Patient Identification with ECG and SaO_2 Time Series

by

Mengnan Zhao

A thesis submitted to The Johns Hopkins University

in conformity with the requirements for the degree of

Master of Science in Engineering

Baltimore, Maryland

May, 2019

Abstract

Sudden cardiac death is the most common cause of death in United States. Primary prevention implantable cardioverter defibrillators (ICDs) have been the first line to reduce mortality for high-risk patients. Previous work of identifying subjects at greater risk is neither sensitive nor specific. The development of more reliable predictors that could help identify patients that could benefit from these devices is of both academic and public health interest.

In this thesis, we study the time series data of both electrocardiogram (ECG) and oxygen saturation (SaO_2) signals from patients who received ICD implantation. This sutdy is part of Prospective Observational Study of Implantable CardioverterDefibrillators (PROSE-ICD).

The features for each subject are generated from some statistics of the ECG and SaO_2 signals respectively. For ECG signal, the analysis is from both geometry and dynamics perspective. For SaO_2 signal, multivariate and dynamics analysis is applied. Our results showed an overall accuracy of 93.2% for patient classification, with no bias towards healthy or HF patients. Further analysis does not show a clear relationship between ECG and SaO_2 signals.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Physiological Signal Fundamentals

1.1.1 Electrocardiogram

An electrocardiogram (ECG) is a test that measures the electrical activity of the heartbeat. It is a plot of voltage versus time which is recorded by electrodes placed on the skin. For tens of years, ECG is one of the fastest and simplest ways to evaluate the heart.

As shown in Fig. 1.1, there are three main components to an ECG: the P wave, which represents atrial depolarization; the QRS complex, which in turn includes Q, R and S waves, corresponds to the depolarization of the right and left ventricles; and the T wave, which represents electrical recovery or the return to a resting state of the ventricles.

The orderly pattern of depolarization of an ECG conveys a large amount of important information about the structure and function of the heart. Moreover, the development of acquisition systems during the past decades has enabled the recording of ECG signal over a long period of time which could be used

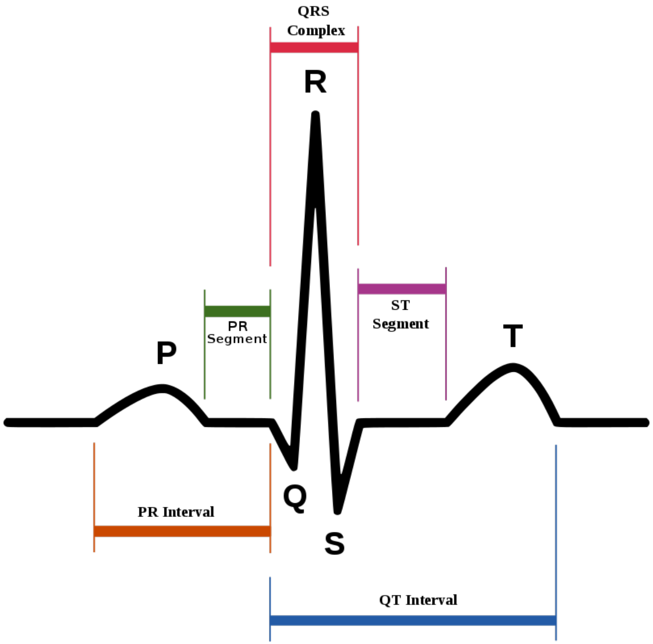


Figure 1.1: ECG of a normal heartbeat (Wikipedia, 2019a)

to detect infrequent abnormalities. Therefore, an ECG signal can be used to diagnose several kind of arrhythmia (Rajpurkar et al., 2017; Owis et al., 2002), damage to the heart's muscle cells (De Capua, Meduri, and Morello, 2010), heart attack (Leijdekkers and Gay, 2008; Acharya et al., 2017) and other anomalies. It is also used to measure the effects of heart drugs (Johannesen et al., 2014) and the function of implanted pacemakers (Jiang and Mangharam, 2011), etc.

A more detailed discussion of the medical uses and interpretation of ECG is beyond the scope of this thesis.

1.1.2 Oxygen Saturation

Oxygen saturation (SaO_2) is the fraction of hemoglobin binding sites occupied by oxygen relative to total hemoglobin in the blood. Normal blood oxygen levels in healthy individuals are 95-100 percent, and tends to maintain around 96 percent. If the SaO_2 (arterial oxygen saturation) value is below 90 percent, it is considered low and the cause of hypoxemia (indicated by cyanosis) (Mayo Clinic, 2018). Blood oxygen saturation levels below 80 percent may impair organ function, such as the brain and heart. Continued low oxygen levels may lead to cardiac or respiratory arrest (Wikipedia, 2019b). A summary of the effects of decreased oxygen saturation is in Table 1.1.

85% and above	No impairment
65% and below	Impaired mental function
55% and below	Loss of consciousness

Table 1.1: Effects of SaO_2.

Oxygen saturation can be measured in different tissues: venous oxygen saturation (SvO_2), tissue oxygen saturation (StO_2), and eripheral oxygen saturation (SpO_2). SpO_2 can be measured with a pulse oximeter device which clips to the body, usually a fingertip. SpO_2 is thought to be a good approximation of SaO_2.

Oxygen saturation levels have been shown to be closely correlated to a variety of diseases, including heart failure (Madsen, Nielsen, and Christiansen, 2000; Ohlsson et al., 2001), sleep apnea (Alvarez et al., 2010; Roebuck et al., 2013; Marcos et al., 2012), vascular complications (Keller, 2009; Lohman et al., 2013), and so on. Although the limitation of oxygen saturation decrease its value as a single diagnostic tool (Netzer et al., 2001), the easy accessibility and high accuracy make it an important complementary noninvasive measurement in the diagnosis of the above diseases.

1.2 Problem Statement

1.2.1 Background

Implantable cardioverter defibrillators (ICDs) are useful in preventing sudden death in patients with ventricular tachycardia or fibrillation. Studies have shown ICD's important role in preventing cardiac arrest in high-risk patients who haven't had, but are at risk for, life-threatening ventricular arrhythmias. However, only a small portion of patients could benefit from implantable ICDs, and the selection of patients for ICD implantation based on ejection fraction criteria lacks sensitivity and specificity (Gehi, Haas, and Fuster, 2005). As a result, there is substantial interest in finding reliable and efficient predictors that could identify patients who could benefit from primary-prevention ICD implantation.

1.2.2 Study Sample and Dataset

The data comes from the project Prospective Observational Study of Implantable Cardioverter Defibrillators (PROSE-ICD), which is a prospective observation study of patients undergoing ICD implantation. The study is being carried out in four medical centers: Johns Hopkins Hospital, University of Maryland Hospital, Washington Hospital Center, and Virginia Commonwealth University Hospital.

The population set includes ICD recipients between 18 and 80 years old who have either ischemic or nonischemic cardiomyopathy. The detailed criteria for inclusion could be found in (Cheng et al., 2013). All patients have received successful ICD implantation. Prior ICD placement, all patients undergo a comprehensive evaluation including history and physical examination, ECG evaluation, cardiac imaging, and blood sampling. Patients are evaluated every 6 months and after every known ICD shock for additional ECG and blood sampling.

The available dataset consists of ECG and SaO_2 data from 484 patients. The patients have been labeled as healthy (388/484) or suffering from heart failure (HF) (96/484). Each patient's data consists of several hours of ECG ($\sim 10^6$ sampling points and $\sim 10^4$ heartbeats) and SaO_2 ($\sim 10^4$ sampling points) signals collected in the same period of time. By checking the quality of data, we found that some snippets of time series are noisy or even purely noise. The data preprocessing phase includes data denoising and automatic segmentation of ECG time series into individual heartbeats.

1.2.3 Objectives

In this study, the goal is to predict each patient's future trend as healthy or HF based only on ECG and SaO_2 time series signals. This would help to develop a reliable, inexpensive and noninvasive method to identify patients to receive primary prevention ICD implantation, and therefore better assist clinical diagnosis and treatment.

1.3 Outline of the Thesis

The remainder of this thesis is organized as follows: Chapter 2 presents geometric and dynamics analysis of ECG signal. Chapter 3 presents results on

SaO_2 signal with both multivariate and dynamics analysis. In Chapter 4, all features from previous sections are ensembled and the patient classification is performed. Also, a discussion on the relationship between ECG and SaO_2 signals is presented. Conclusions of this study are presented in Chapter 5.

References

Wikipedia (2019a). *Electrocardiography — Wikipedia, The Free Encyclopedia*. URL: https://en.wikipedia.org/w/index.php?title=Electrocardiography&oldid=891210210.

Rajpurkar, Pranav, Awni Y Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y Ng (2017). "Cardiologist-level arrhythmia detection with convolutional neural networks". In: *arXiv preprint arXiv:1707.01836*.

Owis, Mohamed I, Ahmed H Abou-Zied, A-BM Youssef, and Yasser M Kadah (2002). "Study of features based on nonlinear dynamical modeling in ECG arrhythmia detection and classification". In: *IEEE transactions on Biomedical Engineering* 49.7, pp. 733–736.

De Capua, Claudio, Antonella Meduri, and Rosario Morello (2010). "A smart ECG measurement system based on web-service-oriented architecture for telemedicine applications". In: *IEEE Transactions on Instrumentation and Measurement* 59.10, pp. 2530–2538.

Leijdekkers, Peter and Valérie Gay (2008). "A self-test to detect a heart attack using a mobile phone and wearable sensors". In: *2008 21st IEEE International Symposium on Computer-Based Medical Systems*. IEEE, pp. 93–98.

Acharya, U Rajendra, Hamido Fujita, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, and Muhammad Adam (2017). "Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals". In: *Information Sciences* 415, pp. 190–198.

Johannesen, Lars, Jose Vicente, JW Mason, Carlos Sanabria, Kristin Waite-Labott, Mira Hong, Ping Guo, John Lin, Jens Stampe Sørensen, Loriano Galeotti, et al. (2014). "Differentiating drug-induced multichannel block on the electrocardiogram: randomized study of dofetilide, quinidine, ranolazine, and verapamil". In: *Clinical Pharmacology & Therapeutics* 96.5, pp. 549–558.

Jiang, Zhihao and Rahul Mangharam (2011). "Modeling cardiac pacemaker malfunctions with the virtual heart model". In: *2011 Annual International*

Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, pp. 263–266.

Mayo Clinic (2018). *Hypoxemia (low blood oxygen)*. URL: https://www.mayoclinic.org/symptoms/hypoxemia/basics/definition/sym-20050930.

Wikipedia (2019b). *Oxygen saturation (medicine) — Wikipedia, The Free Encyclopedia*. URL: https://en.wikipedia.org/w/index.php?title=Oxygen_saturation_(medicine)&oldid=887248032.

Madsen, PL, HB Nielsen, and P Christiansen (2000). "Well-being and cerebral oxygen saturation during acute heart failure in humans." In: *Clinical physiology (Oxford, England)* 20.2, pp. 158–164.

Ohlsson, SH Kubo, D Steinhaus, DT Connelly, S Adler, C Bitkover, R Nordlander, L Ryden, and T Bennett (2001). "Continuous ambulatory monitoring of absolute right ventricular pressure and mixed venous oxygen saturation in patients with heart failure using an implantable haemodynamic monitor: results of a 1 year multicentre feasibility study". In: *European heart journal* 22.11, pp. 942–954.

Alvarez, Daniel, Roberto Hornero, J Victor Marcos, and Felix del Campo (2010). "Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis". In: *IEEE Transactions on Biomedical Engineering* 57.12, pp. 2816–2824.

Roebuck, A, V Monasterio, E Gederi, M Osipov, J Behar, A Malhotra, T Penzel, and GD Clifford (2013). "A review of signals used in sleep analysis". In: *Physiological measurement* 35.1, R1.

