was less than ten, then it was classified as low BPV, whereas a variation greater than or equal to ten resulted in a high-BPV classification.

3.2 Machine Learning Application

We employed three machine learning classifiers to detect a high-BP group and a high-BPV group: a logistic regression model (LM), random forest (RF), and the eXtreme Gradient Boosting (XGBoost) model. LM is the most traditional classifier and the most informative method, as it can determine to what extent a predictor is associated with the outcome as well as the direction of that association. RF is another popular machine learning classifier because it employs many techniques to reduce variance and avoid overfitting; however, due to its lower variance, the model may have a higher bias. XGBoost uses a gradient descent algorithm to minimize loss when adding new models, and models are sequentially added until no further improvements can be made.

As described in Section 3.1, we split the complete dataset into an 80% training and a 20% testing dataset. We trained and developed each model using a training dataset based on five-fold cross-validation. The most effective model was then evaluated using the test dataset. Performance metrics, such as accuracy, recall, precision, F1-score, and

the area under the receiver operating characteristics curve (AUC), were calculated to evaluate the classification models, which were compared with the overall consideration of accuracy, F1-score, and AUC.

Although we selected some variables from the original PSG reports, eighteen yet remained. Reducing variables can improve the speed of building each model and thus render each model easier to interpret without losing the data's predictive power. We employed two variable reduction methods and compared their performances with the original variable sets. LASSO, which is one method that aims to select a small subset of features that minimize redundancy and maximize relevance to the class labels, is based on the ℓ1-norm of the linear model's coefficient (Aggarwal, Kong, Gu, Han, & Yu, 2014). Principle component analysis (PCA), a feature extraction method, projects original features into lower dimensional space and creates new features rather than using original features themselves. We repeated the dataset training with reduced variables and evaluated the models on the same test dataset.

4. Results and Discussion

4.1 Descriptive Statistics

Before building our prediction models, we calculated descriptive statistics of the complete dataset and generated an overall understanding of our own dataset. According to the descriptive table (Table 4-1), the sample was generally middle-aged (mean=52.95 years, SD=14.32) and overweight, according to BMI (mean=35.69 kg/m^2, SD=8.96), and 63.5% of all respondents were female. This group of patients has a higher risk for developing hypertension and cardiovascular disease than the average person due to the former's age and BMI. The patients' amount of sleep averaged about 6 hours, while their supine time averaged about 2.5 hours. AHI, which is the sum of ahi_apnea and ahi_hyponea, averaged higher than 15 events per hour. This AHI is above the normal range and indicates moderate sleep apnea. Therefore, many of these patients were potentially suffering from sleep apnea and had a high risk of developing hypertension.

Table 4-1. *Statistics Table of the Sleep Study Participants*

Category	Variables	Statistics	
Demographic	sex	Female	1715
		Male	1027
	age(year)	52.93±14.32	
Clinical	bmi(kg/m2)	35.69±8.96	
Sleep time	sleep_time(minutes)	355.2±71.23	
	supine_time(minutes)	154.06±108.9	
	sleep_onset_time(minutes)	26.22±27.32	
	first_rem_time(minutes)	126.8±79.40	
	n3_sleep(%)	11.34±9.60	
	rem_sleep(5)	17.20±8.37	
Sleep change	plm_arousal	2.44±7.22	
	spontaneous_arousals	5.83±5.30	
	awakenings	112±94.9	
	stage_changes	87.35±39.44	
AHI	ahi_apnea(events per hour)	9.75±13.54	
	ahi_hypopneas(events per hour)	6.06±7.50	
	ahi_supine(events per hour)	26.65±27.90	
SpO2	spo2_mean(%)	94.53±2.48	
	spo2_min_rem_nrem(%)	84.86±6.87	

4.2 Prediction Model of Blood Pressure

The prediction model's outcome variable is BP level, and our first step was to find the high BP level our data could predict. We compared different thresholds (130, 140, 150, 160 mmHg) and measures (mean, median) for classifying high-BP patients (HBP) and low-BP patients (LBP). Hypertension medication was used for adjustment of BP levels as described in Section 3. The comparison of the predictions' accuracy, F1-score, and AUC based on the LM is summarized in Table 4-2 (see Appendix B for the full comparison).

