A Machine Learning Approach to Assess the Separation of Seismocardiographic Signals by Respiration

2018

Brian Solar
University of Central Florida

Find similar works at: https://stars.library.ucf.edu/honorstheses

University of Central Florida Libraries http://library.ucf.edu

Part of the Biomedical Devices and Instrumentation Commons

Recommended Citation

https://stars.library.ucf.edu/honorstheses/310

This Open Access is brought to you for free and open access by the UCF Theses and Dissertations at STARS. It has been accepted for inclusion in Honors Undergraduate Theses by an authorized administrator of STARS. For more information, please contact lee.dotson@ucf.edu.
A MACHINE LEARNING APPROACH TO ASSESS THE SEPARATION OF SEISMOCARDIOGRAPHIC SIGNALS BY RESPIRATION

by

BRIAN SOLAR

A thesis submitted in partial fulfillment of the requirements for the Honors in Major Program in Mechanical Engineering in the College of Engineering and Computer Science and in the Burnett Honors College at the University of Central Florida Orlando, Florida

Spring Term, 2018

Thesis Chair: Hansen Mansy, PhD
ABSTRACT

The clinical usage of seismocardiography (SCG) is increasing as it is being shown to be an effective non-invasive measurement for heart monitoring. SCG measures the vibrational activity at the chest surface and applications include non-invasive assessment of myocardial contractility and systolic time intervals. Respiratory activity can also affect the SCG signal by changing the hemodynamic characteristics of cardiac activity and displacing the position of the heart. Other clinically significant information, such as systolic time intervals, can thus manifest themselves differently in an SCG signal during inspiration and expiration. Grouping SCG signals into their respective respiratory cycle can mitigate this issue. Prior research has focused on developing machine learning classification methods to classify SCG events as according to their respiration cycle. However, recent research at the Biomedical Acoustics Research Laboratory (BARL) at UCF suggests grouping SCG signals into high and low lung volume may be more effective. This research aimed at comparing the efficiency of grouping SCG signals according to their respiration and lung volume phase and also developing a method to automatically identify the respiration and lung volume phase of SCG events.
DEDICATION

To God, and to the many that suffer from cardiovascular disease.
ACKNOWLEDGMENTS

I thank Dr. Hansen Mansy, my thesis chair, for giving me the opportunity to conduct research in his lab, and supporting my research endeavors beyond what I could have asked for. I thank Dr. Dazhong Wu, my committee member, for inspiring me to pursue a more rigorous approach to machine learning. I thank Dr. Sandler for his important clinical insight. I thank Amir Taebi for giving me the foundation for this research and for the collaboration throughout. I thank the rest of the members of BARL for their suggestions, help, and overall support.
# TABLE OF CONTENTS

**LIST OF FIGURES** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . x

**LIST OF TABLES** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . xii

**CHAPTER 1: INTRODUCTION** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1

1.1 Overview . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1

1.2 Objective and Scope . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

1.3 Outline . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

**CHAPTER 2: Seismocardiography (SCG)** . . . . . . . . . . . . . . . . . . . . . . . . . 5

2.1 Definition . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

2.2 Review . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

2.3 Respiration and Lung Volume . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

2.4 Measurements . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

**CHAPTER 3: Machine Learning** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

3.2 Feature Extraction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2.1</td>
<td>Introduction</td>
<td>13</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Time-Domain Features</td>
<td>15</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Frequency-Domain Features</td>
<td>18</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Three-Dimensional Features</td>
<td>20</td>
</tr>
<tr>
<td>3.2.5</td>
<td>Feature Scaling</td>
<td>22</td>
</tr>
<tr>
<td>3.3</td>
<td>Classification</td>
<td>22</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Introduction</td>
<td>22</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Support Vector Machine</td>
<td>25</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Random Forest</td>
<td>29</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Neural Network</td>
<td>31</td>
</tr>
<tr>
<td>3.3.5</td>
<td>Evaluation</td>
<td>33</td>
</tr>
<tr>
<td>3.3.6</td>
<td>Model Selection</td>
<td>35</td>
</tr>
<tr>
<td>3.3.7</td>
<td>Training Scenarios</td>
<td>36</td>
</tr>
<tr>
<td>3.4</td>
<td>Clustering</td>
<td>36</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Introduction</td>
<td>36</td>
</tr>
<tr>
<td>3.4.2</td>
<td>K-Means</td>
<td>38</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Evaluation</td>
<td>39</td>
</tr>
</tbody>
</table>
CHAPTER 4: Classification Results ................................................. 41

4.1 Introduction ................................................................. 41

4.2 Feature Evaluation ....................................................... 41

  4.2.1 Introduction ....................................................... 41

  4.2.2 Time-Domain .................................................... 42

  4.2.3 Frequency-Domain .......................................... 43

  4.2.4 Time- with Frequency-Domains ......................... 43

  4.2.5 Three-Dimensional SCG .................................. 44

4.3 Classification Algorithm Evaluation ............................... 45

  4.3.1 Introduction .................................................... 45

  4.3.2 Support Vector Machine (SVM) ........................... 45

  4.3.3 Random Forest (RF) .......................................... 45

  4.3.4 Neural Network (NN) ........................................ 47

4.4 Classification Discussion ............................................... 49

  4.4.1 Feature Evaluation ........................................... 49

  4.4.2 Classification Algorithm Evaluation ...................... 50

  4.4.3 Classification Pipeline for SCG signals .................. 51

viii
CHAPTER 5: Clustering Results ........................................... 52

5.1 Introduction ......................................................... 52

5.2 K-Means Clustering Evaluation ................................. 52
  5.2.1 Introduction .................................................. 52
  5.2.2 Time-Domain ................................................ 52
  5.2.3 Frequency-Domain .......................................... 54
  5.2.4 Time- with Frequency-Domains ......................... 55