Marcos, J Víctor, Roberto Hornero, Daniel Alvarez, Mateo Aboy, and Félix Del Campo (2012). "Automated prediction of the apnea-hypopnea index from nocturnal oximetry recordings". In: *IEEE Transactions on Biomedical Engineering* 59.1, pp. 141–149.

Keller, Alex (2009). "A new diagnostic algorithm for early prediction of vascular compromise in 208 microsurgical flaps using tissue oxygen saturation measurements". In: *Annals of plastic surgery* 62.5, pp. 538–543.

Lohman, Robert F, Claude-Jean Langevin, Mehmet Bozkurt, Neilendu Kundu, and Risal Djohan (2013). "A prospective analysis of free flap monitoring techniques: physical examination, external Doppler, implantable Doppler, and tissue oximetry". In: *Journal of reconstructive microsurgery* 29.01, pp. 051–056.

Netzer, Nikolaus, Arn H Eliasson, Cordula Netzer, and David A Kristo (2001). "Overnight pulse oximetry for sleep-disordered breathing in adults: a review". In: *Chest* 120.2, pp. 625–633.

Gehi, Anil, Donald Haas, and Valentin Fuster (2005). "Primary prophylaxis with the implantable cardioverter-defibrillator: the need for improved risk stratification". In: *JAMA* 294.8, pp. 958–960.

Cheng, Alan, Darshan Dalal, Barbara Butcher, Sanaz Norgard, Yiyi Zhang, Timm Dickfeld, Zayd A Eldadah, Kenneth A Ellenbogen, Eliseo Guallar, and Gordon F Tomaselli (2013). "Prospective observational study of implantable cardioverter-defibrillators in primary prevention of sudden cardiac death: study design and cohort description". In: *Journal of the American Heart Association* 2.1, e000083.

Chapter 2

ECG Time Series Modeling

2.1 ECG Segmentation

Accurate ECG segmentation is essential to automatic ECG analysis. By segmenting ECG signal into individual waveform features, it allows extracting informative features which can be used to detect abnormal heartbeats. A variety of segmentation methods have been proposed and validated, including time warping (Vullings, Verhaegen, and Verbruggen, 1998; Vullings, Verhaegen, and Verbruggen, 1997), hidden Markov model(Andreao, Dorizzi, and Boudy, 2006), EM algorithm(Hughes, Roberts, and Tarassenko, 2004), etc. In general, ECG segmentation is the first and most complicated step in automatic analysis.

The ECG data provided consists of tens of thousands of individual heartbeats for each patient. The proposed methods analyze the dynamics of individual heartbeats, so initial signal must be segmented into individual heartbeats (R-R interval). This avoids the difficulty to extract waveform features for each heartbeat but only detection of R peak. However, this is still challenging for a variety of reasons. First, the sampling quality is often low, with substantial noise appearing and disappearing across time. Second, it is typically the case that the end of the time series signal is severely corrupted, and no meaningful data is captured in this region. A similar phenomenon often occurs at the start of the signal. These problems with the data necessitate a careful data cleaning stage prior to applying a segmentation algorithm.

Algorithm 1 Heartbeats Segmentation

```
Input: ECG time series
Divide the whole time series into large chunks (e.g., 1000).
for each chunk do
    Threshold based on the largest value in the chunk to find potential $R$ peaks.
end for
Compute the $R$-$R$ interval $d$ for each potential $R$ peak to the next one.
Determine the smallest and largest R-R peak intervals $s$ and $l$ allowed based on the median of all the $d$'s.
while $\exists d \notin [s, l]$ do
    if $d > l$ then                              ▷ deal with peaks too far away
        Search R peaks between
    else if $d < s$ then                         ▷ deal with peaks too close
        Compare the derivative of peaks to find true R peaks.
    end if
end while
```

Our procedure for segmenting the raw ECG data begins with removing the initial 5% and final 20% of the time series for each patient, since these regions of the signal were often very noisy or corrupted. The percentages 5% and 20% were chosen somewhat arbitrarily, and some patients had additional portions of the beginning and end of their time series discarded after this initial pruning stage. The segmentation algorithm is shown in Alg.1. After pruning of the data, each time series was segmented into individual heatbeats by searching for local maxima in the time series corresponding to the R peak in the data. The local maxima of the signal are detected by thresholding the amplitude, and intervals between local maxima are checked to avoid cases with large T wave (too short interval) or small R peak (too long interval). We distinguish R peak and T wave by comparing the first derivative (slope) of the peak. Finally, individual heartbeats are extracted as the regions demarcated by these local maxima and then normalized to the same length ($D = 110$) with interpolation. An example of the original data and a heartbeat segmented from that data appear in Figure 2.1.

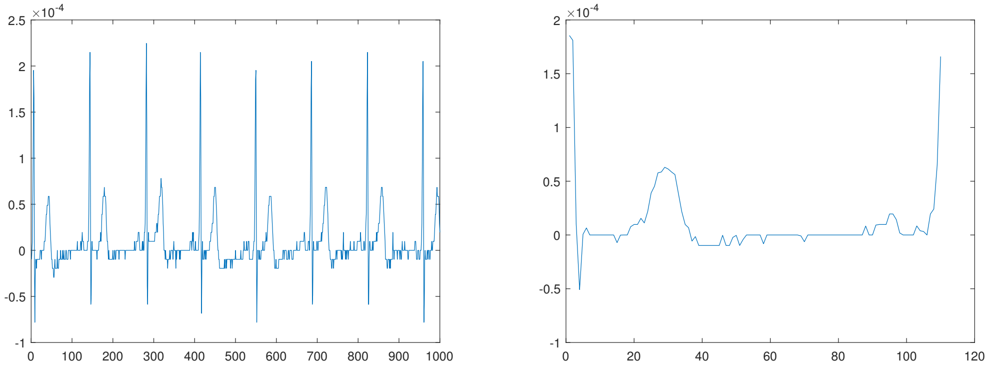
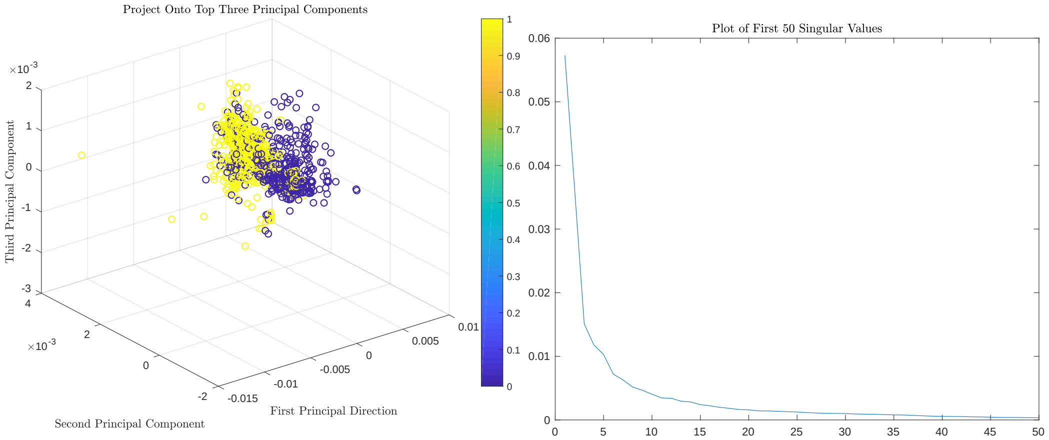


Figure 2.1: (a)An example ECG signal. The first 1000 recorded measurements are displayed. The full signal has length 3948750. A segmented heartbeat with length normalized is shown in (b). The method for segmenting the data searches for the R peak, which is why the heartbeat shown begins and ends with an R peak.

2.2 Learning the Geometry of ECG data

Once the data has been pruned and segmented, it is possible to do analysis of the resulting collection of heartbeats. Our approach is to consider these heartbeats as data in some high dimensional space, which can be analyzed using statistical learning and dimension reduction. We think of the data generated by a patient as a time series of heartbeats $\{x_i\}_{i=1}^{n} \subset \mathbb{R}^D$, where D is the length of a heartbeat, and n is the number of segmented heartbeats. Visual inspection suggests that the space of possible heartbeats may be intrinsically low-dimensional, depending on only a small number of (unobserved) parameters. To investigate this, we constructed data-dependent embeddings of sample of heartbeats from $\mathbb{R}^D$ into $\mathbb{R}^3$ for purposes of visualization. We consider embedding linearly with principal component analysis (PCA), and also nonlinearly by embedding with the eigenvectors of a graph Laplacian; Figures 2.2 and 2.3 show these embeddings.



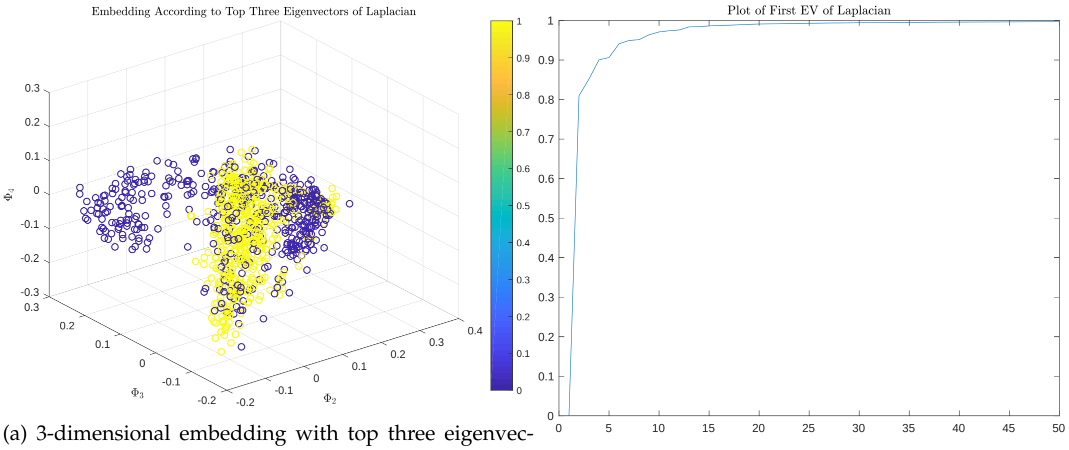
(a) 3-dimensional embedding with top three principal components.

(b) Plot of the singular values of the heartbeat data.

Figure 2.2: Random heartbeats are linearly embedded into 3 dimensions by projecting onto the top three eigenvectors of the covariance matrix of the mean-centered data i.e. the top three principal components. There is some separation between the healthy and HF patients in the linear embedding. The decay of the singular values of the data show some decay, but the data does not appear to live close to a low-dimensional subspace, as even the 50^{th} singular value is nontrivial.

2.2.1 Semi-supervised Graph Classification

One method for labeling a patient as healthy or HF consists in embedding labeled and unlabeled heartbeats from both healthy and HF patients in a common low-dimensional space, and using labeled heartbeats to classify unlabeled heartbeats by proximity. This idea is a form of semisupervised



(a) 3-dimensional embedding with top three eigenvectors of the Laplacian

(b) Plot of the eigenvalues of the graph Laplacian.

Figure 2.3: Random heartbeats represented according to the second, third, and fourth principal eigenvectors of the normalized symmetric graph Laplacian. We see that the data is, with one outlier, quite localized on this three dimensional surface. Moreover, the healthy and HF beats seem to cluster well. We see from the plot of the eigenvectors of the graph Laplacian that the data is approximately low-dimensional, but not with dimension less than say, 20. However, the convergence of the eigenvectors of the Laplacian to 1 is much more rapid than the decay of the singular values toward 0. Though these are not comparable, this suggests that the correlations in the data are nonlinear.

learning on graphs (Belkin and Niyogi, 2002; Belkin and Niyogi, 2004; Szlam, Maggioni, and Coifman, 2008), and bears some resemblance to the method of non-local means (Buades, Coll, and Morel, 2005a; Buades, Coll, and Morel, 2005b). An example of an embedding with the top eigenvectors of the graph Laplacian appear in Figure 2.4.