According to this cross-sectional analysis, designating the HBP group using the median of all BP measures and the threshold of 130 mm Hg yields the most favorable performance of AUC=0.75. We additionally found that reclassifying based on hypertension medication use always yields a more favorable performance than does treating hypertension medication use as a variable. This standard (with a median, 130 as threshold, and by reclassifying hypertension medication use) for defining HBP exhibits a stable performance across all classifiers (see Appendix B for performance on other classifiers).

Table 4-2. *Comparison of Different Classification Standards Based on Accuracy, F1-Score, and AUC*

Threshold	Hypertension medication use	Measure	High%	Accuracy	F1-score	AUC
130	Variable	Mean	49	0.60	0.58	0.63
130	Reclassification	Mean	74	0.67	0.75	0.72
130	Variable	Median	49	0.60	0.59	0.64
130	Reclassification	Median	75	0.66	0.75	0.75
140	Variable	Mean	22	0.59	0.37	0.61
140	Reclassification	Mean	63	0.68	0.74	0.70
140	Variable	Median	22	0.60	0.36	0.62
140	Reclassification	Median	63	0.67	0.72	0.71

We built prediction models for the BP levels generated by this standard. 75% of patients were in the high-BP group. This high percentage is likely due to the fact that many patients who participate in sleep studies have had prior health concerns. We

achieved an unbalanced outcome dataset and performed over-sampling, under-sampling, and SMOTE sampling methods to overcome the problem. Unlike over-sampling and under-sampling, both of which randomly select studies from an existing dataset, SMOTE sampling creates a new and synthetic observation method for modeling. In addition to using the entire variable set for prediction, we generated a subset of thirteen variables with LASSO and a new variable set of eight constructed factors with a PCA. With all these techniques aimed toward improving the model, the most favorable performance models are summarized in Table 4-3. The first two components obtained by the PCA are illustrated in Figure 4-1. Merely 29.2% of the variance was captured by the first two principal components, and thus more components should be included when developing models.

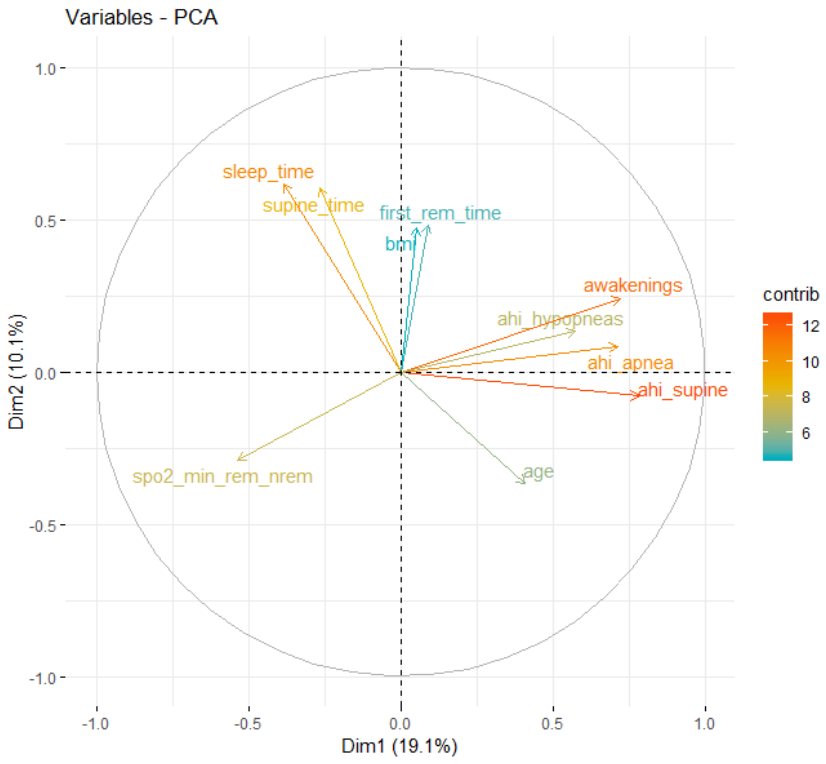

Figure 4-1. *The First Two Components Obtained by the PCA*

As indicated in Table 4-3, models built with LASSO-selected variables generally performed more favorably than did models built with PCA-constructed features. Compared to models that used all variables, the LASSO feature selection method

