Paper 11841-2016

Diagnosing Obstructive Sleep Apnea: Using Predictive Analytics Based on Wavelet Analysis in SAS/IML® Software and Spectral Analysis in PROC SPECTRA

Woranat Wongdhamma, Ph.D., Oklahoma State University

ABSTRACT

This paper presents an application based on predictive analytics and feature-extraction techniques to develop the alternative method for diagnosis of obstructive sleep apnea (OSA). Our method reduces the time and cost associated with the gold standard or polysomnography (PSG), which is operated manually, by automatically determining the OSA's severity of a patient via classification models using the time-series from a one-lead electrocardiogram (ECG). The data is from Dr. Thomas Penzel of Philipps-University, Germany, and can be downloaded at www.physionet.org. The selected data consists of 10 recordings (7 OSAs, and 3 controls) of ECG collected overnight, and non-overlapping-minute-by-minute OSA episode annotations (apnea and non-apnea states). This accounts for a total of 4,998 events (2,532 non-apnea and 2,466 apnea minutes). This paper highlights the nonlinear decomposition technique, wavelet analysis (WA) in SAS/IML® software, to maximize the information of OSA symptoms from ECG, resulting in useful predictor signals. Then, the spectral and cross-spectral analyses via PROC SPECTRA are used to quantify important patterns of those signals to numbers (features), namely power spectral density (PSD), cross power spectral density (CPSD), and coherency, such that the machine learning techniques in SAS® Enterprise Miner™, can differentiate OSA states. To eliminate variations such as body build, age, gender, and health condition, we normalize each feature by the feature of its original signal (that is, ratio of PSD of ECGs WA by PSD of ECG). Moreover, because different OSA symptoms occur at different times, we account for this by taking features from adjacency minutes into analysis, and select only important ones using a decision tree model. The best classification result in the validation data (70:30) obtained from the Random Forest model is 96.83% accuracy, 96.39% sensitivity, and 97.26% specificity. The results suggest our method is well comparable to the gold standard.

INTRODUCTION

Sleep is a crucial part of life. The human body becomes fatigued during the day because of numerous activities and rejuvenates itself during the night while sleeping, creating a daily life cycle of degradation and renewal. Good sleep fosters a good working state for the body. Conversely, a disturbed sleep will not restore the body to its normal working state. Sleep disorders prevent the body from rejuvenating. One form of sleep disorders is sleep apnea, a common disorder marked by frequent pauses in breathing or shallow breaths during sleep. The most prevalent form of apnea, called obstructive sleep apnea (OSA), is due to a partly or completely obstructed airway. Clinically, OSA is identified as a major risk factor for hypertension, arrhythmias, stroke, myocardial infarction, congestive heart failure, and death [1-6]. Approximately 1 in 15 adults, or about 18 million Americans, have moderate or severe OSA [7] and more than half of them remain undiagnosed. An estimated 50 to 70 million Americans suffer from chronic sleep apnea [8], and hundreds of billions of dollars are spent each year in direct medical costs for screening and treatment [9].

OSA has subtle observable symptoms during the day and, more importantly, it is almost impossible for the person with OSA to realize that he or she has the disease because it occurs when the person is asleep. A vast majority of OSA patients seek treatment and/or receive the diagnosis after their condition becomes moderate or severe. Unlike standard diagnosis of chronic diseases such as hypertension or diabetes, in which a test is performed routinely during an annual physical exam, a sleep apnea diagnosis is made only after a patient expresses sleep discomfort or upon a doctor's recommendation. The gold standard for OSA screening and diagnosis involves administering polysomnography (PSG) or a sleep study. PSG involves the patient spending the entire night in a sleep clinic with many sensors attached to several parts of the body for recording several biological signals. To complete the study, sleep apnea episodes (i.e., impeded or difficult breathing events) are marked manually by a sleep specialist or a sleep

doctor by looking for specific patterns in the multiple bio-signal time-series collected overnight. The purpose of a PSG is to determine the severity of the sleep apnea condition called apnea-hypopnea index (AHI) by noting how often a subject stops breathing (apnea) on average in each hour of sleep. Sleep clinics are known to collectively perform 1.17 million screening tests per year in the U.S. [9], a very low number compared to the number of individuals with OSA, estimated at 70 million. The waiting time for diagnosis and screening of suspected patients ranges from 2 to 10 months [9]. Baseline estimates of 5-year diagnosis and treatment charges for a patient with OSA are about $4,210 [10]. The OSA diagnosis process mentioned above has two key problems. 1) Because of the limited equipment and number of facilities for PSG, a relatively small percentage of the OSA population can be tested. 2) Each apnea episode must be manually and individually marked by a sleep technician or sleep doctor, a laborious task that requires a significant amount of time to complete. There is a clear need to expedite the diagnosis process to reduce the overall medical costs and the adverse impact on an individual's health. Currently, with the advancement of the microelectromechanical systems (MEMs), one-lead ECG signal acquisition can be done at home with a small wearable device. From previous studies [11-14], it is clear that the information about cardiorespiratory system is embedded in an ECG signal.

In this paper, we present the method to diagnose an OSA existence and determine its severity from only one-lead ECG signal collected overnight using advance feature extraction and data mining techniques. The organization of this paper is as follows. First, we give a brief background of a wavelet analysis and demonstrate the technique and codes in SAS/IML®. Then, the spectral and cross-spectral analyses via PROC SPECTRA to extract information (features) related to OSA symptoms are explained. We also give a detail about our data used in this analysis and how to preprocess them. Next, the techniques for developing the classification model from the extracted features are depicted. Finally, the results and conclusion are given in the last sections.