5.3 Clustering Discussion ............................................. 56

CHAPTER 6: CONCLUSION ............................................. 58

6.1 Summary ......................................................... 58

6.2 Conclusion ....................................................... 59

6.3 Recommendations for Future Work ............................ 60

LIST OF REFERENCES .............................................. 69
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>SCGz sample event</td>
<td>5</td>
</tr>
<tr>
<td>2.2</td>
<td>Cardiac Fiducial Points</td>
<td>7</td>
</tr>
<tr>
<td>2.3</td>
<td>SCG Events Overlayed on Respiratory Flow</td>
<td>10</td>
</tr>
<tr>
<td>2.4</td>
<td>SCG Events Overlayed on Tidal Volume</td>
<td>11</td>
</tr>
<tr>
<td>3.1</td>
<td>SCGz divided into 16 equal-width bins</td>
<td>15</td>
</tr>
<tr>
<td>3.2</td>
<td>SCGz divided into 16 adaptive-width bins</td>
<td>16</td>
</tr>
<tr>
<td>3.3</td>
<td>SCG Frequency Spectrum ($F(\omega)$)</td>
<td>18</td>
</tr>
<tr>
<td>3.4</td>
<td>16 Frequency Bands</td>
<td>19</td>
</tr>
<tr>
<td>3.5</td>
<td>Three-Dimensional SCG</td>
<td>21</td>
</tr>
<tr>
<td>5.1</td>
<td>K-Means Produced Clusters using Time-Domain Features According to Lung</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>Volume</td>
<td></td>
</tr>
<tr>
<td>5.2</td>
<td>K-Means Produced Clusters using Time-Domain Features According to Respiration</td>
<td>54</td>
</tr>
<tr>
<td>5.3</td>
<td>K-Means Produced Clusters using Frequency-Domain Features According to Lung Volume</td>
<td>54</td>
</tr>
</tbody>
</table>
5.4 K-Means Produced Clusters using Frequency-Domain Features According to Respiration .............................................. 55

5.5 K-Means Produced Clusters using Time- and Frequency-Domain Features According to Lung Volume .............................................. 56

5.6 K-Means Produced Clusters using Time- and Frequency-Domain Features According to Respiration .............................................. 57

6.1 Time-Domain Features ...................................................... 63

6.1 Frequency-Domain Features .............................................. 65

6.1 Time- and Frequency-Domain Features .............................................. 67
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Overview of Subject’s Characteristics (mean ± SD)</td>
<td>8</td>
</tr>
<tr>
<td>3.1</td>
<td>Contingency Table</td>
<td>40</td>
</tr>
<tr>
<td>4.1</td>
<td>Time-Domain Feature Evaluation for Classification</td>
<td>42</td>
</tr>
<tr>
<td>4.2</td>
<td>Frequency-Domain Feature Evaluation for Classification</td>
<td>43</td>
</tr>
<tr>
<td>4.3</td>
<td>Time-with Frequency-Domain Feature Evaluation for Classification</td>
<td>44</td>
</tr>
<tr>
<td>4.4</td>
<td>Three-Dimensional Feature Evaluation for Classification</td>
<td>44</td>
</tr>
<tr>
<td>4.5</td>
<td>SVM SS Classification Results</td>
<td>46</td>
</tr>
<tr>
<td>4.6</td>
<td>SVM LOSO Classification Results</td>
<td>46</td>
</tr>
<tr>
<td>4.7</td>
<td>RF SS Classification Results</td>
<td>47</td>
</tr>
<tr>
<td>4.8</td>
<td>RF LOSO Classification Results</td>
<td>47</td>
</tr>
<tr>
<td>4.9</td>
<td>NN SS Classification Results</td>
<td>48</td>
</tr>
<tr>
<td>4.10</td>
<td>NN LOSO Classification Results</td>
<td>48</td>
</tr>
<tr>
<td>5.1</td>
<td>K-means Clustering Results with Time-Domain Features</td>
<td>53</td>
</tr>
<tr>
<td>5.2</td>
<td>K-means Clustering Results with Frequency-Domain Features</td>
<td>55</td>
</tr>
<tr>
<td>5.3</td>
<td>K-means Clustering Results with Time- and Frequency-Domain Features</td>
<td>56</td>
</tr>
</tbody>
</table>
CHAPTER 1: INTRODUCTION

1.1 Overview

Our world is experiencing an unprecedented technological development coupled with an increasing application of engineering to medicine. Among the results of these phenomena are advancements in sensors, signal processing, machine learning, and computational power being harnessed for clinical utility [2], [21], [12]. Concurrently is the fact that cardiovascular disease is the leading cause of death in the United States for both men and women [5]. Given the medical significance of the heart, it is not surprising that these ongoing technological advancements are actively being explored for their potential use in cardiology. This thesis explores some of these potentialities.

Of the various potentialities are measurement techniques. Measurements are known as "the cornerstone of medical research and clinical practice" [41]. Sensors and other instrumentation transduce physiological signals into an electrical signal to facilitate measurements. Computers can then process the electrical signals into a manageable and interactive format and provide an enhancement in the qualitative measuring capacity of biomedical signals. Signal processing techniques and machine learning can be applied to extract clinically important information and characteristics from the signal that would otherwise not be observable by a human.

The cardiovascular system is one of the most important physiological systems for measurements and biomedical signal analysis. Cardiovascular disease is the leading cause of death in the United States for both men and women, and by 2035, is expected to cost healthcare $1.1 trillion [5]. There is therefore much incentive for the development of technologies for cardiovascular care and improvements in cardiac monitoring is an actively researched topic in biomedical signal
A particular measurement for cardiac health monitoring is Seismocardiography (SCG) [56]. SCG measures the vibrational activity at the chest surface and applications include non-invasive monitoring and detection of cardiovascular conditions [57]. However, compared to other cardiac measurements such as electrocardiography (ECG), SCG signals are relatively complex and not as well understood. Renewed interest in investigating the utility of SCG has accelerated in recent years and benefited from new advances in low-cost micro-electromechanical systems (MEMS) accelerometers, signal processing and machine learning methods.

SCG signals experience significant inter- and intra-subject variability due to activities such as respiration [30]. Relevant clinical cardiac information can be significantly different between the two respiratory phases (inspiration and expiration), and therefore prior studies have focused on separating SCG signals into their respective respiratory phase [3]. However, the physiological changes occurring during respiration that affect the SCG signal may largely be in part due to changes in lung volume (high and low lung volume) as opposed to respiratory flow. Taebi [44] investigated the advantages of using lung volume over respiratory phases and reported that grouping according to lung volume resulted in better grouping. This thesis compares the efficiency of grouping SCG signals according to their respiratory phase (inspiration vs. expiration) and lung volume using machine learning algorithms.