helped exclude variables (e.g., plm_arousal) and awakenings in this situation without impairing performance. The table also indicates that the RF model achieved high accuracy and a high F1-score; thus, with an AUC greater than 0.7, the model was desirable. Over-sampling provided comprehensively more favorable results than did the other sampling methods. Over-sampling expanded the dataset for training when we had a relatively small dataset.

Table 4-3. *Comparison Table of Different Models and Feature Reduction Methods*

Model	Variables	Sampling	Accuracy	F-score	AUC
LM	All	Over	0.66	0.75	0.75
LM	All	Under	0.67	0.75	0.74
RF	All	Over	0.78	0.86	0.71
XGBoost	All	Over	0.71	0.79	0.73
LM	Lasso Selected	Over	0.66	0.75	0.75
LM	Lasso Selected	Under	0.67	0.75	0.75
RF	Lasso Selected	Over	0.78	0.87	0.70
XGBoost	Lasso Selected	Over	0.67	0.76	0.74
LM	PCA constructed	Over	0.63	0.72	0.67
LM	PCA constructed	Under	0.62	0.70	0.67
RF	PCA constructed	Over	0.73	0.83	0.61
XGBoost	PCA constructed	Over	0.65	0.75	0.66

The implementation of the MICE procedure resulted in imputed datasets, which created 260 additional complete observations for training. The models trained with the complete and imputed datasets are displayed in Table 4-4. The most favorable performance of a model trained with the complete dataset was that of RF, with an accuracy of 0.78 and an AUC of 0.71. The accuracy and AUC were increased by 0.01 for the same model upon employing the imputed dataset. In general, the results were similar after including the imputed dataset, which indicated consistency between the original and imputed values. Therefore, the MICE procedure is an appropriate method for filling in missing values when the number of complete cases is low.

The feature contributions are presented in Table 4-5. For the LM, the significance level of 0.05 provides predictors that are statistically, significantly associated with the

response variable, while the odds ratio of 1 indicates the association's direction. RF provides an important measure for each variable during the training process, and two measures assess this importance: Gini importance and mean decrease in impurity. We merged the top ten most important for both measures to determine the RF model's important features. The feature importance of XGBoost is represented by "gain," which is the improvement in accuracy contributed by each feature to each tree in the model. The top five variables were then selected for comparison.

Table 4-4. *Comparison Table of Models Trained with an Imputed Dataset and a Complete Dataset*

Dataset	Model	Sampling	Accuracy	F-score	AUC
Complete	LM	Over	0.66	0.75	0.75
Complete	LM	Under	0.67	0.75	0.74
Complete	RF	Over	0.78	0.86	0.71
Complete	RF	SMOTE	0.68	0.77	0.68
Complete	XGBoost	Over	0.71	0.79	0.73
Complete	XGBoost	SMOTE	0.68	0.77	0.68
Imputed	LM	Over	0.65	0.73	0.74
Imputed	LM	Under	0.64	0.72	0.74
Imputed	RF	Over	0.79	0.87	0.72
Imputed	RF	SMOTE	0.68	0.77	0.70
Imputed	XGBoost	Over	0.70	0.78	0.74
Imputed	XGBoost	SMOTE	0.68	0.76	0.73