WAVELET TRANSFORMATION: BACKGROUND

Wavelet decomposition is a modified short-time Fourier transform that represents the decomposed signals in both time and frequency domain through time windowing function or mother wavelet function [15]. Traditionally, the Fourier transform is normally used for analyzing the signal in frequency domain. However, in nonlinear time-series that contains short duration transients, Fourier transform failed to capture that behavior. When transformed the short transient in time domain to frequency domain, it corresponds to a damped and long-duration vibration [16]. This time-frequency localization advantage is a well-known characteristic of a wavelet transformation. In contrast to Fourier transform, which assumes the signal to be stationary, the wavelet analysis does not have such limitation so that it works well with the nonstationary time-series.

The wavelet transformation process comprises of two main phases, analysis or decomposition and synthesis or reconstruction phases. If the certain condition is met, the signal can be perfectly reconstructed using the coefficients obtained from the analysis or decomposition phase. With these reasons, the wavelet decomposition is popular in a signal denoising application. The user can selectively delete the decomposed coefficients corresponding to the noises and reconstruct the denoised signal back.

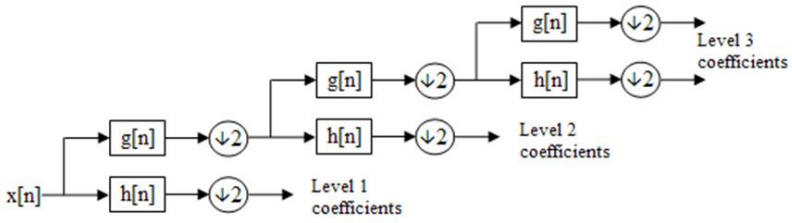

Figure 1. Discrete wavelet transform (DWT) using multiresolution analysis (MRA) with 3 level filter banks

There are several mathematical methods that could be used to achieve a wavelet decomposition. The one that seems to be intuitively easy to understand is a multiresolution analysis (MRA) developed by Mallat in 1989 [17]. In general discrete wavelet transformation (DWT), the signal is passed through a series

of high-pass filters (mathematical tool that allows only fast changing value data to pass, otherwise zero) and low-pass filters (passing slow changing value data, otherwise zero) as shown in Figure 1.

The DWT procedure starts from feeding the time-series $x[n]$ to the half band low-pass filter with an impulse response $g[n]$ and half band high-pass filter with an impulse response h[n]. In mathematical expression, the filtering process is the convolution of the signal with the impulse response of the filter:

$$x[n] * h[n] = \sum_{k=-\infty}^{\infty} x[k] \cdot h[n-k] \quad (1)$$

Regarding to the Nyquist theory, after passing the signal through either a half band low-pass filter or a half band high-pass filter, half of the samples could be eliminated. This denotes by the symbol ↓2 in the Figure 1. The result of the first high-pass filter is level 1 detail coefficients. Likewise, the result of the first low-pass filter is level 1 approximation coefficients. To perform a further analysis, the level 1 approximation coefficients are used as a signal to be passed through another set of half band low-pass and high-pass filters. In theory, the decomposition level could be done for n levels. However, in practice, the analysis levels depend on a number of samples of the original signal. It should be noted that because the decomposition process involves the downsampling with the factor of two. Thus, the number of samples required in the wavelet analysis must be the power of two.

In synthesis (reconstruction) phase, to be able to perfectly reconstruct the signal back from the wavelet coefficients in every decomposed level, the pair of low-pass and high-pass filters must form orthonormal bases. To satisfy that constrain, the relationship between them is [18]:

$$h[L-1-n] = (-1)^n \cdot g[n] \quad (2)$$

Where $h[n]$ is impulse response of a high-pass filter

$g[n]$ is impulse response of a low-pass filter

L is the filter length in number of sample

When the filter pair that satisfies equation 2 is used, the reconstruction process is exactly the reverse process of the analysis process. The coefficients at every level are upsampled with the factor of two then passed through the synthesis filter pairs. The relationship between the analysis and synthesis filters is that they are identical to each other but time reversal. There are many choices of the low-pass and high-pass filter pairs used in wavelet analysis. SAS/IML® [19] provides two choices of wavelet family, the Daubechies Extremal phase family, and the Daubechies Least Asymmetric family (Symmlet family). For further information, reference [18] provides very good information on theory and application of wavelet decomposition.

WAVELET DECOMPOSITION USING PROC IML

The goal of this section is to go through how a wavelet decomposition is processed in SAS/IML®. In this paper, we use SAS/IML® version 12.1 user guide [19] as a guideline, specifically Chapter 19 and 23 for a wavelet analysis. To start, after the time-series data was successfully imported into SAS library. We can activate PROC IML using the following command:

```
proc iml;
```

Unlike others, PROC IML works interactively inside its own shell. Once activated, the commands for computation are little different comparing to the normal commands. Now, we will start to do a wavelet decomposition using the commands bellow:

```
use SASUSER.A15HRV; *indicate the dataset to be used;
   read all var{HRV} into signal;
   optn ={3,.,2,10};*SYMLET10;
     call wavft(decomp,signal,optn);
     call coefficientPlot(decomp, , , , ,"Summary of wavelet decomposition's coefficient");
```

The first line in the code above is to indicate which dataset to be used. In this case, dataset $A15HRV$ from a library $SASUSER$ is assigned. Next, we read all values in variable HRV into a variable name $signal$ in PROC IML shell. Then, the options for our wavelet decomposition is assigned. Briefly, there are 4 options needed to be declared before the wavelet decomposition could be executed.