1.2 Objective and Scope

The main objective is to determine whether separating SCG signals according to respiration or lung volume is more efficient through the use of machine learning algorithms. As stated before, SCG signals suffer from significant variability, and separating them into their respective respiratory
or lung volume phases are methods to mitigate this issue that affects clinical analysis. Machine learning algorithms can not only aid in performing this separation but can also assess the efficiency of each method.

A secondary objective is to develop a machine learning classification pipeline to identify SCG signals as inspiration/expiration (INS/EXP) in their respiratory phase and high/low lung volume (HLV/LLV) in their lung volume phase. A number of feature extraction methods are explored and compared, such as time-domain and frequency-domain features. Three classification algorithms are compared: Support Vector Machine (SVM), Random Forest (RF), and Neural Network (NN).

As part of the main objective, a machine learning method, clustering, is used to assess the original assumptions that respiration and lung volume are valid domains for separation. The K-means clustering algorithm is used to cluster the SCG signals into two clusters. The degree to which the SCG signals are separated into INS/EXP (inspiration and expiration) vs. HLV/LLV (high lung volume and low lung volume) is examined.

1.3 Outline

The following is the structure for the remainder of the thesis:

- Chapter 2 - Seismocardiography (SCG): A detailed introduction to the SCG, including its clinical significance, a review of recent literature pertaining to its advances, and the measurements conducted for this thesis.

- Chapter 3 - Machine Learning (ML): An introduction to ML, including the intuition behind classification and clustering, and an explanation of the feature extraction and ML methods used to develop the ML classification pipeline.
• Chapter 4 - Classification Results: The usefulness of the various feature extraction methods for classification is explored. The results of using three classification algorithms (SVM, RF, and NN) to identify the respiration and lung volume phase are presented.

• Chapter 5 - Clustering Results: The K-means algorithm clustering results are presented while using different sets of features. The degree to which the K-means clustering algorithm separated the SCG signals into expiration/inspiration or high/low lung volume is assessed.

• Chapter 6 - Conclusions: A summary of the thesis is first presented followed by the main conclusions. Physiological explanations of the results are discussed, as well as possibilities for future investigations.
CHAPTER 2: Seismocardiography (SCG)

2.1 Definition

Seismocardiography (SCG) is a noninvasive technique that measures cardiac-induced mechanical vibrations at the chest surface using an accelerometer. SCG signals contain information related to mechanical activity of the cardiac system as opposed to electrical from electrocardiography (ECG) signals [15]. Generally, the z-component (normal to chest surface) of an accelerometer is used for SCG (termed SCGz), but some studies ([23], [33]) have incorporated the x and y-components (SCGx and SCGY) for the case when using a tri-axial accelerometer. For the purposes of this thesis, all three axes are considered, and a single SCG event is synonymous to a single heartbeat. An example of a SCGz event is shown in figure 2.1.

![SCGz sample event](image)

Figure 2.1: SCGz sample event

2.2 Review

SCG signals are believed to be caused by the cardiac mechanical processes (such as cardiac contraction, blood momentum changes, valve closure, etc.) [45], as mentioned before. The charac-
teristics of these signals can contain useful clinical information that correlates with cardiovascular processes and pathologies. Such information would complement methods of detecting heart electrical activity, such as ECG. The SCG therefore has the potential of aiding in the monitoring and detection of cardiovascular health.

A number of studies have applied SCG for patients with cardiovascular disease. These included the diagnosis and monitoring of clinical conditions such as: atrial fibrillation [32], atrial flutter [33], valvular disease [17], coronary artery disease and ischemia [40], myocardial infarction [49], heart failure [18], structural heart disease [53], and heart stress testing [10]. Information from the SCG signal that are believed to be generated by opening and closure of heart valves can be used as a diagnostic marker of these diseases.

An issue surrounding the SCG waveform is the variability. Significant inter and intra-subject variability exists. SCG signals morphology changes with different parameters such as respiratory cycle, different gender, age, sensor location on the chest, health conditions, cardiac contractility, heart rhythm, and postural positions [50]. Studying these confounding parameters can enhance our understanding of SCG, help group SCG cycles into groups with similar SCG events, reduce SCG signal variability and noise, and lead to more accurate definition of its features.

Important cardiac information contained in SCG signals include cardiac time intervals (CTIs). CTIs are a quantitative non-invasive measurement of cardiac function and can be used as markers for cardiac disease [24]. They are found by first locating the fiducial points, as shown in figure 2.2 [48], [46]. The fiducial points are indices corresponding to cardiac events in both systole and diastole. The fiducial points in a SCG signal are related to the opening and closure of the mitral and aortic valves. In systole, these include the closure of the mitral valve (MC) and the opening of the aortic valve (AO). In diastole, these include the closing of the aortic valve (AC) and the opening of the mitral valve (MC). CTIs are affected by respiration [50], and it is one factor that
has inspired researchers to separate SCG signals according to their respiratory phase [54], [3].

Prior studies have mainly focused on using SCGz, but as opposed to only the z-component, the x- and y- components of the SCG signal (SCGx and SC Gy) may also contain clinically important information. Paukkunen et al. [33] showed that 3-D vector trajectory of SCG might be useful in diagnosing atrial flutter, and another study used an averaged 3D SCG signal to assess heart failure patients. Because the accelerometer used in this study is tri-axial, SCGx and SC Gy will be also assessed.

2.3 Respiration and Lung Volume

Respiratory activity can affect the SCG signal by changing the hemodynamic characteristics of cardiac activity and displacing the position of the heart. Clinically significant information, such as CTIs, can thus manifest themselves differently in an SCG signal during inspiration and expiration. Grouping SCG signals into their respective respiratory cycle can mitigate this issue and
help quantify the differences on the SCG signal due to physiological parameters. Prior research therefore has focused on separating according to their respiration cycle [54]. However recent research at the Biomedical Acoustics Research Laboratory (BARL) at UCF suggests grouping SCG signals by high and low lung volume (HLV and LLV) could be superior in isolating these effects [44]. Physiologically, this suggests that intrathoracic pressure changes and heart displacement are significant parameters affecting the SCG morphology, as these are more associated with lung volume than respiration [44].