As depicted in Table 4-5, age and BMI are important across the models and should be adjusted when assessing the association between sleep data and BP level. In addition, sleep time-related features are highly important, namely with regard to the distribution of sleep stages (e.g., such n3_sleep and rem_sleep). According to the LM, these features are negatively related to BP, and thus a sufficient amount of sleep during the N3 and REM stages is important. However, rem_sleep is not significant when using the

imputed dataset; on the contrary, the sleep change-related variable of awakenings is positively related to BP, which is something we must have control over in future patients. SpO2-related variables have been demonstrated as being more relative to BP level than are AHI-related variables. We determined that, while the mean SpO2 during sleep is positively related to the high-BP group, the minimum SpO2 during sleep is negatively related to that group. We may doubt that a sudden drop in SpO2 during sleep poses a greater risk than does the maintenance of a relatively low SpO2 level at all times.

Table 4-5. *Feature Significance and Importance Comparison Across Models*

	Complete			Imputed		
Variables	LM	RF	XGBOOST	LM	RF	XGBOOST
sex(male)	√+					
Age	√+	√	√	√+	√	√
Bmi	√+	√	√	√+	√	√
sleep_time					√	
supine_time		√				
awakenings	√+			√+		
stage_changes	√+					
sleep_onset_time		√			√	
first_rem_time		√			√	
n3_sleep	√-			√-		√
rem_sleep	√-	√	√			√
plm_arousal						
spontaneous_arousals						
ahi_apnea					√	
ahi_hypopneas						
ahi_supine			√			
spo2_mean	√+					
spo2_min_rem_nrem	√-		√	√-		√

4.3 Prediction Model of Blood Pressure Variability

The method of developing a prediction was also applied to predict BPV. Table 4-6 displays the models' performances with BPV as the response variable. AUC values around 0.6 indicate that these models are barely satisfactory for providing classification results for BPV. Although we have increased the number of BP measures for calculating BPV to 5, the evaluation metrics did not achieve desirable values, for which one reason

was that we trained a smaller dataset due to our stricter rules. Unlike BP level, BPV is difficult to adjust with hypertension medication use.

Table 4-6. *Models' Performances When Predicting Visit-to-Visit BPV*

Model	Number of Measures	Accuracy	F-score	AUC
LM	3	0.55	0.35	0.55
RF	3	0.77	0.09	0.60
XGBoost	3	0.62	0.32	0.59
LM	5	0.64	0.20	0.62
RF	5	0.64	0.27	0.62
XGBoost	5	0.61	0.52	0.62

5. Conclusion

The present study poses three limitations, the first of which are the participants. Sleep studies were performed on the group of individuals who had previous health concerns as well as higher-than-normal BP. We involved patients who had either diagnosed or undiagnosed hypertension. This bias in patients weakened the model's predictive value. The second limitation is the reasons for high BP. Although the prediction model can predict that a certain patient has undiagnosed hypertension, the model cannot provide the cause; the reasons why a patient is likely to experience hypertension must be investigated by physicians. The third limitation is the data size. We excluded many sleep studies in consideration of their rules and dates. In the future, we can improve this study in many ways. If we include more data, we can exclusively focus on patients who have undiagnosed hypertension, and we can additionally increase the number of readings for calculating BPV. We may thereby extend our study to not only analyze BP, but also to predict cardiovascular events.

The analyses and comparisons discussed in this paper develop a technique for building a BP level prediction model using PSG data from the UVA Sleep Lab. According to the prediction models' performances, we determined that the PSG data from our Sleep Lab can be used to predict high BP levels; however, our model that predicts high visit-to-visit BPV exclusively provides moderate results. Although a PSG report is not decisive for diagnosing hypertension or other cardiovascular diseases, it provides some valuable clinical information for physicians and patients alike.

Thirty-seven variables from sleep studies in addition to demographic and clinical variables were organized in several clusters, and eighteen variables remained following a two-step feature exclusion process. Furthermore, feature reduction methods were used, for which the results indicate that LASSO is a more appropriate auto-feature selection method than is PCA when selecting representative features and speeding up the training process for this particular dataset.