The first element in vector $optn$ (opt[1]) indicates how the signal boundary is handled. One of the wavelet analysis limitations is that the analysis signal must have a number of data points (N) in the increment of 2^n where n=1,2,3…. SAS/IML has a built-in function to handle this limitation using several options for padding the signal such as padding the signal by zero, the signal reflection, user specified number, and so on. In our case, we use a signal reflection because the extension sequence near boundary is continuous. A user should experiment with all extension methods for the best result. However, to reduce the error introduced in the analysis process, it is suggested that the data should be format to the length of 2^n so that the extension is not needed. In the next option, (opt[2]), the user can indicate the degree of the polynomial to be used in the data padding if the first option (opt[1]) is set to be 2. Since we use the signal reflection, this option will be ignored by PROC IML. For option 3, (opt[3]), the user must specify the method to be used for a decomposition. Symmlet family, (opt[3]=2), was chosen in our case because of its near symmetric property which is desirable in the reconstruction phase.

Finally, the last option, (opt[4]), chooses which wavelet family member to be used in the decomposition. Generally, the wavelet family member indicates how enlarged or compressed the wavelet base function is (the higher number indicates more compressed wavelet base function). The choice for choosing this number depends solely on the user's application. Some experiments may be needed before the final wavelet family member is chosen. For the demonstration, we use Symmlet10 in this case (opt[4]=10). For more information about the aforementioned options, please consult Chapter 19 and Chapter 23 in the SAS/IML® user manual [19].

After required options have been specified, we call a wavelet decomposition $(call\ wavft(\ldots..))$ on variable $signal$ and its decomposition information will be stored in variable $decomp$. To visually inspect the decomposition, $coefficientPlot$ call is used. The result is shown in the Figure 2. We can use the call $wavprint$ to see the summary of the composition also. From this plot, we look for the total number of the decomposed levels. In this case, we have a total of 23 decomposition levels (start level = 0 and top level =22). The lower levels are composed of lower frequencies (slow changing) components extracted from the original data. Likewise, the higher levels are composed of higher frequency (fast changing) components.

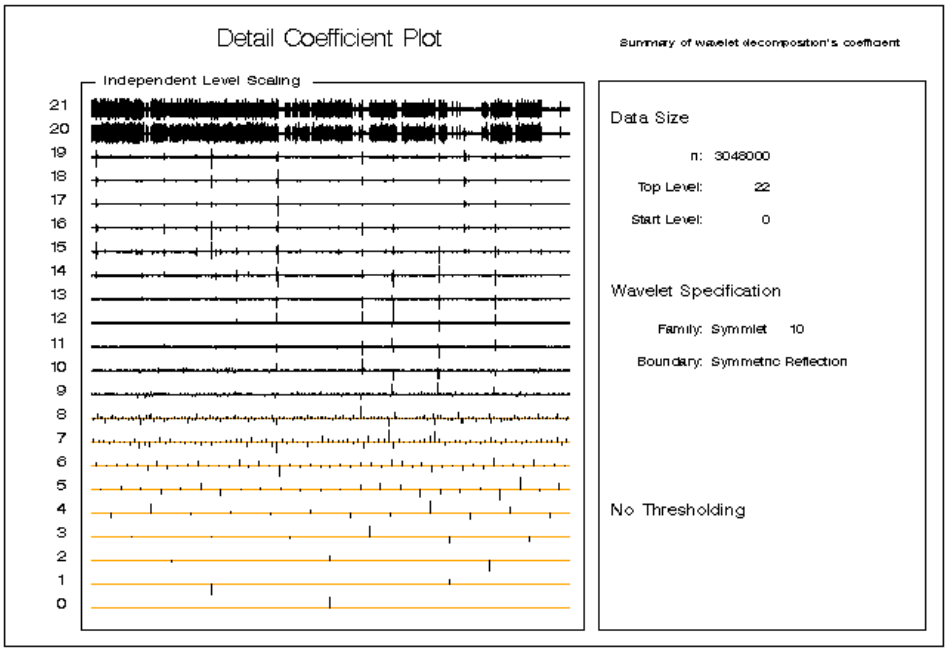

Figure 2. Detailed coefficient plot and decomposition summary

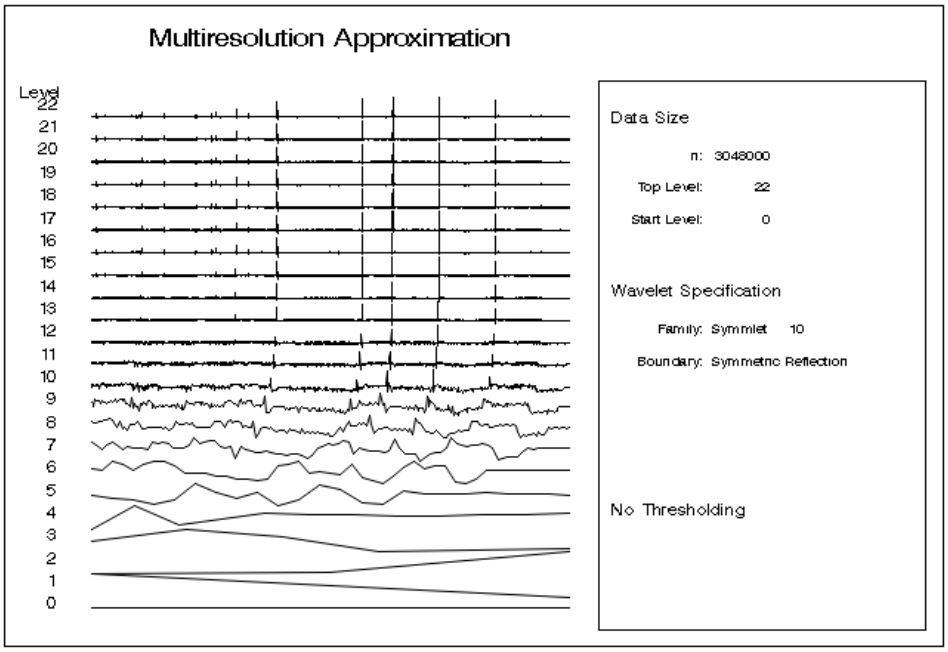

Figure 3. Multiresolution approximation of the signal corresponding to each decomposition level