2.4 Measurements

A dataset of SCG events from an ongoing study at the BARL is used for this thesis. A summary of the experimental protocol follows:

(A) Participants:

Experimental protocol used in this study was approved by the institutional review board of the University of Central Florida, Orlando, FL. A total of 10 adults participated in our study. The subjects provided their informed consent and reported no history of cardiovascular disease verbally. The age, height, and weight of the subjects were obtained and are reported in Table 2.1.

(B) Experiments:
For consistency in breathing pattern and tidal volume, the subjects were first trained on how to breathe. For this purpose, a volume controlled ventilator (Model: 613, Harvard Apparatus, South Natick, MA) was used to instruct subjects to breath with the same inspiratory:expiratory (INS:EXP) ratio and respiratory rate. The INS:EXP and respiratory rate were set to 1:3 and 12 breath per minute, respectively. For each subject, the tidal volume (TV) was calculated in real time as the integral of the respiratory flow rate. The TV was displayed to the subjects on a computer screen during the experiment and was kept at 10 to 15 mL/kg. The subjects were asked to lay down on a folding bed with their chest tilted at 45 degrees and the signals of interest were recorded for 2 trials of 5 minutes each.

(C) Instrumentation:

A triaxial accelerometer (Model: 356A32, PCB Piezotronics, Depew, NY) was used to record all SCG signals. A signal conditioner (Model: 482C, PCB Piezotronics, Depew, NY) was used to amplify the accelerometer output with a gain factor of 100. The sensor was attached at the left sternal border and the 4th intercostal space using a double-sided medical-grade tape. This location was chosen due to its high signal-to-noise ratio [13], [14]. The accelerometer’s z-axis was aligned perpendicular to the subject chest surface, while the x- and y-axes were aligned parallel to the axial and mediolateral directions, respectively. The respiratory flow rate of the subjects was measured using a pre-calibrated spirometer (Model: A-FH-300, iWorx Systems, Inc., Dover, NH). The inspiration and expiration produced positive and negative flow rate signal amplitudes, respectively. The voltage signal for respiratory flow rate, ECG, and SCG signals were acquired simultaneously using a Control Module (Model: IX-TA-220, iWorx Systems, Inc., Dover, NH).

All the signals were acquired simultaneously at a sampling frequency of 10 kHz. The SCG signals were then filtered using a low-pass filter with a cut-off of 100 Hz to remove the remaining respiratory sound noise, which have significant energy above this cut-off fre-
quency [15]. Matlab (R2015b, The MathWorks, Inc., Natick, MA) was used to process all signals.

(D) SCG Signal Pre-Processing:

The SCG cycles in each signal were found using a matched filtering with a template consisting of a previously identified SCG. The respiratory and tidal volume signal was then used to group SCG events into two groups for two scenarios: 1) inspiration/expiration (INS/EXP) and 2) high and low lung volume (HLV and LLV). To assign whether the SCG signal occurred during INS or EXP and HLV or LLV, the respiration flow rate (figure 2.3) and the tidal volume (figure 2.4) at the peak of the SCG signals were used. Inspiration was defined as a respiration flow rate of greater than or equal to 0, and expiration as less than 0. HLV was defined as a tidal volume of greater than or equal to 0, and LLV as less than 0. The SCG events were mean padded to a length of 8192. This was to ensure that the lengths of all signals were identical and of a power of 2, for homogeneity and mathematical convenience. For the two separation scenarios (HLV/LLV and INS/EXP), HLV and INS were labeled as 1, and LLV and EXP as 0. A total of 5960 SCG events were recorded, with an average of $596 \pm 58$ events per person.
Figure 2.4: SCG Events Overlayed on Tidal Volume
CHAPTER 3: Machine Learning

3.1 Introduction

The signals obtained from the body are generally disturbed with noise and interference, and due to these factors and others, characteristics of the signal that provide clinically important information may not be observable by a human [37]. However, improvements in biomedical signal analysis have been made through the use of machine learning techniques [11]. Machine learning has the potential to objectively interpret the biomedical signal, and in some cases can perform better than a human with many years of experience [37].

Machine learning is generally defined as an “automate process that extracts patterns from data.” [20]. The goal of machine learning, after observing a set of examples, is to learn regularities from these examples through the use of computational methods in order to make predictions on a future set of data. Succinctly, the goal of these methods is to produce a prediction following a reception of inputs. Two large fields of machine learning are supervised and unsupervised learning [6].

Supervised learning is concerned with accurately modeling the mapping of inputs to outputs. After training with a set of inputs labeled with their actual outputs, a supervised learning algorithm can then produce predictions of the output for a new set of inputs. The term “supervised” is used because we know, at least for the training examples, what the true outputs should be. Two common supervised learning tasks are regression and classification [6]. Regression algorithms seek to map the input to a continuous set of outputs. Classification algorithms seek to map the input to a discrete set of outputs. This thesis will make use of classification algorithms and are explained in section 3.3.
Unsupervised learning [6] receives a set of inputs and is not concerned with mapping them to a specific output; instead it is concerned with making inferences about the structure of unlabeled data. Three common unsupervised learning tasks are density estimation, dimensionality reduction, and clustering. Density estimation is concerned with modeling the probability distribution of the unlabeled data. Dimensionality reduction is concerned with finding a lower dimensional space that may better represent the data. Clustering is concerned with separating the data into groups, or clusters, where the data within each cluster are more similar with each other compared to data within other clusters. This thesis will make use of clustering and is explained in section 3.4.

A term used in this thesis is “machine learning pipeline”, and it refers to the entire process of receiving raw data to the execution of a machine learning algorithm. It consists of multiple parts, which are all described in this thesis. The raw data first goes through an initial preprocessing, which may include the labeling of the data (see section 2.4.4). The data then undergoes what is called feature extraction (see section 3.2), where the input goes through a transformation into a new representation of inputs. Finally, the input is fed to the machine learning algorithm where a specific task is performed, and in this case is either classification or clustering (see sections 3.3 and 3.4, respectively).

3.2 Feature Extraction

3.2.1 Introduction

The inputs to machine learning algorithms are generally transformed to a new space through a process called feature extraction [6]. Feature extraction is a phase in the machine learning pipeline that attempts to find the most defining and informative characteristics about the raw input data. These characteristics are called features and are ultimately used as the input to the machine learning
algorithm. Features can be quantitative or qualitative and should also be fixed in number (and in most cases, less than the total samples to be used for training). The hope is that after feature extraction, the performance and/or computational efficiency of the machine learning process will be greater; a larger dimensional input leads to a higher computational cost.