Comparisons for deciding high BP levels indicate that the prediction model performs best when designating the high-BP group using the median of all BP measures

above 130 mm Hg and reclassifying BP levels with hypertension medication use. This result is inconsistent with the standard, which set 140mm Hg as the cutoff in earlier studies (Appleton et al., 2016; Dean et al., 2013). Many patients' SBP levels fall within the 130 mm Hg–140 mm Hg range, and our model suggests that PSG data can predict low-stage hypertension within this range. In addition to patients with low-stage hypertension, the model can differentiate patients who have developed hypertension and taken hypertension medicine.

The comparison between the complete dataset and the imputed dataset supports the idea of including additional observations with the MICE procedure when an acceptable number of missing variables are present in a sleep study. Due to these possibly missing values during the data collection and extraction procedure, MICE offers an option that avoids wasting these studies when training the prediction model and rather filling them up based on other variables.

With regard to this study's feature importance results, several features might have contributed to high BP. Age and BMI are two risk factors that must be considered when assessing BP; the distribution of different sleep stages—namely the percentage of REM sleep—is associated with BP level. The mean and minimum SpO2 levels are associated with BP but exhibited opposite effects, which might indicate the risk of a sudden drop in SpO2. Furthermore, the AHI-related variables were not proven to have strong links with BP in any prediction model although nevertheless contributed to some models.

Appendix A

Variable group	Variable name	Description	# of NAs	# of 0s	Include	Reason
AHI	ahi	Apnea-Hypopnea Index (AHI)	7	42	No	Correlation:ahi=ahi_apnea+ahi_hypopnea
	ahi_obstructive	Obstrucrive per hour	7	344	No	Correlation:ahi_apnea=ahi_obstructive+ahi_central+ahi_mixed
	ahi_central	Central per hour	7	1907	No	Correlation:ahi_apnea=ahi_obstructive+ahi_central+ahi_mixed
	ahi_mixed	Mixed per hour	7	2603	No	Correlation:ahi_apnea=ahi_obstructive+ahi_central+ahi_mixed
	ahi_apnea	Apnea per hour	6	256	Yes	Correlation:ahi=ahi_apnea+ahi_hypopnea
	ahi_hypopneas	Hypopneas per hour	6	312	Yes	Correlation:ahi=ahi_apnea+ahi_hypopnea
	ahi_rera	RERA per hour	13	1845	No	# of 0s >20%
	ahi_supine	AHI during supine time	272	181	Yes	Supine is the main sleep position
	ahi_prone	AHI during prone time	3049	106	No	Missing caused by no prone time during sleep, # of missing+0s>20%
	ahi_left	AHI during left time	777	491	No	Missing caused by no left time during sleep, # of missing+0s>20%
	ahi_right	AHI during right time	709	462	No	Missing caused by no right time during sleep, # of missing+0s>20%
	ahi_rem	AHI during rem sleep	223	193	No	Physician recommendation as less important
	ahi_nrem	AHI during non-rem sleep	13	137	No	Physician recommendation as less important
SpO2	spo2_mean	Mean oxyhemoglobin saturation	5	0	Yes	Oxygen levels are related to sleep apnea..
	spo2_min_rem	Minimum oxyhemoglobin saturation in REM	223	3	No	Correlation: spo2_min_rem_nrem=min(spo2_min_rem,spo2_min_nrem)
	spo2_min_nrem	Minimum oxyhemoglobin saturation in NREM	13	9	No	Correlation: spo2_min_rem_nrem=min(spo2_min_rem,spo2_min_nrem)