Next, the wavelet decomposition based on a multiresolution analysis (MRA) is called using the command below:

call mraApprox(decomp,,0,,);

The result of the $mraApprox$ call is shown in Figure 3. The time-series in each level is corresponding to the reconstruction based solely on the detail coefficient of that particular level. If no loss was introduced in analysis and synthesis phases, the summation of every level shown in Figure 3 is our original signal. Unfortunately, to the best of our knowledge, with the PROC IML version used in this analysis (12.1), there is no direct way to obtain any decomposed time-series from $mraApprox$ call. However, with available commands in PROC IML, we can reconstruct each corresponding time-series based on the MRA concept from the process that will be described as follows.

By using the obtained coefficients in variable $decomp$ and its wavelet base function, we can reconstruct the original signal back or choose not to use some levels that corresponding to the noise in the signal in the reconstruction process. PROC IML has a built-in function to help eliminate noises in the signal with $wavift$ call which is the Inverse Fast Wavelet Transform (WAVIFT) via several thresholding methods such as hard, soft, and garrote thresholding methods (see [19] for more information). Now, to continue our decomposition, our next goal is to reconstruct the time-series corresponding to each decomposed level using MRA concept. Another way around is to manually keep each level of the reconstructed time-series by manually thresholding other non-desired level. It may sound simple but the $wavift$ call was not originally designed to do such task. First, the usage of $wavift$ call is as follows [19]:

call WAVIFT(result,decomp<,opt>< ,level>);

The options we used is the hard threshold which corresponds to the equation 3 below [19]:

$$\delta_T^{hard}(x) = \begin{matrix} 0 \; if \; |x| \leq T \\ x \; if \; |x| > T \end{matrix} \tag{3}$$

Intuitively, if the absolute value (magnitude) of the signal $(|x|)$ is smaller than or equal to the threshold value (T), that data point will be set to zero, but if it is larger than the threshold, it will be set to itself. Thus, we will use a very high threshold on the levels that we would like to eliminate. The code used in this case is as follows (still in PROC IML shell):

```
n=nrow(signal);*declare array in proc iml;
    signal1=j(n,23,0);
    effect=j(n,23,0);
    temp=j(n,1,0);
        opt=j(4,23,0);
        opt[1,]=1;
        opt[2,]=0;
        opt[3,]=1000;
        opt[4,]=0:22;
        call wavift(buffer,decomp,opt[,1]);
    signal1[,1]=buffer;
 *reconstruct wavelet decomposed signal to all levels;
 do i=1 to 23;
    call wavift(buffer,decomp,opt[,i]);
    signal1[,i]=buffer;
 end;
```

To thoroughly explain the code above, the calculation in PROC IML is done in a matrix fashion so that it is a good practice to declare a dimension of the matrix that will be used. For example, a number of data point in $signal$ is looked up and kept in variable n. Then, we will store the reconstruction results in the variable name $signal1$ so that we declare the size of this variable to be n row and 23 columns which is corresponding to the decomposition levels.

For $WAVIFT$ call options, the first option is to specify that the hard thresholding method will be used. Then, we specify option two to be 0 to use the global user-defined threshold. For the third option, this is a threshold value $(T = 1000)$ which is pretty high comparing to our signal. Finally, the last option will specify the number of levels that the thresholding will be applied to, starting from the highest level. This means that we cannot apply the hard threshold exclusively on each detail coefficient. Again, the calculation method to get the individual reconstructed time-series will be explained later. For now, we will apply the threshold to the detail coefficient and reconstruct the decomposed signal starting from the lowest level iteratively until we reach the highest level of the decomposition. This is done by do-loop in the code above.

In the first loop, we applied the threshold to the highest level at level 0, meaning that the thresholding was done to level 0 only. Therefore, the reconstructed signal in this loop is the signal that does not contain any effect from the detail coefficient at level 0. In the next loop, the thresholding was done to the highest level at level 1 and 0 so that the reconstructed signal does not contain any effect from level 0 and level 1 detail coefficients. The process is executed until we reach the last level. The reconstructed signals without the effects are stored in variable $signal1$ in the hierarchy fashion (highest to lowest). Finally, we can obtain the exclusively reconstructed time-series from each level detail coefficient (effect) by:

```
*calculate effects;
do i=1 to 23;
    effect[,i]=signal-signal1[,i]-temp;
    temp=temp+effect[,i];
end;
```

The idea is that, the first column of variable $signal1$ is the time-series that does not contain any reconstructed component from level 0. Thus, if we subtract this time-series off the original signal $(signal)$, what left is actually the reconstructed time-series exclusively from the level 0 detail coefficient (we will call this the effect 0). Thus, in the next iteration, the reconstructed time-series exclusively from the level 1 coefficient (effect 1) could be derived from subtracting the time-series that does not contain any reconstructed component from level 0 and 1 (second column of variable $signal1$) and effect 0 from the original signal $(signal)$. This process is executed until we obtain all effect time-series equal to the number of decomposed levels. These decomposed time-series are much less complex than the original signal and contain different central dominant frequencies (from high to low). The process to quantify the characteristics of these signals in a frequency domain using PROC SPECTRA is explained in next section.