Formally, feature extraction begins with an input, $X \in \mathbb{R}^N$, such as a SCG signal with $N$ data points ($N$ representing the length of the signal), and transforms it to $X \in \mathbb{R}^n$, where $n$ are the number of features extracted. If the input represents a multitude of $m$ samples of size $N$, $X \in \mathbb{R}^{m \times N}$, such as $m$ samples of SCG events each of size $N$, then the transformation would be $X \in \mathbb{R}^{m \times N} \rightarrow X \in \mathbb{R}^{m \times n}$.

To understand feature extraction, one can think of how to classify a common animal. Without knowing the animal’s name or being able to see it, what questions could be asked to help determine the animal? Some questions one may ask are: Does the animal live on land? Does it have a vertebrate? Etc. An effort should also be made to not ask redundant questions, such as: Does it exist on Earth? Does it use energy? Etc. Redundant questions in this case are those with answers that do not change from one animal to another and is therefore not useful for the purpose of distinguishing. Similarly, another example could be medical diagnosis; perhaps it would be helpful to determine what symptoms the patient is experiencing, and it would probably not be helpful in most circumstances to determine if the patient has hair. For these examples, the characteristic information obtained can be compared to feature extraction.

Feature extraction for biological signals such as an SCG event requires a similar approach. A review of literature provides insight as to what features can be extracted. Broadly speaking, feature extraction has generally been limited to the time-domain [55], and the frequency-domain [16] through the Fast Fourier Transform (FFT). The time-domain features included statistical features of the SCG signal, such as mean, median, and standard deviation, and features related to cardiac mechanics [38], such as Cardiac Time Intervals (CTIs). Frequency domain features included sta-
tistical features and frequency coefficients obtained through the FFT. This study seeks to find the most useful features to be used in the SCG classification pipeline. Sections 3.2.2 through 3.2.4 provides a description of these features.

3.2.2 Time-Domain Features

The SCG signal is originally represented in the time-domain, and not surprisingly a number of time-domain features can be extracted. An investigation of prior literature shows a number of time-domain features that can be extracted, including statistical features such as mean and standard deviation [22], and those related to cardiac mechanics such as CTIs [38]. This thesis will focus on evaluating statistical time-domain features.

A statistical time-domain feature extraction method, proposed by Zakeri [55] for SCG signals, is to divide each SCG cycle into a specific number of equal-width bins, and extract the arithmetic mean of each bin as a feature. The means of each bin are then concatenated and correspond to the feature vector. An SCG signal divided into 16 equal-width bins is shown in figure 3.1.

However, in this method there may be segments of the SCG signal that may not contain useful information. More specifically, that information may not have significant contribution to

![Figure 3.1: SCGz divided into 16 equal-width bins](image)
the ability of the machine learning method to correctly classify the signals. The dimensionality reduction was done without prioritizing useful parts of the SCG signal; all segments of the SCG signal are equally represented in the reduced feature space.

To aim at solving this potential issue, a preliminary study divided the signal into bins, however binning of the signal was performed discriminately, where areas of the signal corresponding to higher variation received a higher concentration of bins [47]. A visualization of this is shown in figure 3.2, where an SCG signal is binned with adaptive-width bins.

![Figure 3.2: SCGz divided into 16 adaptive-width bins](image)

Starting with a SCG signal, $SCG$, of length $N$, and a desired number of bins, $p$, let $Indices$ be a list of $n$ tuples containing indexes of $SCG$ corresponding to divisions, or bins:

$$Indices = [(a_0, a_1), (a_1 + 1, a_2), ..., (a_{n-1} + 1, a_n)],$$

(3.1)

where $a_0 = 1$, $a_n = N$, and $a_{i-1} < a_i$ for all $i \in [1, n]$. The adaptive-width binning algorithm begins with $indices = [(1, a_1)]$, signifying 1 bin ($a_1 = N$), and divides the bin or segment of the signal ($SCG[a_{i-1} + 1 : a_i]$) corresponding the highest standard deviation in a recursive fashion until reaching the desired number of $p$ bins. The division occurs by first finding the $kth$ tuple, which
corresponds to the bin with the largest standard deviation:

\[ k = \arg\max_i \text{Std}(SCG[a_{i-1} + 1 : a_i]) \]  
\[ (3.2) \]

The \( k \)th tuple is then divided into two tuples and replaced by them in \textit{Indices}:

\[ (a_{k-1} + 1, a_k) \rightarrow (a_{k-1} + 1, a^*), (a^* + 1, a_k) \]  
\[ (3.3) \]

where,

\[ a^* = \text{Floor}\left(\frac{a_k - a_{k-1} - 1}{2}\right) \]  
\[ (3.4) \]

Using adaptive-width bins as opposed to equal-width led to a higher classification accuracy for identifying lung volume phase [47]. It was shown that the use of adaptive-width bins resulted in similar accuracy when using 16 (90%) bins up to 1024 (92%). When using regular sized bins, the accuracy began at 65% at 16 bins and converged to a accuracy similar to when using adaptive-width bins only after using 1024 bins (91%). Using the adaptive width bins allows extraction of most of the performance gained from increasing the number of bins with only 16 bins. Because of this, considering that the number of features should not exceed the sample size and a lower number of features improves computational speed, the feature extraction method employing the adaptive-width bins will be used to obtain the statistical time-domain features.

The statistical time-domain features used in this thesis were obtained by dividing the signal into 32 bins using the adaptive-width algorithm then calculating statistical features from each bin. A total 96 time-domain features were obtained, corresponding to 32 each from the mean, median, and standard deviation of each bin. The efficacy of each statistical metric will be assessed (see section 4.2.2).
3.2.3 Frequency-Domain Features

The frequency-domain is often used for feature extraction of time-series signals to include biomedical signals [45]. Frequency domain features from prior studies applying machine learning to SCG signals included statistical features and amplitudes from the signal’s frequency spectrum [42], [55]. The frequency spectrum can be obtained through the Fast Fourier Transformation (FFT). The Fourier transformation, \( F(\omega) \), of a time-domain signal, \( f(t) \), can be defined as:

\[
F(\omega) = \int_{-\infty}^{\infty} f(t) e^{i\omega t} \quad (3.5)
\]

The FFT of the time-domain signal results in the frequency spectrum of the signal, \( F(\omega) \), from which can then be used to extract features. Of the types of frequency features that can be obtained, this study will use statistical features and the amplitudes of the frequency spectrum. A sample frequency spectrum for an SCG signal obtained through the FFT is shown in figure 3.3. The Python library SciPy [19] was used to obtain the FFT.