Variable group	Variable name	Description	# of NAs	# of 0s	Include	Reason
	spo2_min_rem_nrem	Minimum oxyhemoglobin saturation in Rem and NREM	2	0	Yes	Correlation: spo2_min_rem_nrem=min(spo2_min_rem,spo2_min_nrem)
	spo2_level	Level (less than xx%) of time spent with an oxygen separation	1595	0	No	Indicator of a Constant level
	spo2_minutes	Time spent with an oxygen separation of less than xx%	1595	542	No	# of missings + 0s >20%
Sleep time	recording_time	Total recording time	1	0	No	Lab study offset
	sleep_time	Total sleep time	1	4	Yes	efficiency=sleep_time/recording_time.
	supine_sleep	supine sleep(%)	10	260	No	Highly correlated to supine_time
	supine_time	Total supine sleep time	10	260	Yes	Supine is the main sleep position
	efficiency	Sleep efficiency (%)	1	4	Yes	efficiency=sleep_time/recording_time.
	sleep_onset_time	Time to first sleep onset	10	41	Yes	Physician recommendation as important
	first_rem_time	Time to first REM	215	0	Yes	Related to rem_sleep. Take sleep_time for NAs.
	n1_sleep	Stage N1 sleep (%)	21	65	No	N1+N2+N3+REM=100.N3 and REM are related to sleep apnea.
	n2_sleep	Stage N2 sleep (%)	5	3	Yes	N1+N2+N3+REM=100.N3 and REM are related to sleep apnea.
	n3_sleep	Stage N3 sleep (%)	6	666	Yes	N1+N2+N3+REM=100.N3 and REM are related to sleep apnea.
	rem__sleep_minutes	Rem time	5	211	No	Highly correlated to rem_sleep
	rem_sleep	REM (%)	5	211	Yes	N1+N2+N3+REM=100.N3 and REM are related to sleep apnea.
Sleep changes	resp_arousals	Respiratory arousals per hour sleep	11	156	No	Highly correlated to awakenings
	spontaneous_arousals	Spontaneous arousal per hour of sleep	11	50	Yes	Important factor of sleep changes
	plm_arousal	Periodic limb movements per hour	11	1924	Yes	Important factor of sleep changes

Variable group	Variable name	Description	# of NAs	# of 0s	Include	Reason
		of sleep associated with arousal				
	plm_nonarousal	Periodic limb movements per hour of sleep NOT associated with arousal	3346	0	No	No records
	awakenings	Awakenings	5	5	Yes	awakenings/sleep_time=sum of arousals
	stage_changes	Total stage changes	2	3	Yes	Important factor of sleep changes
Demographic	age	Age	0	0	Yes	Important factor of hypertension
	sex	Sex	41	0	Yes	Important factor of hypertension
Clinical	wt	Weight	70	0	No	Highly correlated to BMI
	bmi	Body mass index	70	0	Yes	Important factor of hypertension

Appendix B

				LM					RF					XGBoost(auc)				
Cut off	**Hypertension medication use**	**Mean or Median**	**High Ratio**	**Accuracy**	**F-score**	**F-score (Low)**	**Average F-score**	**AUC**	**Accuracy**	**F-score**	**F-score (Low)**	**Average F-score**	**AUC**	**Accuracy**	**F-score**	**F-score (Low)**	**Average F-score**	**AUC**
130	Not considered	Mean	49%	0.58	0.54	0.61	0.57	0.63	0.58	0.57	0.58	0.58	0.61	0.58	0.57	0.59	0.58	0.63
	Variable	Mean	49%	0.60	0.58	0.62	0.60	0.63	0.56	0.55	0.57	0.56	0.60	0.58	0.56	0.60	0.58	0.62
	Reclassification	Mean	74%	0.67	0.75	0.51	0.63	0.72	0.66	0.74	0.51	0.63	0.71	0.69	0.77	0.52	0.65	0.72
	Not considered	Median	49%	0.61	0.60	0.62	0.61	0.63	0.55	0.58	0.52	0.55	0.58	0.58	0.58	0.58	0.58	0.61
	Variable	Median	49%	0.60	0.59	0.61	0.60	0.64	0.58	0.61	0.55	0.58	0.59	0.59	0.60	0.58	0.59	0.63
	Reclassification	Median	75%	0.66	0.75	0.50	0.62	0.75	0.65	0.74	0.49	0.61	0.72	0.71	0.79	0.52	0.65	0.73
140	Not considered	Mean	22%	0.59	0.35	0.70	0.53	0.60	0.55	0.38	0.64	0.51	0.59	0.61	0.34	0.72	0.53	0.61
	Variable	Mean	22%	0.59	0.37	0.69	0.53	0.61	0.54	0.39	0.62	0.51	0.60	0.62	0.37	0.73	0.55	0.62
	Reclassification	Mean	63%	0.68	0.74	0.58	0.66	0.70	0.64	0.70	0.55	0.63	0.68	0.69	0.76	0.57	0.66	0.70
	Not considered	Median	22%	0.59	0.35	0.70	0.52	0.60	0.58	0.38	0.69	0.53	0.58	0.63	0.34	0.74	0.54	0.60
	Variable	Median	22%	0.60	0.36	0.71	0.53	0.62	0.57	0.36	0.68	0.52	0.58	0.64	0.35	0.75	0.55	0.63
	Reclassification	Median	63%	0.67	0.72	0.60	0.66	0.71	0.64	0.69	0.56	0.62	0.68	0.66	0.72	0.55	0.64	0.69
150	Not	Mean	7%	0.64	0.21	0.77	0.49	0.66	0.51	0.17	0.65	0.41	0.64	0.59	0.20	0.73	0.46	0.69