This semi-supervised classification method proceeds as follows. A set of healthy and HF patients are selected as a test set. The goal is to label the heartbeats for these patients. A training set consisting of heartbeats, both healthy and HF, are sampled from patients not among the training patients. Labels for the training set are provided, while labels for the test set are not. All heartbeats are concatenated into a single data matrix. All heartbeats are embedded into $\mathbb{R}^m$ according to the Laplacian eigenmaps algorithm, where the weight matrix is constructed with Euclidean distances. Each unlabeled

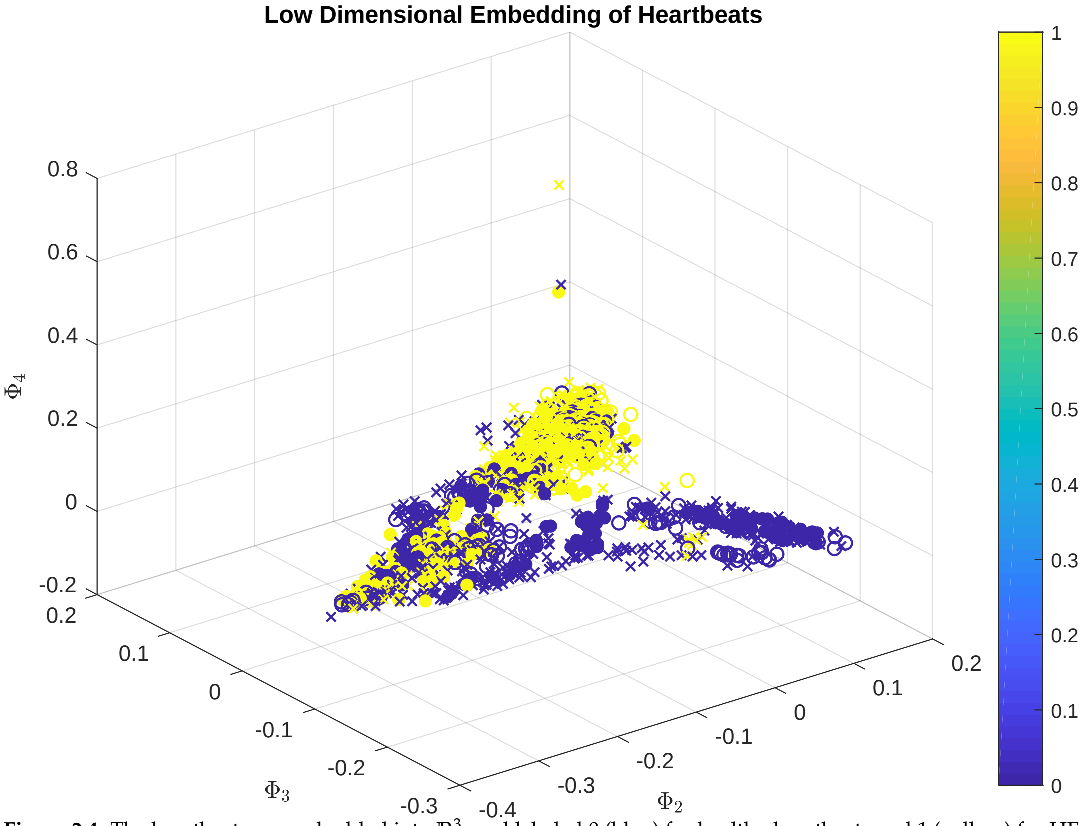


Figure 2.4: The heartbeats are embedded into $\mathbb{R}^3$, and labeled 0 (blue) for healthy heartbeats and 1 (yellow) for HF heartbeats. In the semisupervised labeling method, we use training points ('x' marks), validating data ('o' marks) and testing data (filled 'o' marks). The localization of the colors suggests that a simple classification of healthy or HF based on nearest neighbor in the embedded space may lead to a reasonable classification of heartbeats as healthy or HF. Indeed, using the labels of a heartbeat's nearest neighbors in the low dimensional embedding provides relative good classification accuracy, as indicated in Figure 2.5.

heartbeat is labeled as the most common label among the k nearest neighbors in the embedded domain, excluding heartbeats coming from the same patient as the heartbeat under classification. Once a heartbeat is labeled, a patient may be labeled according to the most common label of their heartbeats.

More precisely, we are given a set of patients Z_{test} that we want to classify as healthy or HF. These patients consist of a collection of heartbeats X_{test} that we want to classify as healthy or HF. Let X_{validate} be heartbeats belonging to patients disjoint from those in Z_{test}. Let X_{train} be heartbeats from patients disjoint

Algorithm 2 Linear Heartbeat Labeling

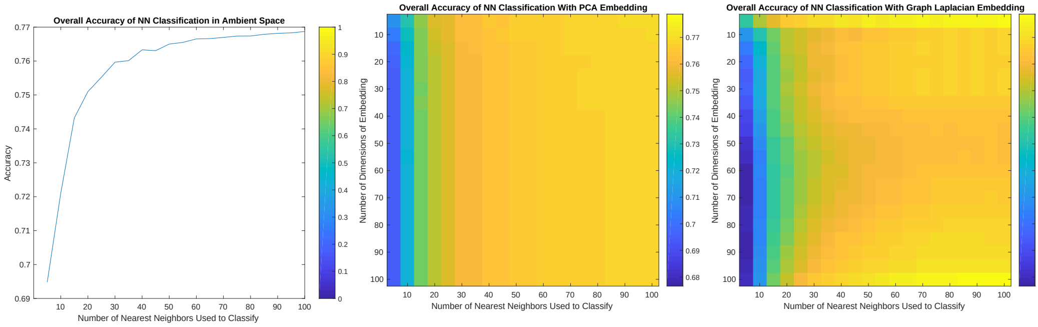
```
Input: $X_{\text{train}}$, $X_{\text{validate}}$, $X_{\text{test}}$; $Y_{\text{train}}$, $Y_{\text{test}}$; $K$, $k$, $m$, $\epsilon$
Output: $\hat{Y}_{\text{validate}}$, $\hat{Y}_{\text{test}}$.
1: Set $X = X_{\text{train}} \cup X_{\text{validate}} \cup X_{\text{test}}$.
2: Compute the principal components of $X$, i.e. the eigenvectors of $X^T X$, call them $u_1, ..., u_D$.
3: Project $X$ onto its top $m$ principal components; call the embedded data $\tilde{X}$.
4: For each point $x^*$ in the validation set, compute its $k$ nearest neighbor among the training data in $\tilde{X}$. Call these nearest neighbors $x_1, ..., x_k$.
5: Label $\hat{Y}_{\text{validate}}(x^*) = \text{mode}(\{Y_{\text{train}}(x_1), Y_{\text{train}}(x_2), ..., Y_{\text{train}}(x_k)\})$.
6: For each point $x^{**}$ in the test set, compute its $k$ nearest neighbor among the training data in $\tilde{X}$. Call these nearest neighbors $x_1^*, ..., x_k^*$.
7: Label $\hat{Y}_{\text{test}}(x^{**}) = \text{mode}(\{Y_{\text{train}}(x_1^*), Y_{\text{train}}(x_2^*), ..., Y_{\text{train}}(x_k^*)\})$.
```

from those in X_{validate} and X_{test}. We have access to the labels for the heartbeats in X_{train} and X_{validate}, call these labels Y_{train} and Y_{validate}, respectively. Our main semisupervised algorithm estimates Y_{test}, the labels for the heartbeats of the test patients Z_{test}. The labels of Z_{test} are subsequently estimated from $\hat{Y}_{\text{test}}$. We consider nearest neighbor classification with the dimension reduced both linearly and nonlinearly.

A patient is then classified as healthy or HF according to the majority rule of her heartbeats' labels. Thus, we compute $\hat{Z}_{\text{test}}$ simply from $\hat{Y}_{\text{test}}$. We note that the algorithm has dependencies on K, k, m, ϵ. Choosing these to maximize the accuracy of the labels of $\hat{Y}_{\text{validate}}$ performs a kind of cross validation, which we employ. Some typical results the accuracy of estimating Z_{test} with $\hat{Z}_{\text{test}}$ appear in Figure 2.5. These results show m, k varying, but $\epsilon = .05$ fixed and K to be the number of data points, so that the graph is fully connected. Similar results hold for $\epsilon = .01, .15, .20, .25$.

Algorithm 3 Nonlinear Heartbeat Labeling

```
**Input:** $X_{\text{train}}, X_{\text{validate}}, X_{\text{test}}; Y_{\text{train}}, Y_{\text{test}}; K, k, m, \epsilon$
**Output:** $\hat{Y}_{\text{validate}}, \hat{Y}_{\text{test}}$.
Set $X = X_{\text{train}} \cup X_{\text{validate}} \cup X_{\text{test}}$.
Form the $K$ nearest neighbors graph $\mathcal{G}$ on $X$ with distances given by the Euclidean distance.
Form the graph Laplacian $L$ of $\mathcal{G}$ with scale parameter $\epsilon$.
Compute the $m$ principal eigenvectors of $L$, call them $\Phi_1, ..., \Phi_m$.
For each point $x^*$ in the validation set, compute its $k$ nearest neighbor among the training data in the embedded space determined by $\Phi_1, ..., \Phi_m$. Call these nearest neighbors $x_1, ..., x_k$.
Label $\hat{Y}_{\text{validate}}(x^*) = \text{mode}(\{Y_{\text{train}}(x_1), Y_{\text{train}}(x_2), ..., Y_{\text{train}}(x_k)\})$.
For each point $x^{**}$ in the test set, compute its $k$ nearest neighbor among the training data in the embedded space determined by $\Phi_1, ..., \Phi_m$. Call these nearest neighbors $x_1^*, ..., x_k^*$.
Label $\hat{Y}_{\text{test}}(x^{**}) = \text{mode}(\{Y_{\text{train}}(x_1^*), Y_{\text{train}}(x_2^*), ..., Y_{\text{train}}(x_k^*)\})$.
```



(a) Classification error in ambient space. Optimal accuracy is 76.87%.

(b) Classification error using linear PCA embedding. Optimal accuracy is 77.02%.

(c) Classification error using nonlinear Laplacian embedding. Optimal accuracy is 77.78%.

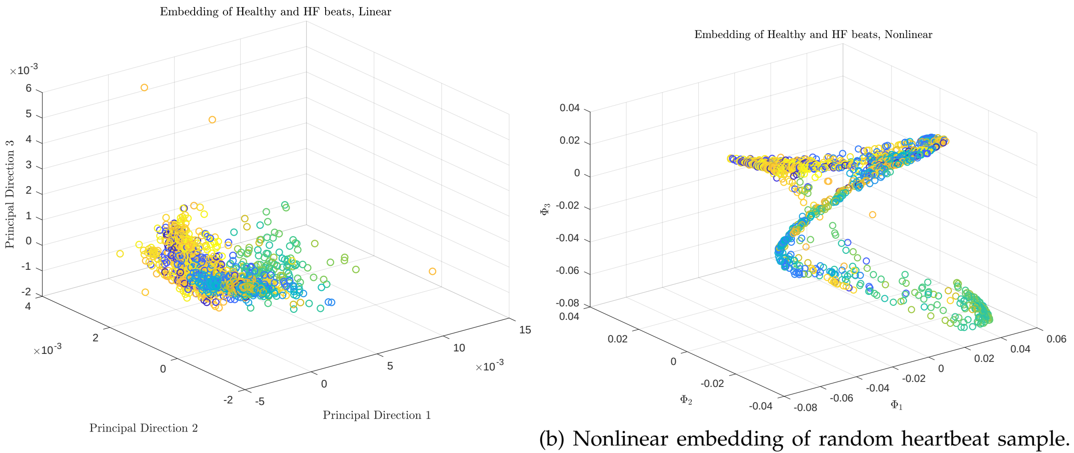
Figure 2.5: Classification accuracy of the proposed methods are shown, with variation depending on the number of eigenvectors used in the low dimensional embedding, along with the number of nearest neighbors used to classify. 1000 trials of training and testing sets are shown, with results averaged. Results are fairly consistent after a sufficient number of eigenvectors are used. The optimal choice of parameters for the linear embedding slightly improves over classifying in the ambient space, while using nonlinear embedding improves over linear embedding by .75%.