![Figure 3.3: SCG Frequency Spectrum (\( F(\omega) \))](image)

The frequency statistical features are obtained similarly to those from the time-statistical. Prior studies divided the frequency spectrum into frequency bands and calculated statistical fea-
tures from each band. Pandia [31] indicated a significant difference between the average power of
the frequency bands for SCG signals occurring during expiration and respiration. Another study
[16] calculated the median of various frequency bands for a machine learning application involving
SCG signals of heart failure patients. Therefore, this study will assess both the median and average
power.

To extract both the median and average power features, first the frequency spectrum is divided
into frequency bands of equal frequency resolution. Because there exists negligible frequency
content beyond 35 Hz, the first 64 data points corresponding to a frequency range of 0 to 100
Hz will be divided into 16 frequency bands (4 data points in each band) each with a frequency
resolution of approximately 6.25 Hz (0-6.25, 6.26-12.50, . . . , 93.76-100 Hz). A visualization of
the 16 frequency bands for a frequency spectrum of an SCG signal is shown in figure 3.4.

From these frequency bands, the median and average power were calculated. The average
power (equation 3.7) is obtained by squaring the magnitudes of the amplitudes to obtain the power
spectrum, and then calculating the average of each band.

\[
\text{Average Power} = \frac{\sum_{i=a}^{b} |F(\omega_i)|^2}{b-a}
\]  

(3.6)
Another type of frequency feature that can be obtained are the amplitudes of the frequency spectrum. A prior study classifying SCG signals into their respiratory cycle [55] used the amplitudes of the frequency spectrum as features. As mentioned before, because of the lack of frequency content beyond 35 Hz, the first 32 frequency coefficients, corresponding to 0-50 Hz, will be used as features.

A total of 64 frequency features were obtained, 16 each by calculating the median and average power of 16 frequency bands, and 32 from amplitudes of the frequency spectrum. The efficacy of each type of frequency feature will be assessed as well as the concatenation of all three types.

### 3.2.4 Three-Dimensional Features

The SCG signal was measured by a tri-axial accelerometer, resulting in three SCG signals in the Euclidean dimensions of x, y, and z (SCGx, SCGy, and SCGz) as shown in figure 3.5. Typically, only the z component of the SCG signal is analyzed. An investigation whether the additional components may provide valuable information is made in this thesis. Valuable information being defined as information that leads to an increase in performance for our machine learning algorithms. Two ways of utilizing the additional components are investigated here. One is to create a 3D vector from the three axial components of the SCG signal, from which the magnitude and orientation can be obtained. The other is to concatenate the feature vectors obtained from each axial component.

The 3D vector is obtained by finding the magnitude, $|SCG|$, and the orientations, $\alpha$ and $\beta$. 
These are defined as:

\[ |SCG| = \sqrt{SCG_x^2 + SCG_y^2 + SCG_z^2} \]  
(3.7)

\[ \alpha = \tan^{-1} \frac{SCG_x}{SCG_y} \]  
(3.8)

\[ \beta = \tan^{-1} \frac{SCG_z}{\sqrt{SCG_x^2 + SCG_y^2}} \]  
(3.9)

These three items, \(|SCG|\), \(\alpha\), and \(\beta\), preserve the same shape as the original SCG signal, and features can be extracted from them with an identical process to the original SCG signals. Therefore the same set of features were extracted from the six items: \(SCG_x, SCG_y, SCG_z, |SCG|, \alpha,\) and \(\beta\).

The efficacy of extracting features from the resulting three-dimensional signals was assessed, both individually and from concatenating. Concatenating was assessed by the concatenation of the feature vectors extracted from \(SCG_x, SCG_y, SCG_z\) as well as the concatenation of feature vectors
3.2.5 Feature Scaling

The values for the extracted features can cover vastly different ranges. This can lead to some features being overrepresented and others underrepresented. One solution to mitigate this issue is through a technique called feature scaling. Feature scaling scales all features so that they fall in a similar range. The relative differences between values for a given feature remain. One way to achieve feature scaling is through standardization [20], which is defined as:

\[
x_{\text{scaled}} = \frac{x - \mu}{\sigma},
\]

where \(x\) is the original feature, \(x_{\text{scaled}}\) is the scaled feature, \(\mu\) is the mean, and \(\sigma\) is the standard deviation. Before classification, all features are scaled in this manner.

3.3 Classification

3.3.1 Introduction

Classification is about predicting or identifying what particular class or group an input belongs to; it is mapping an input to a class out of a finite group of predefined classes. Classification can help the medical professional arrive at a clinical diagnosis [37]. A common approach is to receive a biological signal, execute a classification algorithm, and label the signal as belonging to a particular class – called classification. Using classification, we can also attempt to classify the SCG signal as occurring during HLV/LLV or INS/EXP. Popular machine learning classification algorithms include Support Vector Machines (SVM), Random Forests (RF), and Artificial Neural
Networks (NN). Before providing details about these classification algorithms, the intuition behind classification is explored.

Living organisms in large part have two important cognitive functions, the function of perceiving and judging [27]. We perceive the space of our existence, whether it is the internal space of our thoughts or the physical space that encapsulates us. From perceptions, interpretations and judgments are made: whether they are involuntary judgments like shivering from our response to cold, or voluntary judgments like turning on the light to read at night; the two are intrinsically linked. Human beings perceive through their senses, which transforms the perception into an electrical signal to be received by the brain. The brain interprets the signal and subsequently performs a judgment.

Signal processing has an analogous process: with the use of sensors a change in the environment can be measured and transmitted electrically to be received by a computer. The signal is interpreted by the computer and analyzed; generally, a judgment is made concerning the signal by a human user or as the result of an algorithm. The process of perception to judgment, or the mapping of an input to an output, can be completed in innumerable ways. One such mapping is classification.