SPECTRAL ANALYSIS USING PROC SPECTRA

Briefly, a spectral analysis is an analysis of the spectral contents or components in a frequency domain (i.e., the distribution of power over different frequencies) of a time-series [20]. Popular components are spectral and cross-spectral densities. The spectral density also known as a power spectral density (PSD) is the method for explaining how the variance of a time-series is distributed in the frequency domain. It quantifies the variance of the time-series at all frequencies (for more information please see [21] and [22]). The cross-spectral density also known as a cross power spectral density (CPSD) uses the same concept. It quantifies the variance shared by a given frequency for the two time-series using its amplitude squared and the phase shift between them at a given frequency. PROC SPECTRA estimates the spectral and cross-spectral densities using a periodogram and cross-periodogram obtained from a finite Fourier transform [23]:

$$x_t = \frac{a_0}{2} + \sum_{k=1}^{m-1} f_k(a_k cos\omega_k t + b_k \sin \omega_k t)$$
$$f_t = \begin{cases} 0.5 & if\ n\ is\ even\ and\ k = m-1 \\ 1 & otherwise \end{cases} \quad (4)$$

Where

- t is the time subscript, t = 0,1,2,…, n - 1
- x_t are the equally spaced time-series data
- n is the number of observations in the time-series
- m is the number of frequencies in the Fourier transform, $m = \frac{n+2}{2}$ if n is even and $m = \frac{n+1}{2}$ if n is odd
- k is the frequency subscript, $k = 0,1,2,\ldots,m-1$
- a_0 is the mean term, $a_0 = 2\bar{x}$
- a_k are the cosine coefficients
- b_k are the sine coefficients
- ω_k are the Fourier frequencies, $\omega_k = \frac{2\pi k}{n}$

The Fourier transform in equation (4) represents the time-series in terms of sine and cosine functions in different amplitudes and frequencies. A periodogram is a plot of functions of the Fourier coefficient a_k and b_k against frequency. It is a sequence of the amplitude periodogram, J_k, below [23]:

$$J_k = \frac{n}{2}(a_k^2 + b_k^2) \quad (5)$$

For the amplitude cross-periodogram, J_k^{xy}, can be defined below [23]:

$$J_k^{xy} = \frac{n}{2}\left(a_k^x a_k^y + b_k^x b_k^y\right) + i\frac{n}{2}\left(a_k^x a_k^y + b_k^x b_k^y\right) \quad (6)$$

Where

- i Is an imaginary part, $i = \sqrt{-1}$
- a_k^x are the cosine coefficients of time-series, x_t
- a_k^y are the cosine coefficients of time-series, y_t
- b_k^x are the sine coefficients of time-series, x_t
- b_k^y are the sine coefficients of time-series, y_t

Finally, the spectral and cross-spectral densities are estimated by smoothing the periodogram and cross-periodogram respectively using a weight or window function. In this study, we use Tukey-Hanning window as a weight function because it gives very low aliasing and has good frequency resolution compared

to other weight functions. To demonstrate the spectral and cross spectral analysis using PROC SPECTRA, we simply run the code below:

```
PROC spectra DATA=SC.ECGHRV_&I out=SC.ECGHRVPSD_&I CROSS A P S K PH S ADJMEAN;
    VAR ECG HRV;
    weights TUKEY 0.5 0;
RUN;
```

The example above computes cross-spectral density estimates between variable ECG and HRV and other features as shown in Table 1 below (for complete option set please see [23]):

PROC SPECTRA options	
Option	**Result**
CROSS	output cross-spectral analysis results
A	output the amplitudes of the cross-spectrum
P	output the periodogram
S	output the spectral density estimates
K	output squared coherency of the cross-spectrum
PH	output the phase of the cross-spectrum
S	output the spectral density estimates
ADJMEAN	subtract the series mean
WEIGHTS TUKEY	specify the Tukey-Hanning for weight or window function used for smoothing spectral and cross-spectral periodogram

Table 1. PROC SPECTRA options used in the analysis

From the options specified in the code about, there are 9 features computed from bivariate time-series ECG, and HRV as follows [23]:

Features	**Mathematical form**	**Description**
P_nn	$J_k^x = \frac{n}{2}[(a_k^x)^2 + (b_k^x)^2]$	periodogram of time-series x_t (periodograms of the bivariate time-series are calculated individually)
S_nn	$F_k^x = \sum_{j=-p}^{x} W_j J_{k+j}^x$ (except across endpoint)	spectral density estimate of time-series x_t
RP_nn_mm	$Re(J_k^{xy}) = \frac{n}{2}(a_k^x a_k^y + b_k^x b_k^y)$	real part of cross-periodogram of time-series x_t and y_t
IP_nn_mm	$Im(J_k^{xy}) = \frac{n}{2}(a_k^x a_k^y + b_k^x b_k^y)$	imaginary part of cross-periodogram of time-series x_t and y_t
CS_nn_mm	$C_k^{xy} = \sum_{j=-p}^{p} W_j\ Re(J_{k+j}^{xy})$ (except across endpoint)	cross-spectrum estimate (real part of cross-spectrum) of time-series x_t and y_t
QS_nn_mm	$Q_k^{xy} = \sum_{j=-p}^{p} W_j\ Im(J_{k+j}^{xy})$ (except across endpoint)	quadrature spectrum estimate (imaginary part of cross-spectrum)
A_nn_mm	$A_k^{xy} = \sqrt{(C_k^{xy})^2 + (Q_k^{xy})^2}$	amplitude (modulus) of cross-spectrum of time-series x_t and y_t
K_nn_mm	$K_k^{xy} = (A_k^{xy})^2/(F_k^x F_k^y)$	coherency squared of time-series x_t and y_t
PH_nn_mm	$\Phi_k^{xy} = \arctan(Q_k^{xy}/C_k^{xy})$	phase spectrum in radians of time-series x_t and y_t

Table 2. PROC SPECTRA options used in the analysis

Besides spectral and cross-spectral density quantities, we also calculate for the coherency which tells the degree of relationship between two time-series as a function of frequency. All features in Tables 2 will be used as predictors for building classification models later.