With an appropriately chosen number of eigenvectors used in the low-dimensional embedding, d, and number of nearest neighbors, k, both the linear and nonlinear methods exceed 78% in accuracy, with the nonlinear method performing better.

2.2.2 Space of All Heartbeats

Instead of using manifold learning to classify patients as healthy or at risk of heart failure, one can analyze the space of all possible heartbeats. For simplicity, we will refer the space of all heartbeats to "global" space. Natural clusters may form in this space, and it is interesting to observe the trajectories of a single patient in this larger space.

We first consider 3-dimensional linear and nonlinear embeddings of a random sample of heartbeats. The graph we consider is fully connected, so we are limited in the number of beats we can consider in a single sample. We consider 10000 randomly sampled healthy and HF heartbeats, for a total of 40000 samples. We then embed the data according to the top three principal directions and the top three eigenvectors of the graph Laplacian; these images are in Figure 2.6.

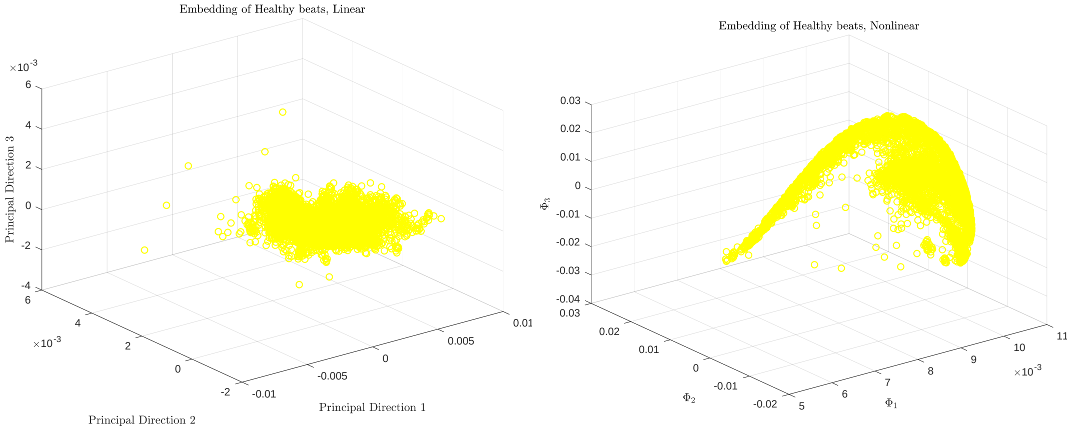


(a) Linear embedding of random heartbeat sample. There is clear separation between the healthy and HF beats.

(b) Nonlinear embedding of random heartbeat sample. There is clear separation between the healthy and HF beats.

Figure 2.6: Whether the data is embedded linearly or nonlinearly, there is obvious separation between healthy and HF beats. This global separation suggests the value in the proposed semisupervised manifold learning method.

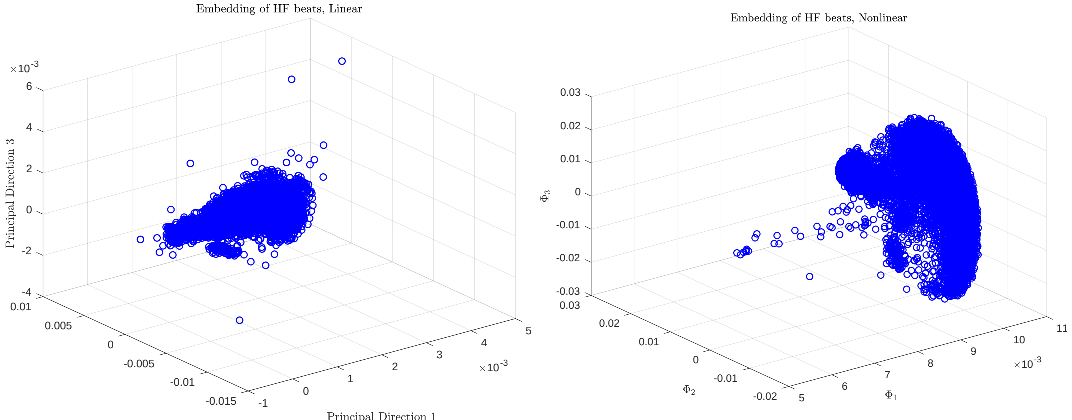
It is also of interest to observe the healthy and HF embedding spaces separately. To do so, we take samples of only healthy or HF patients, and compute the low-dimensional embeddings. Example embedded datasets are in Figures 2.7 and 2.8, respectively. Notice that the shapes of the embeddings are comparable to those in Figure 2.6, despite the completely different sampling methods used.



(a) Linear embedding of random healthy heartbeat sample.

(b) Nonlinear embedding of random healthy heartbeat sample.

Figure 2.7: The healthy data forms a relatively compact cluster in the linear embedding, but there is substantial variation in the nonlinear embedding. Outliers appear to be present in both cases.



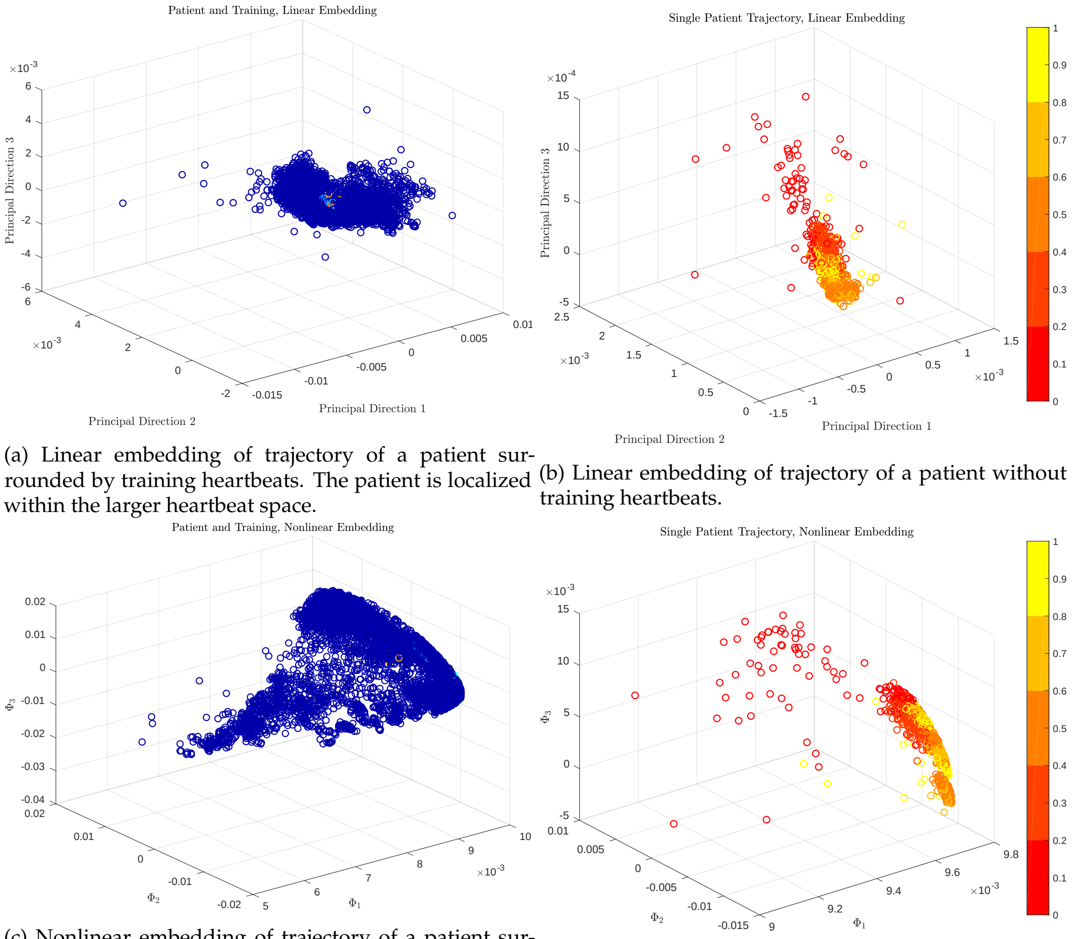
(a) Linear embedding of random HF heartbeat sample.

(b) Nonlinear embedding of random HF heartbeat sample.

Figure 2.8: The HF data appears compact in both the linear and nonlinear embeddings.

We also consider mapping the trajectory of the heartbeats of a single patient.

To do so, we take a random sample of 10000 training heartbeats, as well as 1000 samples from the time series of a single patient, and embed them jointly. We then observe the patient's trajectory in the larger embedding; illustrations for both linear and nonlinear embeddings are in Figure 2.9.



(a) Linear embedding of trajectory of a patient surrounded by training heartbeats. The patient is localized within the larger heartbeat space.

(b) Linear embedding of trajectory of a patient without training heartbeats.

(c) Nonlinear embedding of trajectory of a patient surrounded by training heartbeats. The patient is localized within the larger heartbeat space.

(d) Linear embedding of trajectory of a patient without training heartbeats.

Figure 2.9: The trajectory of a single patient's time series is relatively well-localized in the ambient embedding, but does show some time-correlated variation.

2.3 Modeling ECG Dynamics with Markov Model

A number of previous studies have shown the difference in dynamics of physiological signals from healthy and unhealthy patients(Buchman, 2002; Ivanov et al., 1996). Statistical analysis of the hidden dynamics (Goldberger, 1990; Ivanov et al., 1999; DeMazumder et al., 2016) revealed that healthy subjects are dynamically stable over a wide range of timescale. On the other hand, automatic ECG analysis has been applied to facilitate decision making and reduce costs as early as 1970s. However, its sensitivity has been limited in the cases of ST elevation myocardial infarction (STEMI). Recent advances in machine learning has enabled great improvement of performance in many difficult problems in this area (Rajpurkar et al., 2017; Voisin et al., 2018).

The proposed manifold learning method detailed and evaluated in Section 2.2 essentially characterizes a heartbeat as depending on a small number of parameters that lie near a low-dimensional manifold. Evaluating a patient as healthy or unhealthy was determined based on local proximity in this manifold embedding, and hence depended essentially on the typical shape of a patient's heartbeat. An alternative characterization of a patient as healthy or in danger of heart failure is to study the *dynamics* of a patient's heartbeats, and make predictions on the wellness of a patient based on subsequent dynamical statistics. One advantage of analyzing the dynamics of heartbeats, compared to manifold learning, is that dynamical analysis explicitly accounts for the time-evolving nature of the heartbeats. Whereas the manifold learning method discards the time structure, we propose a method that incorporate this time structure into statistics on the data.