As humans, we are continuously classifying: from a reception of inputs through our eyes and ears we are classifying objects as cars, computers, people, friend, foe, etc., almost seamlessly. We can, generally, correctly identify objects to a high degree of accuracy. It was not always like this, there was an initial stage of learning; the first time we become exposed to a new object, how do we classify a perception that is totally novel? Parents, or other guides, facilitate the learning process by identifying for us the correct classification of an object. After being guided as to what this particular perception/input maps to, we can begin to classify without the need of a guide.

Computers can also learn to correctly classify signals autonomously. Signals labeled with
their correct classification, such as HLV/LLV, can be given to a computer, which is analogous to guiding (also called supervised learning), and after sufficient learning (a sufficient amount of data being used for training), can begin to classify similar signals correctly. The mechanism of learning in the brain facilitates learning for humans. Computers also need a mechanism of learning, of which is provided through classification algorithms. In summary, Classification algorithms in machine learning potentially provide a means for computers to learn to correctly identify signals.

For this thesis, three machine learning classification algorithms were used for binary classification: Support Vector Machine (SVM), Random Forest (RF), and Feed-Forward Neural Network (which is referred here simply as Neural Network (NN)). These were used to classify the SCG events as INS or EXP and HLV or LLV (two different classification scenarios). The binary classification procedure can be generalized as the mapping of a feature vector with \(n\) features, \(x \in \mathbb{R}^n\), to a single output, \(y \in \{0,1\}\). The output is a prediction of what output the input maps to, either 1 or 0 for binary classification, representing classes. For a number of \(m\) feature vectors with \(n\) features, \(X \in \mathbb{R}^{m \times n}\) (such as from extracting \(n\) features from \(m\) number of SCG signals), and the labels for each sample , \(Y \in \mathbb{R}^m\) (such as 1 for HLV and 0 for LLV), our classification function can be represented as \(F(X, \Theta, \Gamma)\), where \(\Theta\) are the weighed parameters, which are optimized automatically through training, and \(\Gamma\) are the hyperparameters, which are tuned by the user.

If we take a subset of samples \(p\) as training data, \(X_{\text{train}} \in \mathbb{R}^{p \times n}\), and a subset as testing data , \(X_{\text{test}} \in \mathbb{R}^{q \times n}\) where \(q = m - p\), we can train and test the classification model respectively. The model is trained by minimizing a cost function, \(J(X, Y, \Theta)\), to obtain the optimized parameters, \(\Theta^*\); which compares the the output or prediction of the classification model, to the actual labels for the training inputs, \(Y_{\text{train}}\).

\[
\Theta^* = \arg\min_{\Theta} J(F, X_{\text{train}}, Y_{\text{train}}, \Theta)
\]  

(3.11)
The cost function varies depending on the classification model, but it is generally a metric of how well the model fits the data it is constrained to, \( Y_{train} \).

Note that an intermediate validation subset can be used for choosing a \( \Gamma \) that results in optimizing the classification accuracy on the cross-validation subset. However, in this study we used the K-fold cross-validation (section 3.3.6) method to make use of all the training samples for optimizing both \( \Theta \) and \( \Gamma \).

We can test the accuracy of the classification model by predicting the classification, \( Y_{pred} \), of the test input subset:

\[
Y_{pred} = F(X_{test}, \Theta^*, \Gamma)
\]  \( (3.12) \)

The classification accuracy can be measured by summing the total amount of correct predictions over the total amount of samples:

\[
Accuracy = \frac{1}{q} \sum_{i=1}^{q} (1 - (Y_{test,i} - Y_{pred,i})^2)
\]  \( (3.13) \)

### 3.3.2 Support Vector Machine

Support vector machines are commonly used for classification of biomedical signals [51], [9]. For the case of binary classification, the goal is to create a decision boundary separating two classes. Samples that lie on one side of the decision boundary are identified as belonging to one class, and those that lie on the other side as the other class. Formally, the goal is to create a canonical hyperplane dividing the feature space in order to provide maximum separation between samples belonging to different classes [28]. Depending on where unknown samples lie in relation to the decision boundary, they are classified as one of the two classes. Taking two classes, for this
example being positive and negative classes, we can define the following function:

\[ f(x) = w \cdot x + b \]  

(3.14)

With a weight vector \( w \) orthogonal to the decision boundary, input vector \( x_i \) for \( i = (1, 2, \ldots, m) \), and a constant \( b \). When the function is negative, then the sample belongs to the negative class; if the function is positive, the sample belongs to the positive class. The goal is to find \( w \) and \( b \) such that the decision boundary has the widest distance between the two classes. To do this we start by setting constraints:

\[ w \cdot x^+ + b \geq 1 \]  

(3.15)

\[ w \cdot x^- + b \leq -1 \]  

(3.16)

Where \( x^+ \) being positive samples and \( x^- \) being negative samples. We then define \( y \) to be 1 for positive samples and \(-1\) for negative samples. Therefore:

\[ y_i(w \cdot x_i + b) - 1 \geq 0 \]  

(3.17)

For all samples along the decision boundaries, for both positive and negative samples, the following constraint is made:

\[ y_i(w \cdot x_i + b) - 1 = 0 \]  

(3.18)

Looking at the decision boundary, having a vector pointing to a positive sample on the decision boundary, \( x^+ \), and a vector pointing to a negative sample on the decision boundary, \( x^- \), and the vector orthogonal to the decision boundary, \( w \), it can be seen that the width of the decision boundary is:

\[ width = (x^+ + x^-) \cdot \frac{w}{\|w\|} \]  

(3.19)
Substituting the results from equation 3.18 into 3.19:

$$width = \frac{2}{\|w\|} \quad (3.20)$$

The objective is to maximize the width, and this is done by minimizing $\frac{1}{2}\|w\|^2$. This can be done with the method of Lagrange multipliers using the constraint given by equation 3.18:

$$L = \frac{1}{2}\|w\|^2 - \sum_i \alpha_i [y_i(w \cdot x_i + b) - 1] \quad (3.21)$$

$$\frac{\partial L}{\partial w} = w - \sum_i \alpha_i y_i \cdot x_i = 0 \quad (3.22)$$

$$\frac{\partial L}{\partial b} = -\sum_i \alpha_i = 0 \quad (3.23)$$

Where $\alpha_i$ are Lagrangian multipliers. Substituting equations 3.22 and 3.23 into 3.21 and simplifying:

$$L = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (3.24)$$

This can then be maximized with respect to $\alpha_i$ through numerical optimization techniques (such as Sequential Minimal Optimization [29]). However, it will only provide a solution if a hyperplane can be made separating the two classes. If a separation cannot be made, then there will be no adequate solution.