DATA DESCRIPTION AND PREPARATION

The data used in this study is from Dr. Thomas Penzel of Philipps-University, Germany, and can be downloaded at www.physionet.org. The selected data consists of 10 recordings (7 OSAs, and 3 controls) of ECG collected overnight (more than 7 hours in each case), and non-overlapping-minute-by-minute OSA episode annotations diagnosed by a sleep physician (apnea and non-apnea states). All R-peaks in each ECG time-series were detected and converted to the HRV time-series (also available for download at the same website). The HRV tells how fast the heart pumps the blood to the body via blood vessels. The sampling rate for an ECG signal is 100 Hz (100 samples per second) so that 6,000 samples is equal to one-minute length of data which accompanies with the OSA episode annotation (diagnosed as apnea or non-apnea minute). After segmenting of the data into one-minute length windows, the overall data of the 10 recordings account for a total of 4,998 events (2,532 non-apnea and 2,466 apnea minutes).

METHODOLOGY

The idea is to deeply extract information that relates to OSA events from ECG and HRV data. We start from decomposing ECG and HRV time-series using wavelet analysis in SAS/IML. This results in 40 decomposed time-series (20 wavelet decomposition levels of ECG and HRV time-series). Each decomposed, and original ECG and HRV time-series are then segmented into one-minute length data windows. Then, we do a spectral analysis on each time-series, and a cross-spectral analysis on each pair of all time-series. Moreover, to eliminate variations such as body build, age, gender, and health condition, we normalize each feature by the feature of its original signal (that is, ratio of spectral density of wavelet decomposed ECGs by spectral density of ECG). This accounts for 2,900 independent variables (features).

Because the OSA symptoms are present not only in the apneic minute but some significant symptoms also appear before and after the apneic minute (i.e., increased heart rate and overshot respiration), we also included the features of ECG and HRV time-series of two minutes before and after the apneic minute into each data vector. Before having machine learning techniques to learn for patterns in the data to differentiate between apnea and non-apnea states, we use a Decision Tree to select only the variables that relate to the target variable (apnea and non-apnea states). Most of the variables are rejected, and there are 62 variables remain in an analysis. The data is partitioned into training and validation partitions (70:30). We use machine learning techniques namely, Logistic Regression, Decision Tree, Neural Networks, Support Vector Machine (SVM), and Random Forest, to build classification models having OSA states (apnea or non-apnea) as a target variable.

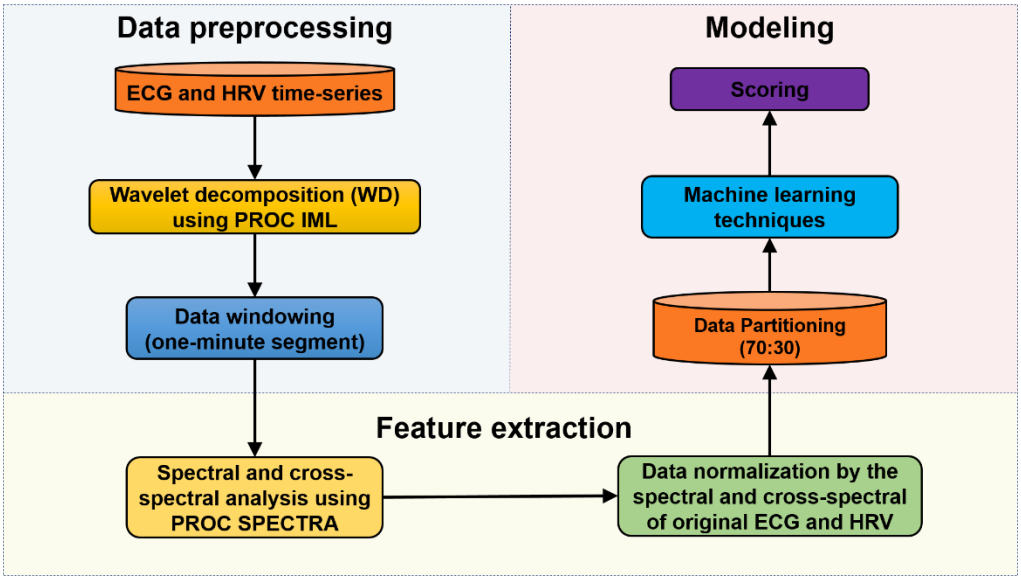

Figure 4. Research methodology

The best model is determined by the overall accuracy in the validation data partition. Finally, we use the best model to score the apnea event from features extracted from each subject individually to determine each subject's AHI. The overall process is depicted in Figure 4.