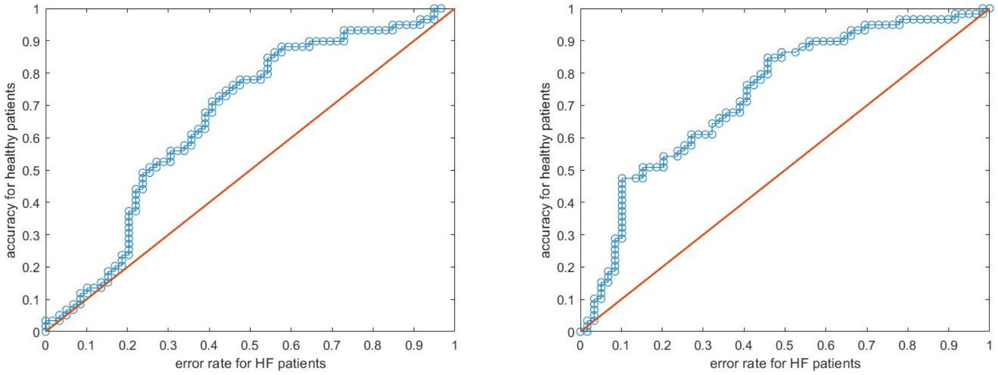
We consider modeling a patient's heartbeat state $\mathcal{H}$ as a Markov model, in which their space of heartbeats is partitioned into a set of possible states, $C_1, C_2, ..., C_K$. One can then discuss the probabilities of transitioning from cluster C_i to C_j as encoding a transition probability on $\mathcal{H}$. This yields a Markov transition matrix $P \in \mathbb{R}^{K \times K}$, where P_{ij} corresponds to the probability of transitioning from state C_i to C_j. The empirical matrices P generated by different patients can then be used for classification. As in Section 2.2, we consider two finite state space in the Markov model: one consists of heartbeats of a single patient ("local" heartbeat space), the other consists of heartbeats from all possible patients ("global" heartbeat space). The goal is to find useful features that could capture the ECG dynamics and facilitate decision making.

2.3.1 Local Heartbeat Space

The predictability of the above Markov chain model (the transition matrix P) is evaluated by comparing with naive transition matrix $\bar{P}$ where each row is the stationary distribution π. The motivation for comparing with the stationary distribution π is as follows. Under mild assumptions, a Markov chain has a stationary distribution $\pi \in \mathbb{R}^{1 \times K}$ such that $\pi P = \pi$. This encodes the long-term probability of being at a given state: for any initial distribution $\pi_0, \lim_{t \to \infty} \pi_0 P^t = \pi$. We run the following tests to validate the proposed model. For each patient, we train a Markov transition matrix P using the first half of the time series and compute the stationary distribution π which is used to construct the transition matrix $\bar{P}$. Divide the second half of the time series equally into small pieces with 10 heartbeats, and compute the log likelihood for each piece using P and $\bar{P}$ respectively. The result shows that for all patients, log likelihood computed from P is greater than $\bar{P}$ for most pieces, suggesting that the Markov chain model could effectively capture the dynamics of the random process.

We consider two tests on the **long-term dynamics** of a patient's heartbeats, by examining the *second (Fiedler) eigenvalue,* λ_2 of the Markov transition matrix P, and also the *stationary distribution* π of the Markov transition matrix P on $\mathcal{H}$. The motivation for considering $\lambda_2 < 1$ is that this quantity governs the mixing time of the Markov process, in the sense that the rate of convergence of the chain towards its stationary distribution is exponential with base λ_2 (Sinclair and Jerrum, 1989). So, the smaller λ_2 is, the more rapidly the Markov chain is mixing. We hypothesized that healthy patients would take less time to converge since their dynamics more stable, and therefore π would be more concentrated in the case of HF patients. We thus consider using $\|\pi\|_\infty$ to discriminate between healthy and unhealthy patients. We consider a simply binary classifier for distinguishing between healthy and HF patients, based on λ_2 and $\|\pi\|_\infty$ being above or below a given threshold. Receiver operating characteristic (ROC) curves for these classifiers appear in Figure 2.10.

The area under the ROC curves (AUC) for classifying based on λ_2 is small, and illustrate poor classification performance. However, the AUC for classifying based on $\|\pi\|_\infty$ indicates a better performance. These tests were run on time series from 59 healthy and 59 HF patients. For each patient with data $\{x_i\}_{i=1}^N$, we cluster the heartbeats with K-means. We chose $K = 20$ to allow for many clusters of varying sizes. We then build the Markov transition matrix



(a) ROC curve corresponding to classifying based on λ_2. Area under ROC curve is 0.627.

(b) ROC curve corresponding to classifying based on $\|\pi\|_\infty$. Area under ROC curve is 0.734.

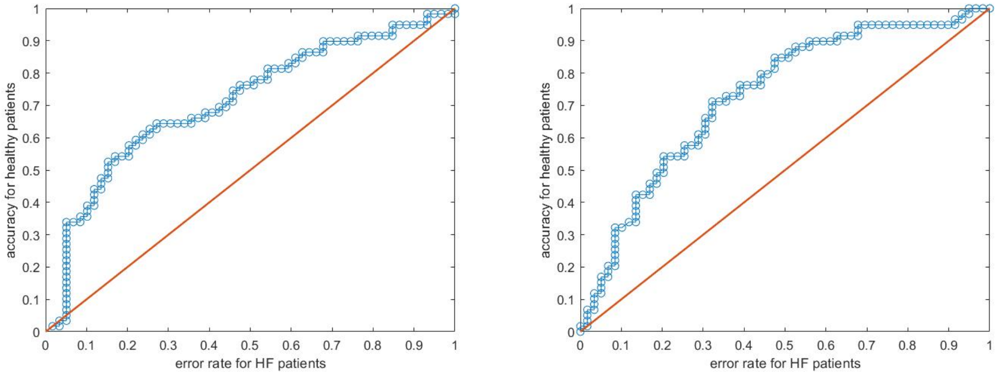
Figure 2.10: The ROC curve with classifier λ_2 and $\|\pi\|_\infty$ in local heartbeats space.

by setting $P_{ij} = |\{t \mid x_t \in C_i, x_{t+1} \in C_j\}|$. We then randomize this matrix by adding a small perturbation to avoid the case of missing observations, and normalize this weight matrix to be row stochastic, i.e. Markovian. In this way, each patient's dynamics is characterized by one Markov matrix P. We threshold on λ_2 and $\|\pi\|_\infty$ respectively to distinguish the two groups of patients.

We also consider two tests on the **short-term dynamics**. Similarly, we would use the above transition matrix P to compute the statistics for the second half of the time series. Here we use log-likelihood as the statistics to model the random process:

$$\begin{aligned}\tau &= \frac{1}{T-1} \log \prod_{t=1}^{T-1} P(id(x_t), id(x_{t+1})) \\ &= \frac{1}{T-1} \sum_{t=1}^{T-1} \log P(id(x_t), id(x_{t+1}))\end{aligned}$$

, since this quantity characterize the probability of seeing a specific trajectory. The entire time series is equally divided into small pieces of size T, and τ is computed for each piece. Afterwards, we compute the mean μ and standard deviation σ of the τ's for all pieces for one patient. The ROC curves by thresholding on μ and σ are shown in Fig. 2.11 .



(a) ROC curve corresponding to classifying based on μ. Area under ROC curve is 0.712.

(b) ROC curve corresponding to classifying based on σ. Area under ROC curve is 0.730.

Figure 2.11: The ROC curve with classifier μ and σ of τ in local heartbeats space.

The algorithm is specified in Alg. 4. The performance of all four statistics is summarized in Table . The accuracy is from linear support vector machine (SVM) and 5-fold cross-validation.

Statistics	accuracy	AUC
λ_2	0.593	0.627
$\|\pi\|_\infty$	0.678	0.734
μ	0.644	0.712
σ	0.678	0.730

Table 2.1: Performance of statistics in local heartbeat space

Algorithm 4 Classifier based on dynamics in local heartbeat space

```
Input: Time series $\{x_i\}_{i=1}^N$ for each patient, $K$, $T$
Output: $\lambda_2$,$\pi$, $\mu$, $\sigma$ for each patient.
1: Take the time series data of one patient, cut it into two parts with equal length.
2: Use the first half as training set. Do $k$-means with $K = 20$ and return the centroid of each cluster $\{C_i\}_{i=1}^K$. Construct the Markov transition matrix $P \in \mathbb{R}^{K \times K}$. Randomize $P$ by adding a small perturbation $10^{-6}I$, and then normalize each row.
3: Compute $\lambda_2$ and $\pi$ of $P$.
4: Use the second half as testing set. Assign label $\{id_i\}_{i=1}^N$ to each heartbeat based on its distance to $\{C_i\}_{i=1}^K$.
5: Divide the testing set equally into $\lfloor \frac{N}{2T} \rfloor$ fragments of equal size $T = 10$ and within fragment $i$, we compute
   $$\tau_i = \log_{10} \prod_{t=1}^{T-1} P(id(x_t), id(x_{t+1})) = \sum_{t=1}^{T-1} \log_{10} P(id(x_t), id(x_{t+1}))$$
   and the mean $\mu$ and standard deviation $\sigma$ of the vector $\left[\tau_1 \cdots \tau_{\lfloor \frac{N}{2T} \rfloor}\right]$.
```

2.3.2 Global Heartbeat Space

The above experiments could also be done in global heartbeat space. The global heartbeat space is constructed by sampling a large number of heartbeats from all patients. One key difference in global heartbeat space is that the distance metric we used is correlation, not Euclidean as in local heartbeat space:

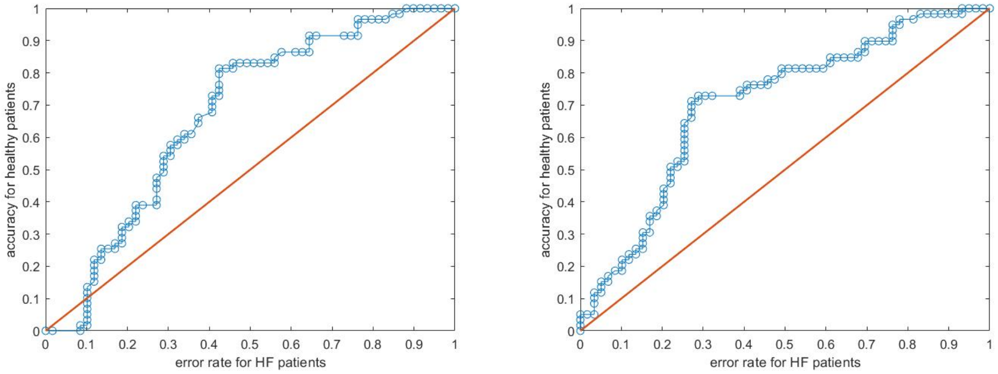
$$d(x,y) = 1 - \frac{(x-\bar{x})^T(y-\bar{y})}{\sqrt{(x-\bar{x})^T(x-\bar{x})}\sqrt{(y-\bar{y})^T(y-\bar{y})}} \tag{2.1}$$

Euclidean distance is sensitive to translation which makes the heartbeats from one patient concentrate in only a few clusters, or even one and thus the Markov chain could not model the transitions between different states.