				LM					RF					XGBoost(auc)				
Cut off	**Hypertension medication use**	**Mean or Median**	**High Ratio**	**Accuracy**	**F-score**	**F-score (Low)**	**Average F-score**	**AUC**	**Accuracy**	**F-score**	**F-score (Low)**	**Average F-score**	**AUC**	**Accuracy**	**F-score**	**F-score (Low)**	**Average F-score**	**AUC**
	considered																	
	Variable	Mean	7%	0.63	0.22	0.72	0.47	0.68	0.56	0.19	0.69	0.44	0.66	0.71	0.23	0.82	0.52	0.71
	Reclassification	Mean	58%	0.65	0.67	0.59	0.63	0.68	0.65	0.70	0.58	0.64	0.69	0.66	0.72	0.58	0.65	0.69
	Not considered	Median	7%	0.64	0.15	0.77	0.46	0.57	0.56	0.14	0.70	0.42	0.55	0.68	0.15	0.80	0.47	0.58
	Variable	Median	7%	0.65	0.16	0.78	0.47	0.61	0.56	0.14	0.71	0.42	0.60	0.56	0.18	0.70	0.44	0.65
	Reclassification	Median	58%	0.64	0.67	0.61	0.64	0.68	0.66	0.70	0.60	0.65	0.67	0.66	0.72	0.56	0.64	0.69
160	Not considered	Mean	3%	0.63	0.07	0.77	0.42	0.61	0.48	0.04	0.64	0.34	0.61	0.68	0.06	0.82	0.44	0.58
	Variable	Mean	3%	0.65	0.07	0.78	0.43	0.64	0.49	0.04	0.65	0.34	0.57	0.86	0.07	0.92	0.50	0.59
	Reclassification	Mean	57%	0.67	0.70	0.64	0.67	0.72	0.68	0.72	0.62	0.67	0.72	0.68	0.72	0.63	0.68	0.72
	Not considered	Median	3%	0.60	0.05	0.75	0.40	0.57	0.48	0.05	0.65	0.35	0.58	0.83	0.06	0.90	0.48	0.66
	Variable	Median	3%	0.62	0.06	0.76	0.41	0.62	0.53	0.07	0.68	0.37	0.58	0.61	0.08	0.75	0.41	0.71
	Reclassification	Median	57%	0.66	0.69	0.61	0.65	0.71	0.67	0.72	0.59	0.66	0.71	0.67	0.73	0.59	0.66	0.71

Reference

Aggarwal, C. C., Kong, X., Gu, Q., Han, J., & Yu, P. S. (2014). Active learning: A survey. *Data Classification: Algorithms and Applications*, 571–605. https://doi.org/10.1201/b17320

Appleton, S. L., Vakulin, A., Martin, S. A., Lang, C. J., Wittert, G. A., Taylor, A. W., … Catcheside, P. G. (2016). Hypertension Is Associated With Undiagnosed OSA During Rapid Eye Movement Sleep. *CHEST*, *150*(3), 495–505. https://doi.org/10.1016/j.chest.2016.03.010