RESULTS

To choose which model is the best for an OSA diagnosis using one-lead ECG signal, we select the model that gives the highest predictive accuracy in the validation data partition. All model results are shown in Table 3 below:

Model	True positive rate (TPR)	True negative rate (TNR)	False positive rate (FPR)	False negative rate (FNR)	Accuracy
Logistic Regression	82.85%	84.87%	15.13%	17.15%	83.86%
Decision Tree (CART)	90.06%	87.03%	12.97%	9.94%	88.54%
Neural Network	89.77%	94.23%	5.77%	10.23%	92.00%
Support Vector Machine (SVM)	86.31%	85.88%	14.12%	13.69%	86.10%
Random Forest	96.39%	97.26%	2.74%	3.61%	96.83%

Table 3. Classification model performances in a validation data partition

Based on model performances in the validation data partition shown in Table 3 above, the classification model trained from a Random Forest modeling method gives the best accuracy of 96.83% with low FPR (2.74%) and FNR (3.61%). Furthermore, we also apply this model to each subject's features extracted from an ECG data to see how it performs individually. The desired result for this study is a predicted apnea-hypopnea index (AHI) which is used for determining a severity of OSA clinically. It is calculated by:

$$AHI = \frac{total\ number\ of\ apnea\ and\ hypopnea\ events}{total\ number\ of\ sleep\ time\ in\ minute} \cdot 60 \quad (7)$$

The summary of the best model's performances is reported individually in Table 4 below:

Apnea group	Accuracy	Total sleep min	Actual Apnea min	Predicted Apnea min	Actual non-apnea min	Predicted non-apnea min	Actual AHI	Predicted AHI
a01	96.38%	487	470	485	17	2	57.90	59.75
a02	96.86%	526	420	432	106	94	47.91	49.27
a03	97.34%	517	246	249	271	268	28.55	28.89
a15	91.55%	505	368	399	137	106	43.72	47.40
a18	96.37%	487	438	443	49	44	53.96	54.57
a19	97.10%	500	204	197	296	303	24.48	23.64
a20	96.86%	508	315	319	193	189	37.20	37.67
Control group	Accuracy	Total sleep min	Actual Apnea min	Predicted Apnea min	Actual non-apnea min	Predicted non-apnea min	Actual AHI	Predicted AHI
c03	100	448	0	0	448	448	0	0
c07	99.52	427	4	1	423	426	0.56	0.14
c08	99.76	509	0	1	509	508	0	0.12

Table 4. Summary of the classification model performance on each individual ECG data

Shown in Table 4, the accuracy of the model to predict the OSA event (apnea or non-apnea) using only one-lead ECG data is high (>96% in most cases). Comparing the actual AHI calculated by the actual

apnea minutes, with the predicted AHI calculated by the predicted apnea minutes, our model tends to slightly overestimate the AHI. However, the differences are very subtle in most cases.

CONCLUSION

This paper presents an application based on predictive analytics and feature-extraction techniques to develop the alternative method for diagnosis of obstructive sleep apnea (OSA). Our method reduces the time and cost associated with the gold standard or polysomnography (PSG), which is operated manually, by automatically determining the OSA's severity of a patient via classification models using the time-series from a one-lead electrocardiogram (ECG) that can be collected overnight using on-the-shelf wearable devices.

We use the nonlinear decomposition technique, wavelet analysis (WA) in SAS/IML® software, to maximize the information of OSA symptoms from ECG, resulting in useful predictor signals. Then, the spectral and cross-spectral analyses via PROC SPECTRA are used to quantify important patterns of those wavelet decomposed signals to numbers (features), namely power spectral density (PSD), cross power spectral density (CPSD), and coherency. To eliminate variations such as body build, age, gender, and health condition, we normalize each feature by the feature of its original signal (that is, ratio of PSD of ECGs WA by PSD of ECG). Moreover, because different OSA symptoms occur at different times, we account for this by taking features from adjacency minutes into analysis, and select only important ones using a decision tree model.

To build classification models from those features to differentiate OSA states, we use machine learning techniques namely, Logistic Regression, Decision Tree, Neural Networks, Support Vector Machine (SVM), and Random Forest, in SAS® Enterprise Miner™. The best classification result in the validation data (70:30) obtained from the Random Forest model is 96.83% accuracy, 96.39% sensitivity, and 97.26% specificity. Furthermore, each subject's apnea-hypopnea index (AHI) which is used for determining a severity of OSA clinically is calculated from the predicted OSA events and compared with the actual ones. The results suggest our method is well comparable to the gold standard.