Algorithm 5 Classifier based on dynamics in global heartbeat space

```
Input: Time series $\{x_i\}_{i=1}^{N}$ for each patient, $K$, $T$
Output: $\lambda_2$,$\pi$, $\mu$, $\sigma$, $\gamma$ for each patient.
Load time series signals of all patients. For each patient, divide the time series equally into two parts.
Use the first half of time series from each patient to construct a global heartbeat space $\mathcal{H}$. Do $k$-means with $K = 50$ and return the centroid of each cluster $\{C_i\}_{i=1}^{K}$.
For each patient, construct the Markov transition matrix $P \in \mathbb{R}^{K \times K}$ with the first half of time series. Randomize $P$ by adding a small perturbation $10^{-6}I$, and normalize each row.
Compute $\lambda_2$ and $\pi$ of $P$.
Divide the second half of the time series equally into $\lfloor \frac{N}{2T} \rfloor$ fragments of equal size $T = 10$. Within fragment $i$, compute
    $\tau_i = \log_{10} \prod_{t=1}^{T-1} P(id(x_t), id(x_{t+1})) = \sum_{t=1}^{T-1} \log_{10} P(id(x_t), id(x_{t+1}))$
    and the mean $\mu$, standard deviation $\sigma$ and skewness $\gamma$ of the vector $\left[\tau_1 \cdots \tau_{\lfloor \frac{N}{2T} \rfloor}\right]$.
```

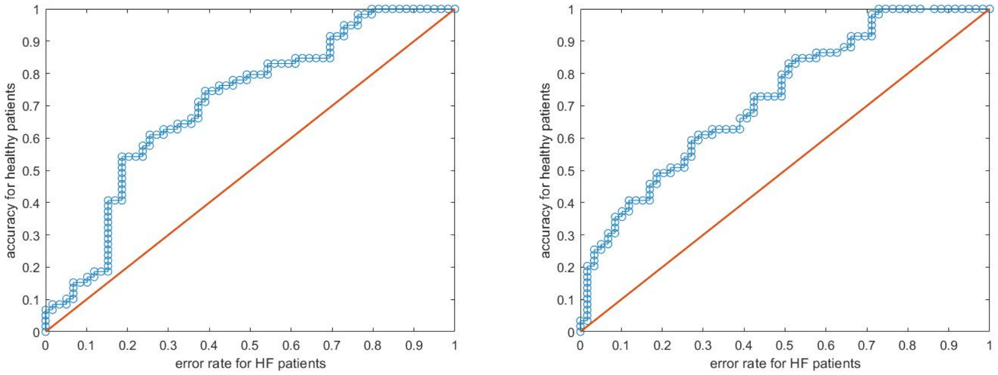
We found that the performance of σ is poor, so instead we plot the performance of skewness γ. The ROC curves of the four statistics are shown in Fig. 2.12 and Fig. 2.13.



(a) ROC curve corresponding to classifying based on λ_2. Area under ROC curve is 0.672.

(b) ROC curve corresponding to classifying based on $\|\pi\|_\infty$. Area under ROC curve is 0.688.

Figure 2.12: The ROC curve with classifier λ_2 and $\|\pi\|_\infty$ in global heartbeats space.



(a) ROC curve corresponding to classifying based on μ. Area under ROC curve is 0.704.

(b) ROC curve corresponding to classifying based on γ. Area under ROC curve is 0.726.

Figure 2.13: The ROC curve with classifier μ and γ of τ in global heartbeats space.

References

Vullings, HJLM, MHG Verhaegen, and HB Verbruggen (1998). “Automated ECG segmentation with dynamic time warping”. In: *Engineering in Medicine and Biology Society, 1998. Proceedings of the 20th Annual International Conference of the IEEE*. IEEE, pp. 163–166.

Vullings, HJLM, MHG Verhaegen, and Henk B Verbruggen (1997). “ECG segmentation using time-warping”. In: *International Symposium on Intelligent Data Analysis*. Springer, pp. 275–285.

Andreao, Rodrigo Varejao, Bernadette Dorizzi, and Jérôme Boudy (2006). “ECG signal analysis through hidden Markov models”. In: *IEEE Transactions on Biomedical engineering* 53.8, pp. 1541–1549.

Hughes, Nicholas P, Stephen J Roberts, and Lionel Tarassenko (2004). “Semi-supervised learning of probabilistic models for ECG segmentation”. In: *IEEE Engineering in Medicine and Biology Conference (EMBC)*.

Belkin, M. and P. Niyogi (2002). “Laplacian eigenmaps and spectral techniques for embedding and clustering”. In: *Advances in neural information processing systems*, pp. 585–591.

Belkin, M. and P. Niyogi (2004). “Semi-supervised learning on Riemannian manifolds”. In: *Machine Learning* 56.1-3, pp. 209–239.

Szlam, A.D., M. Maggioni, and R.R. Coifman (2008). “Regularization on graphs with function-adapted diffusion processes”. In: *Journal of Machine Learning Research* 9.Aug, pp. 1711–1739.

Buades, A., B. Coll, and J.M. Morel (2005a). “A non-local algorithm for image denoising”. In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 2. IEEE, pp. 60–65.

Buades, A., B. Coll, and J.M. Morel (2005b). “A review of image denoising algorithms, with a new one”. In: *Multiscale Modeling & Simulation* 4.2, pp. 490–530.

Buchman, Timothy G (2002). “The community of the self”. In: *Nature* 420.6912, p. 246.

Ivanov, Plamen Ch, Michael G Rosenblum, C-K Peng, Joseph Mietus, Shlomo Havlin, H Eugene Stanley, and Ary L Goldberger (1996). “Scaling behaviour of heartbeat intervals obtained by wavelet-based time-series analysis”. In: *Nature* 383.6598, p. 323.

Goldberger, Ary L (1990). “Fractal Electrodynamics of the Heartbeat”. In: *Annals of the New York Academy of Sciences* 591.1, pp. 402–409.

Ivanov, Plamen Ch, Luis A Nunes Amaral, Ary L Goldberger, Shlomo Havlin, Michael G Rosenblum, Zbigniew R Struzik, and H Eugene Stanley (1999). “Multifractality in human heartbeat dynamics”. In: *Nature* 399.6735, p. 461.

DeMazumder, D., W.B. Limpitikul, M. Dorante, S. Dey, B. Mukhopadhyay, Y. Zhang, J.R. Moorman, A. Cheng, R.D. Berger, E. Guallar, and S.R. Jones (2016). “Entropy of cardiac repolarization predicts ventricular arrhythmias and mortality in patients receiving an implantable cardioverter-defibrillator for primary prevention of sudden death”. In: *EP Europace* 18.12, pp. 1818–1828.

Rajpurkar, Pranav, Awni Y Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y Ng (2017). “Cardiologist-level arrhythmia detection with convolutional neural networks”. In: *arXiv preprint arXiv:1707.01836*.

Voisin, Maxime, Yichen Shen, Alireza Aliamiri, Anand Avati, Awni Hannun, and Andrew Ng (2018). “Ambulatory Atrial Fibrillation Monitoring Using Wearable Photoplethysmography with Deep Learning”. In: *arXiv preprint arXiv:1811.07774*.

Sinclair, A. and M. Jerrum (1989). “Approximate counting, uniform generation and rapidly mixing Markov chains”. In: *Information and Computation* 82.1, pp. 93–133.

Chapter 3

SaO_2 Data Analysis

Oxygen saturation has been proposed as an useful tool to provide cardiorespiratory information due to its low cost, convenience and ability to provide immediate and continuous values. However, important technical limitations, lack of interpretation of data, as well as lack of sensitivity all decrease the value of oxygen saturation as a single diagnostic tool (Netzer et al., 2001; Mower et al., 1996; DeMeulenaere, 2007; Sinex, 1999; Hayes and Smith, 2001; Runciman et al., 1993). Therefore, the oxygen saturation data cannot replace the ECG as the sole standard for patient selection in ICD implantation, but rather provide additional and complementary information to better identify patient set.

As introduced in Chapter 1, SaO_2 tends to remain constant around 96%. A closer look reveals that the SaO_2 samples take 21 discrete values in the range between 80 and 100, with fewer samples under 90. By checking the available dataset, we also found that there are SaO_2 samples with values near 0. We would like to treat these points as outliers due to measurement instruments and overlook these samples. For each patient, the number of samples in a typical SaO_2 time series is at the order of 10^4.

3.1 Nonparametric Statistical Analysis

The intuition of performing multivariate analysis of SaO_2 data stems from two aspects. First, previous studies have revealed that the levels and significant changes of oxygen saturation are closely related to certain kinds of cardiorespiratory diseases, such as sleep apena, breathing disorder, and pickwickian syndrome (Lloyd-Owen et al., 1999; Javaheri et al., 1999; Olson, Ambrogetti, and Gyulay, 1999). In addition, common statistics from time and frequency domain analyses of blood oxygen saturation recordings have shown to be simple and accurate in the diagnosis of obstructive sleep apnea (Alvarez et al., 2010).

We consider first to fourth order statistical moments in the time domain (Alvarez et al., 2010):

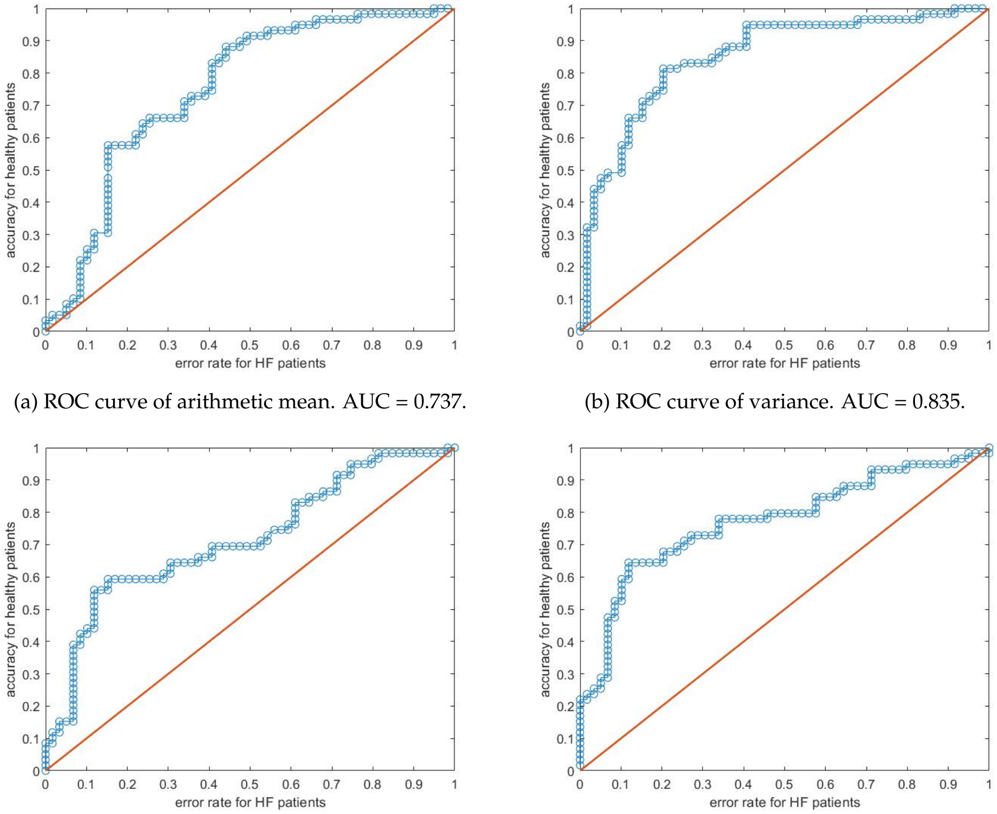
$$M1 = E[x] = \mu = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{3.1}$$

$$M2 = E[(x-\mu)^2] \tag{3.2}$$

$$M3 = \frac{1}{\sigma^3}E[(x-\mu)^3] \tag{3.3}$$

$$M4 = \frac{1}{\sigma^4}E[(x-\mu)^4] \tag{3.4}$$

I.e., we use arithmetic mean ($M1$), variance ($M2$), skewness ($M3$), and kurtosis ($M4$) in the time domain which were derived from each SaO_2 recording to quantify central tendency, amount of dispersion, symmetry/asymmetry, and



(a) ROC curve of arithmetic mean. AUC = 0.737.

(b) ROC curve of variance. AUC = 0.835.

(c) ROC curve of skewness. AUC = 0.716.

(d) ROC curve of kurtosis. AUC = 0.774.