Araujo, M., Bhojwani, R., Srivastava, J., Kazaglis, L., & Iber, C. (2018). ML Approach for Early Detection of Sleep Apnea Treatment Abandonment. https://doi.org/10.1145/3194658.3194681

Azur, M. J., Stuart, E. A., Frangakis, C., Leaf, P. J., & Washington, D. C. (2011). Multiple imputation by chained equations : what is it and how does it work ?, *20*(1), 40–49. https://doi.org/10.1002/mpr

Dean, D. A., Wang, R., Jacobs, D. R., Duprez, D., Punjabi, N. M., & Zee, P. C. (2013). ASSOCIATION OF PSG INDICES WITH BLOOD PRESSURE A Systematic Assessment of the Association of Polysomnographic Indices with Blood Pressure : The Multi-Ethnic Study of Atherosclerosis (MESA).

Harding, S. M. (2000). Complications and consequences of obstructive sleep apnea. *Current Opinion in Pulmonary Medicine*. https://doi.org/10.1097/00063198-200011000-00004

Hassan, A. R. (2016). A comparative study of various classifiers for automated sleep apnea screening based on single-lead electrocardiogram. In *ICEEE 2015 - 1st International Conference on Electrical and Electronic Engineering*. https://doi.org/10.1109/CEEE.2015.7428288

Khalighi, S., Sousa, T., Pires, G., & Nunes, U. (2013). Expert Systems with Applications Automatic sleep staging : A computer assisted approach for optimal combination of features and polysomnographic channels. *Expert Systems With Applications*, *40*(17), 7046–7059. https://doi.org/10.1016/j.eswa.2013.06.023

Kononenko, I. (2001). Machine learning for medical diagnosis : history , state of the art and perspective, *23*.

Marrone, O., & Bonsignore, M. R. (2018). Blood-pressure variability in patients with obstructive sleep apnea : current perspectives, 229–242.

Rothwell, P. M., Howard, S. C., Dolan, E., O'Brien, E., Dobson, J. E., Dahlöf, B., … Poulter, N. R. (2010). Prognostic significance of visit-to-visit variability, maximum systolic blood pressure, and episodic hypertension. *The Lancet*. https://doi.org/10.1016/S0140-6736(10)60308-X

Shiina, K., Tomiyama, H., Takata, Y., Matsumoto, C., Odaira, M., Kato, K., … Yamashina, A. (2016). Obstructive Sleep Apnea as Possible Causal Factor for Visit-to-Visit Blood Pressure Variability. *Circulation Journal*. https://doi.org/10.1253/circj.cj-16-0302

TANIGAWA, T., TACHIBANA, N., YAMAGISHI, K., MURAKI, I., KUDO, M., OHIRA, T., … ISO, H. (2004). Relationship between Sleep-Disordered Breathing and Blood Pressure Levels in Community-Based Samples of Japanese Men. *Hypertension Research*, *27*(7), 479–484. https://doi.org/10.1291/hypres.27.479

Vasan, R. S., Larson, M. G., Leip, E. P., Evans, J. C., O'Donnell, C. J., Kannel, W. B., & Levy, D. (2002). Impact of High-Normal Blood Pressure on the Risk of Cardiovascular Disease. *New England Journal of Medicine*. https://doi.org/10.1056/nejmoa003417

Xiao, M., Yan, H., Song, J., Yang, Y., & Yang, X. (2013). Sleep stages classification based on heart rate variability and random forest. *Biomedical Signal Processing and Control*. https://doi.org/10.1016/j.bspc.2013.06.001

Zinchuk, A. V, Jeon, S., Koo, B. B., Yan, X., Bravata, D. M., Qin, L., … Yaggi, H. K. (2018). Polysomnographic phenotypes and their cardiovascular implications in obstructive sleep apnoea, (Cvd), 472–480. https://doi.org/10.1136/thoraxjnl-2017-210431