REFERENCES

1. Malhotra, A. and J. Loscalzo, *Sleep and Cardiovascular Disease: An Overview.* Progress in Cardiovascular Diseases, 2009. **51**(4): p. 279.
2. Leung, R.S.T. and T.D. Bradley, *Sleep Apnea and Cardiovascular Disease.* American Journal of Respiratory and Critical Care Medicine, 2001. **164**(12): p. 2147-2165.
3. Ip, M.S.M., B. Lam, M.M.T. Ng, W. Lam, K.W.T. Tsang, and K.S.L. Lam, *Obstructive Sleep Apnea Is Independently Associated with Insulin Resistance.* American Journal of Respiratory and Critical Care Medicine, 2002. **165**(5): p. 670-676.
4. Marin, J.M., S.J. Carrizo, E. Vicente, and A.G. Agusti, *Long-Term Cardiovascular Outcomes in Men with Obstructive Sleep Apnoea—Hypopnoea with or without Treatment with Continuous Positive Airway Pressure: An Observational Study.* The Lancet, 2005. **365**(9464): p. 1046-1053.
5. Shahar, E., C. WHITNEY, S. REDLINE, E. LEE, A. NEWMAN, F. JAVIERNIETO, G. O'CONNOR, L. BOLAND, J. SCHWARTZ, and J. SAMET, *Sleep-Disordered Breathing and Cardiovascular Disease: Cross-Sectional Results of the Sleep Heart Health Study.* American Journal of Respiratory and Critical Care Medicine, 2001. **163**(1): p. 19-25.
6. Go, A.S., D. Mozaffarian, V.L. Roger, E.J. Benjamin, J.D. Berry, W.B. Borden, D.M. Bravata, S. Dai, E.S. Ford, C.S. Fox, S. Franco, H.J. Fullerton, C. Gillespie, S.M. Hailpern, J.A. Heit, V.J. Howard, M.D. Huffman, B.M. Kissela, S.J. Kittner, D.T. Lackland, J.H. Lichtman, L.D. Lisabeth, D. Magid, G.M. Marcus, A. Marelli, D.B. Matchar, D.K. McGuire, E.R. Mohler, C.S. Moy, M.E. Mussolino, G. Nichol, N.P. Paynter, P.J. Schreiner, P.D. Sorlie, J. Stein, T.N. Turan, S.S. Virani, N.D. Wong, D. Woo, and M.B. Turner, *Heart Disease and Stroke Statistics—2013 Update: A Report from the American Heart Association.* Circulation, 2013. **127**(1): p. e6-e245.
7. Young, T., J. Skatrud, and P.E. Peppard, *Risk Factors for Obstructive Sleep Apnea in Adults.* Journal of the American Medical Association, 2004. **291**(16): p. 2013-2016.
8. Altevogt, B.M. and H.R. Colten, *Sleep Disorders and Sleep Deprivation: An Unmet Public Health*

Problem. 2006, Washington, DC: National Academies Press
9. Flemons, W.W., N.J. Douglas, S.T. Kuna, D.O. Rodenstein, and J. Wheatley, *Access to Diagnosis and Treatment of Patients with Suspected Sleep Apnea.* American Journal of Respiratory and Critical Care Medicine, 2004. **169**(6): p. 668-672.
10. Chervin, R.D., D.L. Murman, B.A. Malow, and V. Totten, *Cost-Utility of Three Approaches to the Diagnosis of Sleep Apnea: Polysomnography, Home Testing, and Empirical Therapy.* Annals of Internal Medicine, 1999. **130**(6): p. 496-505.
11. Karandikar, K., T.Q. Le, A. Sa-ngasoongsong, W. Wongdhamma, and S.T. Bukkapatnam. *Detection of Sleep Apnea Events Via Tracking Nonlinear Dynamic Cardio-Respiratory Coupling from Electrocardiogram Signals*. in *Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference on*. 2013. IEEE.
12. Le, T.Q., C. Cheng, A. Sangasoongsong, W. Wongdhamma, and S.T. Bukkapatnam, *Wireless Wearable Multisensory Suite and Real-Time Prediction of Obstructive Sleep Apnea Episodes.* Translational Engineering in Health and Medicine, IEEE Journal of, 2013. **1**: p. 2700109-2700109.
13. Bukkapatnam, S.T., T. Le, and W. Wongdhamma, *Device and Method for Predicting and Preventing Obstructive Sleep Apnea (Osa) Episodes*. 2013, Google Patents.
14. Wongdhamma, W., T.Q. Le, and S.T. Bukkapatnam. *Wireless Wearable Multi-Sensory System for Monitoring of Sleep Apnea and Other Cardiorespiratory Disorders*. in *Automation Science and Engineering (CASE), 2013 IEEE International Conference on*. 2013. IEEE.
15. Shim, I., J. Soraghan, and W. Siew, *Detection of Pd Utilizing Digital Signal Processing Methods. Part 3: Open-Loop Noise Reduction.* Electrical Insulation Magazine, IEEE, 2001. **17**(1): p. 6-13.
16. Chow, S.-M., E. Ferrer, and F. Hsieh, *Statistical Methods for Modeling Human Dynamics: An Interdisciplinary Dialogue*. 2012: Taylor & Francis
17. Mallat, S.G., *A Theory for Multiresolution Signal Decomposition: The Wavelet Representation.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1989. **11**(7): p. 674-693.
18. Gao, R.X. and R. Yan, *Wavelets: Theory and Applications for Manufacturing*. 2010: Springer Science & Business Media
19. SAS Institute Inc, *Sas/Iml 12.1 User's Guide*. 2012, Cary, NC: SAS Publishing.https://support.sas.com/documentation/cdl/en/imlug/65547/PDF/default/imlug.pdf
20. Stoica, P. and R.L. Moses, *Spectral Analysis of Signals*. 2005: Pearson/Prentice Hall Upper Saddle River, NJ
21. Oppenheim, A.V. and R.W. Schafer, *Discrete-Time Signal Processing*. 3rd ed. 2010, Upper Saddle River, NJ: Prentice Hall
22. Ljung, L., *System Identification*. 1998, Englewood Cliffs, NJ: Springer
23. SAS Institute Inc, *Sas/Ets® 13.2 User's Guide the Spectra Procedure*. 2014, Cary, NC.https://support.sas.com/documentation/onlinedoc/ets/132/spectra.pdf

ACKNOWLEDGMENTS

I would like to express my gratitude to Dr. Goutam Chakraborty, Professor, Department of Marketing and founder of SAS and OSU Data Mining Certificate program, Oklahoma State University and my academic advisor, Dr. Sunderesh Heragu, Head and Regents Professor, Department of Industrial Engineering and Management, Oklahoma State University, for their support throughout the research.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Woranat Wongdhamma
Industrial Engineering and Management, Oklahoma State University
woranat@okstate.edu