Figure 3.1: ROC curves of first to fourth order statistics in the time domain.

tail extremity, respectively. We found that healthy subjects have higher arithmetic mean and kurtosis but lower variance and skewness. This is consistent with the intuition that healthy subjects have higher SaO_2 levels and lower fluctuation and asymmetry. The ROC curve of all four statistics in time-domain are shown in Fig. 3.1.

The difference in time domain statistics suggest that it might be helpful to use the (empirical) probability mass function and cumulative distribution function of data as feature vector. We applied *k*-nearest neighbors combined with 100 repeated random sub-sampling validation on the patient set, with an average accuracy of 73.8% and 74.0% respectively.

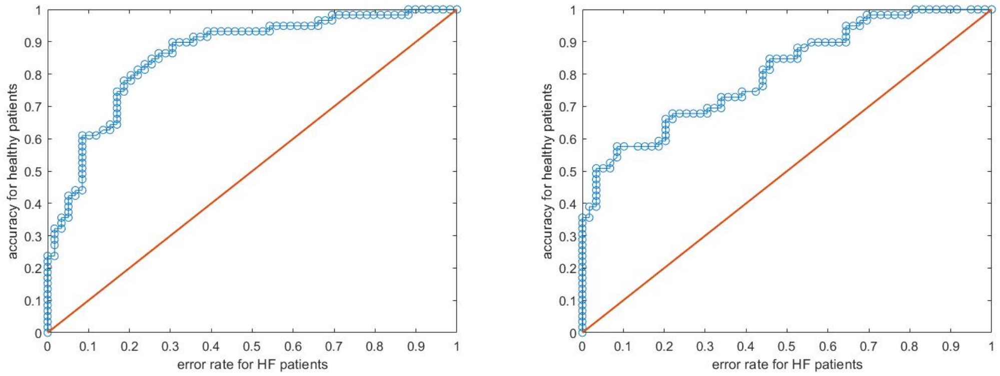
3.2 Dynamics of SaO_2 with Markov Chain

In Section 2.3 we have introduced how to model the ECG dynamics with Markov chain, where the state space is constructed by clusters of heartbeats in $\mathbb{R}^D$. Similar method could be applied to model the dynamics of SaO_2 signal as well. Here we consider the "global" SaO_2 space consisting of all possible SaO_2 values which is one-dimensional. It is natural to choose $K = 20$ for the global SaO_2 space since there are 21 discrete values with few samples with values under 90. The log likelihood is computed for pieces of length $T = 10$. Afterwards, the first to third order moments are computed for all the τ's, and ROC curves are plotted for each of the three moments respectively. The algorithm is specified in Alg. 6.

We hypothesized that the dynamics for healthy subjects is more stable since the SaO_2 values maintain at a high level. As a result, the Markov transition matrix built from earlier time could better model the dynamics in the future and thus the mean of log likelihood is higher for healthy subjects. For the same reason, we also hypothesized that the log likelihood for healthy subjects will slightly fluctuate around the mean and the skewness will be lower. As shown in Fig. 3.2, the mean and skewness of log-likelihood show good performance while the standard deviation performs poorly and thus not plotted here.

Algorithm 6 Classifier based on dynamics in global SaO_2 space

```
Input: $SaO_2$ time series $\{x_i\}_{i=1}^N$ for each patient, $K$, $T$
Output: $\lambda_2$, $\pi$, $\mu$, $\sigma$, $\gamma$ for each patient.
1: Load time series signals of all patients. For each patient, divide the time series equally into two parts.
2: Use the first half of time series from each patient to construct a global $SaO_2$ space $\mathcal{H}$. Do $k$-means with $K = 20$ and return the centroid of each cluster $\{C_i\}_{i=1}^K$.
3: For each patient, construct the Markov transition matrix $P \in \mathbb{R}^{K \times K}$ with the first half of time series. Randomize $P$ by adding a small perturbation $10^{-6}I$, and normalize each row.
4: Divide the second half of the time series equally into $\lfloor \frac{N}{2T} \rfloor$ fragments of equal size $T = 10$. Within fragment $i$, compute
       $\tau_i = \log_{10} \prod_{t=1}^{T-1} P(id(x_t), id(x_{t+1}))$
       $= \sum_{t=1}^{T-1} \log_{10} P(id(x_t), id(x_{t+1}))$
   and the mean $\mu$, standard deviation $\sigma$ and skewness $\gamma$ of the vector $\left[\tau_1 \cdots \tau_{\lfloor \frac{N}{2T} \rfloor}\right]$.
```



(a) ROC curve corresponding to classifying based on μ. Area under ROC curve is 0.860.

(b) ROC curve corresponding to classifying based on γ. Area under ROC curve is 0.804

Figure 3.2: The ROC curve with classifier μ and γ of τ.

References

Netzer, Nikolaus, Arn H Eliasson, Cordula Netzer, and David A Kristo (2001). “Overnight pulse oximetry for sleep-disordered breathing in adults: a review”. In: *Chest* 120.2, pp. 625–633.

Mower, WR, C Sachs, EL Nicklin, P Safa, and LJ Baraff (1996). “A comparison of pulse oximetry and respiratory rate in patient screening”. In: *Respiratory medicine* 90.10, pp. 593–599.

DeMeulenaere, Susan (2007). “Pulse oximetry: uses and limitations”. In: *The Journal for Nurse Practitioners* 3.5, pp. 312–317.

Sinex, James E (1999). “Pulse oximetry: principles and limitations”. In: *The American journal of emergency medicine* 17.1, pp. 59–66.

Hayes, Matthew James and Peter R Smith (2001). “A new method for pulse oximetry possessing inherent insensitivity to artifact”. In: *IEEE Transactions on Biomedical Engineering* 48.4, pp. 452–461.

Runciman, WB, RK Webb, L Barker, and M Currie (1993). “The pulse oximeter: applications and limitations − an analysis of 2000 incident reports”. In: *Anaesthesia and intensive care* 21.5, pp. 543–550.

Lloyd-Owen, SJ, A Crawford, MR Partridge, and CM Roberts (1999). “Clinical value and cost of a respiratory sleep-related breathing disorders screening service for snorers referred to a District General Hospital ENT department”. In: *Respiratory medicine* 93.7, pp. 454–460.

Javaheri, Shahrokh, Maqbool Ahmed, Thomas J Parker, and Candice R Brown (1999). “Effects of nasal O2 on sleep-related disordered breathing in ambulatory patients with stable heart failure”. In: *Sleep* 22.8, pp. 1101–1106.

Olson, LG, A Ambrogetti, and SG Gyulay (1999). “Prediction of sleep-disordered breathing by unattended overnight oximetry”. In: *Journal of sleep research* 8.1, pp. 51–55.

Alvarez, Daniel, Roberto Hornero, J Victor Marcos, and Felix del Campo (2010). “Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis". In: *IEEE Transactions on Biomedical Engineering* 57.12, pp. 2816–2824.

Chapter 4

Patient Identification with Physiological Signals

4.1 Feature Selection

From previous chapters, we have found statistics for ECG and SaO_2 signals respectively. Therefore, for each subject, we could ensemble all predictive statistics to represent each patient as a vector. We included all features from previous sections that have AUC higher than 0.7. Each patient is represented by a feature vector $x = [x_1 \dots x_{12}]$, where each x_i is shown in Table. 4.1.

Since the total number of features is small, we could simply use greedy

x_1	ratio of healthy heartbeats
x_2	$\|\pi\|_\infty$ in local heartbeat space
x_3	mean of log-likelihood in local heartbeat space
x_4	standard deviation of log-likelihood in local heartbeat space
x_5	mean of log-likelihood in global heartbeat space
x_6	skewness of log-likelihood in global heartbeat space
x_7, x_8, x_9, x_{10}	first to fourth moment of SaO_2 in time domain
x_{11}	mean of log-likelihood in global SaO_2 space
x_{12}	skewness of log-likelihood in global SaO_2 space

Table 4.1: Feature set used in patient classification.

forward selection to find the optimal feature subset. The idea of the algorithm is to start from an empty feature set and greedily search for the best feature set with j components in each step, where j is the step number. The algorithm is shown in Alg. 7.

Algorithm 7 Forward Selection

```
Initialize $\mathcal{F} = \emptyset$
loop
for $i = 1, \ldots, n$ do
    if $i \notin \mathcal{F}$ then
        $\mathcal{F}_i = \mathcal{F} \cup \{i\}$
        Use cross validation to evaluate feature set $\mathcal{F}_i$
    end if
end for
Set $\mathcal{F}$ to be the best feature subset found from all $\mathcal{F}_i$'s.
end loop
```

The termination condition for the outer loop can be either $\mathcal{F} = \{1, \ldots, n\}$ or number of features selected reach the expectation.

We used logistic regression and 5-fold cross validation in the above forward selection algorithm. With feature subset $[x_1, x_2, x_5, x_{12}]$, we got the accuracy of 93.2%. Notice that the features are from geometry, dynamics of ECG as well as dynamics of SaO_2, which means that the information contained from these statistics are complementary.

The confusion matrix is shown in Fig. 4.1. The algorithm performs well on both healthy and HF subjects, with no bias towards any of the two classes.

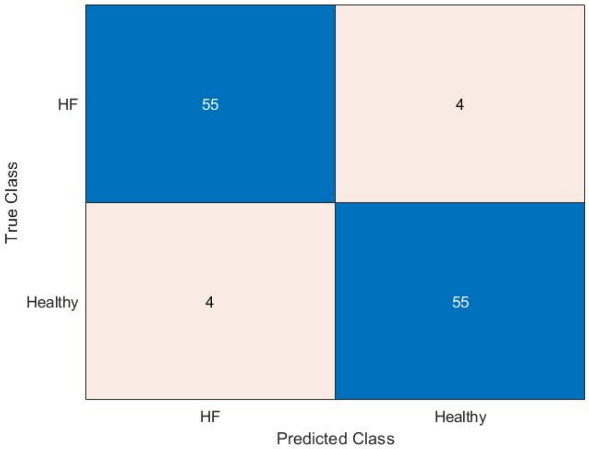


Figure 4.1: Confusion matrix of logistic regression on optimal feature set $x = [x_1, x_2, x_5, x_{12}]$ with 5-fold cross validation.

4.2 Regression Analysis of ECG and SaO_2

Previous results suggest that the information contained in ECG and SaO_2 time series signals can enhance one another regarding the classification accuracy. It is of interest to explore the relationships between the two signals, as they are collected simultaneously. In this section we will discuss two regression analysis done on these signals.

The first regression analysis is on the original ECG and SaO_2 time series. For each SaO_2 sample, we will locate the corresponding ECG sample and find the nearest R peak value. The result is shown in Fig. 4.2. For visualization purpose, the SaO_2 from healthy subjects are right shifted by 0.1. From the figure we can see that for both healthy and HF subjects, each SaO_2 value maps to the ECG peak of 0.0012. However, for HF subjects, this is not the case when oxygen saturation level is high. We also compute the mapping accuracy which is defined as the ratio of number of samples with most ECG values divided by the total number of samples for each SaO_2 level. For healthy subjects the

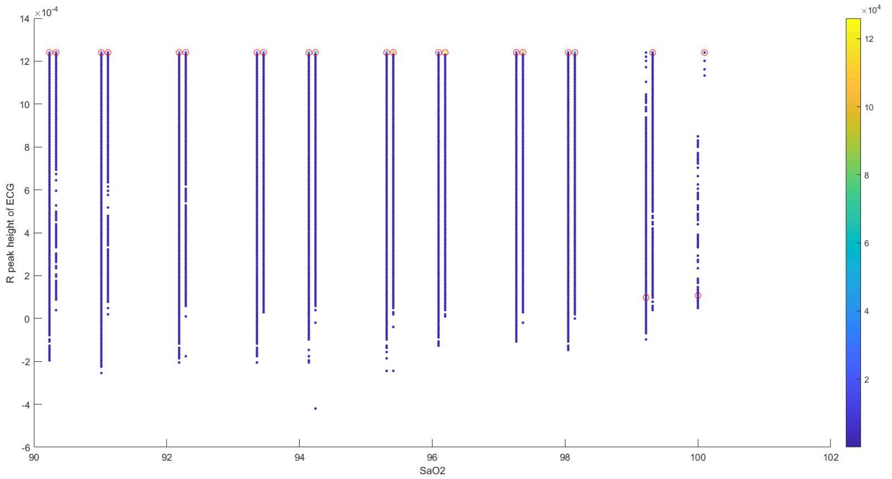


Figure 4.2: Relationship between oxygen saturation levels and heights of *R* peaks of ECG. The left vertical lines are from HF subjects and right lines are from healthy subjects. Color bar corresponds to the number of samples. For each oxygen saturation level, the ECG voltage with most number of samples is marked with red circle.

average mapping accuracy is 51.7% while for HF subjects the average is only 16.9%, which means that for healthy subjects the R peak heights are consistent regardless of oxygen saturation level, while for HF subjects the relationship is not clear. The clinical explanation of this result is not clear though.

The second regression is on the clusters of ECG and SaO_2, both in global space. For convenience we choose $K = 20$ for both data. The relationship is not clear, as in Fig. 4.3.

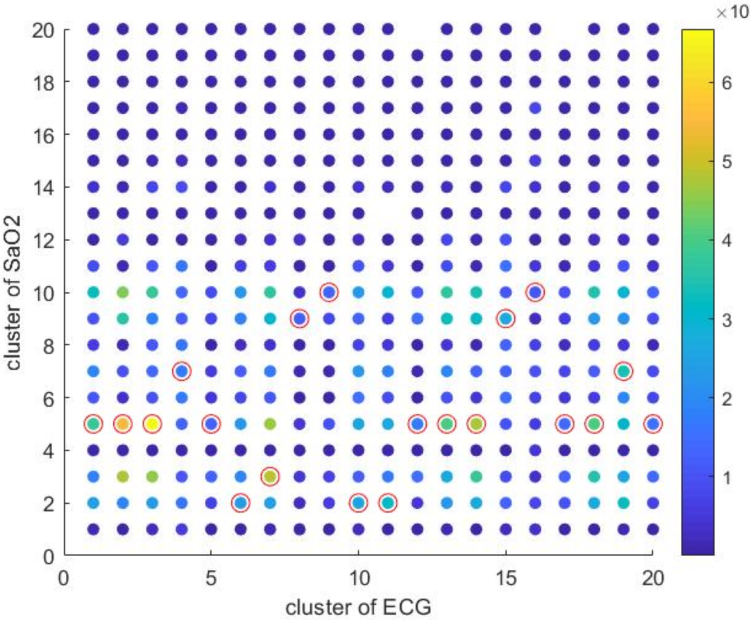


Figure 4.3: Regression analysis of clusters of ECG and SaO_2. The color bar shows the number of samples. For each cluster of ECG, the cluster of SaO_2 with the largest number of samples is marked with red circle.

Chapter 5

Discussion and Conclusion

The development of economic, reliable and non-invasive risk prediction methods of individuals for primary prevention ICD implantation is of clinical and public healthy priority. The present study allowed us to identify patients who would benefit the most from ICD implantation with non-invasive ECG and oxygen saturation test.

Current strategies for risk stratification based on deterministic linear measures have demonstrated limited clinical utility (Rashba et al., 2006; Berger et al., 1997; Grimm et al., 2003; Hohnloser et al., 2003). Large patient datasets and novel machine learning methods have facilitated the development and validation of new risk prediction models (Rajpurkar et al., 2017; Pourbabaee, Roshtkhari, and Khorasani, 2017; Acharya et al., 2017), which are able to perform more complicated tasks with higher accuracy. Our research combines machine learning algorithms, stochastic processes and nonparametric statistics into a cohesive measure which is robust and accurate in identification of patients with high heart failure risk. Our findings on the nonlinear dynamics are consistent with previous results on the same dataset (DeMazumder et al., 2016). The methods we introduced not only can be used to identify patients who would benefit from ICD implantation, but have the potential to be widely applied in other clinical problems.

A sensitivity, specificity and accuracy of 93.2% were reached. This could not be achieved with any single approach. The optimal feature set outperforms the accuracy of each single feature. Therefore, ECG and SaO_2 data could provide complementary information in the context of patient identification and thus enhance diagnostic ability.

References

Rashba, Eric J, NA Mark Estes, Paul Wang, Andi Schaechter, Adam Howard, Wojciech Zareba, Jean-Philippe Couderc, Juha Perkiomaki, Joseph Levine, Alan Kadish, et al. (2006). “Preserved heart rate variability identifies low-risk patients with nonischemic dilated cardiomyopathy: results from the DEFINITE trial”. In: *Heart rhythm* 3.3, pp. 281–286.

Berger, Ronald D, Edward K Kasper, Kenneth L Baughman, Eduardo Marban, Hugh Calkins, and Gordon F Tomaselli (1997). “Beat-to-beat QT interval variability: novel evidence for repolarization lability in ischemic and nonischemic dilated cardiomyopathy”. In: *Circulation* 96.5, pp. 1557–1565.

Grimm, Wolfram, Michael Christ, Jennifer Bach, Hans-Helge Müller, and Bernhard Maisch (2003). “Noninvasive arrhythmia risk stratification in idiopathic dilated cardiomyopathy: results of the Marburg Cardiomyopathy Study”. In: *Circulation* 108.23, pp. 2883–2891.

Hohnloser, Stefan H, Thomas Klingenheben, Daniel Bloomfield, Omar Dabbous, and Richard J Cohen (2003). “Usefulness of microvolt T-wave alternans for prediction of ventricular tachyarrhythmic events in patients with dilated cardiomyopathy: results from a prospective observational study”. In: *Journal of the American College of Cardiology* 41.12, pp. 2220–2224.

Rajpurkar, Pranav, Awni Y Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y Ng (2017). “Cardiologist-level arrhythmia detection with convolutional neural networks”. In: *arXiv preprint arXiv:1707.01836*.

Pourbabaee, Bahareh, Mehrsan Javan Roshtkhari, and Khashayar Khorasani (2017). “Deep convolutional neural networks and learning ECG features for screening paroxysmal atrial fibrillation patients”. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 99, pp. 1–10.

Acharya, U Rajendra, Hamido Fujita, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, and Muhammad Adam (2017). “Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals”. In: *Information Sciences* 415, pp. 190–198.

DeMazumder, Deeptankar, Worawan B Limpitikul, Miguel Dorante, Swati Dey, Bhasha Mukhopadhyay, Yiyi Zhang, J Randall Moorman, Alan Cheng, Ronald D Berger, Eliseo Guallar, et al. (2016). “Entropy of cardiac repolarization predicts ventricular arrhythmias and mortality in patients receiving an implantable cardioverter-defibrillator for primary prevention of sudden death”. In: *Ep Europace* 18.12, pp. 1818–1828.

Mengnan Zhao

Contact Information

116 W University Pkwy, Unit 1108
Baltimore, MD 21210

(607) 379-0676
mzhao21@jhu.edu

Research Interests

Mathematical Data Science, Machine Learning, Signal Processing.

Education

Johns Hopkins University, Baltimore, MD

M.S.E. candidate, Applied Mathematics and Statistics, expected May 2019

M.S.E (Ph.D. program), Electrical and Computer Engineering, September 2016 to December 2018

- Advisors: Sang (Peter) Chin, Ph.D and Mauro Maggioni, Ph.D
- Johns Hopkins University Fellowship

Cornell University, Ithaca, NY

M.S., Applied Physics, August 2013 to May 2015

- Thesis Topic: *Enhancement of the spin Hall effect in Platinum/Hafnium alloys*
- Advisor: Robert A. Buhrman, Ph.D

Sun Yat-sen University, Guangzhou, China

B.S., Optical Information Science and Technology, September 2009 to June 2013

- *Outstanding Undergraduate Thesis Award (Top 1%)*
- *Dean's List*

Research Experience

Research Assistant September 2017 to present

Johns Hopkins University, Department of Applied Mathematics and Statistics

- Conducted geometric machine learning of ECG time series data for primary prevention of sudden death
- Performed statistical analysis of hidden dynamics of ECG and SaO_2 data with Markov model

Research Assistant November 2016 to present

Johns Hopkins University, Department of Electrical and Computer Engineering

- Researched sparse recovery over graph incidence matrices
- Studied generalization of the nullspace property with applications in some special kinds of measurement matrices and regularizers

Research Assistant June 2014 to April 2016

Cornell University, School of Applied and Engineering Physics

- Investigated the giant spin Hall effect in thin film platinum/hafniumf alloys and spin transfer torque in magnetic heterostructures

Research Assistant September 2011 to August 2013

Sun Yat-sen University, State Key Laboratory of Optoelectronic Materials and Technologies

- Conducted theoretical and experimental studies on the diffraction of optical vortex beams
- Demonstrated a Pancharatnam-phase based interferometric method to recover phase pattern of a Gaussian optical vortex beam with different topological charges
- Simulated the polarization and phase pattern of diffracted optical vortex beams with numerical method based on angular spectrum representation

Teaching Experience

Teaching Assistant August to December 2014

Cornell University, School of Applied and Engineering Physics

- Graded homework and exams for AEP 4210 - Mathematical Physics I. Topics included: complex analysis, differential equations, etc.

Work Experience

Equity Research Analyst Externship January to April 2017
T. Rowe Price
Research Analyst Intern May to August 2016
CITIC Securities Company Limited

Publications

1. **M. Zhao**, J. M. Murphy, and M. Maggioni, “Classification of Physiological Signal Dynamics for Primary Prevention of Sudden Death”, *paper in preparation.*
2. **M. Zhao**, D. M. Kaba, R. Vidal, D. P. Robinson, and E. Mallada, “Sparse Recovery over Graph Incidence Matrices”, *57th IEEE Conference on Decision and Control (CDC)*, 2018.
3. D. M. Kaba, **M. Zhao**, R. Vidal, D. P. Robinson, and E. Mallada, “Abstract Simplicial Complexes and Sparse Recovery”, *paper in preparation.*
4. **M. Zhao**, L. Yang, X. Xie, H. Sun, and J. Zhou, “Pancharatnam-phase-based characterization for the diffraction of an optical vortex beam”, *Journal of Physics B: Atomic, Molecular and Optical Physics*, 47, 115401 (2014). [**J. Phys. B annual highlight of 2014**]
5. M. H. Nguyen, **M. Zhao**, D. C. Ralph, and R. A. Buhrman, “Enhanced spin Hall torque efficiency in $\mathrm{Pt}_{100\text{-x}}\mathrm{Al}_{\mathrm{x}}$ and $\mathrm{Pt}_{100\text{-x}}\mathrm{Hf}_{\mathrm{x}}$ alloys arising from the intrinsic spin Hall effect”, *Applied Physics Letters*, 108, 242407 (2016).
6. M. H. Nguyen, **M. Zhao**, D. C. Ralph, and R. A. Buhrman, “Enhanced spin Hall ratios by Al and Hf impurities in Pt thin films”, *APS March Meeting*, 2016.
7. **M. Zhao** and X. Xie, “Generation of arbitrary vector beams with a spatial light modulator”, *In Proceedings of International Conference on Electronic Materials and Packaging (EMAP)*, 2012. [**Oral Presentation**]

Service

Reviewer:
American Control Conference
IEEE Transactions on Signal Processing

Skills

MATLAB, Python, Java, LaTeX

Certifications

Python Data Structures
Introduction to Data Science in Python