To solve this problem of being unable to create a decision boundary adequately separating the classes, the feature space can be transformed into a higher dimension which allows separation of the classes. This is done by replacing the inner product, $x_i \cdot x_j$, with the inner product of transformation functions, $\Phi(x_i) \cdot \Phi(x_j)$. Finding the transformation functions could prove to be difficult, but the inner product of the transformation functions is what is needed, not the transformation functions themselves. The inner product of the transformation functions can be found using kernel
methods.

A kernel function here can be defined as the inner product of the transformation functions:

\[ K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j) \]  

(3.25)

Specifically, in this study, the Radial-Basis-Function (RBF) kernel is used and is defined as:

\[ K(x_i, x_j) = e^{-\gamma ||x_i - x_j||} \]  

(3.26)

Where \( \gamma \) is a chosen hyperparameter. Applying the kernel function to the inner product in equation 3.24:

\[ L = \sum_i^m \alpha_i - \frac{1}{2} \sum_i^m \sum_j^m \alpha_i \alpha_j y_i y_j e^{-\gamma ||x_i - x_j||} \]  

(3.27)

And equation 3.22 into 3.14 to get the new decision function:

\[ f(z) = \sum_i^m y_i \alpha_i K(x_i, z) + b \]  

(3.28)

\[ Y_{test,i} = \sum_i^q y_i \alpha_i K(X_{train,i}, X_{test,i}) + b \]  

(3.29)

Which is assigned the sample according to the sign of its value (negative or positive.) The optimized bias term, \( b^* \), can be readily found by solving equation 3.18 for points along the decision boundary:

\[ b^* = -\frac{1}{2} \left( \max_i w^* \cdot x_i + \min_i w^* \cdot x_i \right) \]  

(3.30)

While not formally introduced, a hyperparameter \( C \) exists in common implementations of the SVM. The hyperparameter \( C \) governs how much penalty is given for the error present during training. A larger \( C \) generally increases variance, while a lower \( C \) generally increases the bias.
The implementation of the SVM in this thesis is through the sci-kit learn library [34] in Python.

3.3.3 Random Forest

Random forests can be used for most types of learning tasks, including classification [8]. The classification at the output of a random forest is obtained from the average output of an ensemble of decision trees, \( \{T_b\}_1^B \), each which are of themselves classification models.

A decision tree is a graphical structure composed of nodes. The nodes are arranged in a hierarchical fashion where one node successively splits into different nodes until the terminal nodes are reached. For this study, binary trees are used, where nodes split into exactly two daughter nodes: left and right. A node \( j \) begins with a set of samples, \( S_j \), which represents a sample of training pairs \( (x, y) \), and splits into two daughter nodes, \( S^L_j \) and \( S^R_j \), such that:

\[
S_j = S^L_j \cup S^R_j, \quad S^L_j \cap S^R_j = \emptyset
\]  

(3.31)

Every node represents a boolean question about a specific feature from the input data point. If the value of this feature is above or below a certain threshold, the input traverses down to the left or right daughter node. This continues until the terminal nodes, also called the leaf nodes. The label that is most present at a terminal node represents the classification output at that node, and is represented as which class, \( y^* \), it has the highest probability of being:

\[
y^* = \arg \max_{y \in \{0, 1\}} p(y|x)
\]  

(3.32)

The splitting decision made at each node is automatically learned through a weak learner.
function:

\[ h(x, \Theta) \mapsto \{\text{left}, \text{right}\}, \tag{3.33} \]

where \( \Theta \) are the parameters for the weak learner. The objective function the weak learner attempts to maximize is the information gained through splitting the node in a particular way. The more distinct the separation is, the higher the information gain. The information gain, \( I \), is represented as:

\[ I = H(S) - \sum_{i \in \{L,R\}} \frac{|S'_i|}{|S|} H(S'_i), \tag{3.34} \]

where \( H \) is entropy, and is represented as:

\[ H(S) = - \sum_{y \in \{0,1\}} p(y) \log(p(y)). \tag{3.35} \]

The node can split the data by aligning axes (axis-aligned weak learner) through a subset of \( p \leq n \) features, and determining which axis splits the samples resulting in the highest information gain. Successive nodes are split in this fashion until certain stopping criteria are met, such as tree depth. The stopping criteria in this study was when the splitting resulted in only one type of class remaining in a daughter node.

A random forest is an ensemble of many decision trees that are generated through a random process. This results in higher generalization. The decision tree can be randomized by selecting a bootstrap sample \( N \) from \( m \) samples, and randomly selecting a subset of \( p \) features at every node to be analyzed. The number of trees in a random forest is itself a hyperparameter, with a higher number of trees generally resulting in a higher generalization at the expense of computational run-time.
The training algorithm can be summarized as:

1. From $b = 1$ to $B$:

   (a) Select a bootstrap sample of size $N$ from the training data.

   (b) Build a decision tree, $T_b$, for the bootstrapped data, by repeating these steps until the stopping criteria is met:

   i. Select $p$ variables from the available $n$.

   ii. Find the best feature/axis-aligned split among the $p$.

   iii. Split the node into two daughter nodes.

2. Output the ensemble of trees $\{T_b\}_{1}^{B}$

When classifying the test data, the results from all the individual trees are averaged, and the input data is classified as the prediction which has the highest value:

$$Y_{pred} = \arg \max_{y \in \{0,1\}} \frac{1}{B} \sum_{i=1}^{B} p_i(y, X_{test,i}) \quad (3.36)$$

The implementation of the RF in this thesis is through the sci-kit learn library [34] in Python.

3.3.4 Neural Network

Neural Networks (NN) are a class of machine learning models that are loosely modeled after the physiological neuron process [13]. For this study we used a feed-forward neural network for classification, which are the most commonly implemented, but will simply refer to it as a Neural Network (NN). A NN maps an input $x$ to a class, $y$, through a series of intermediate computations based on the architecture of the NN.
The architecture of a NN can be defined by the first layer, number of hidden layers, number of units per hidden layer, and the last layer. The first layer of \( m \) units is the input, \( X \in \mathbb{R}^{m \times n} \), and the last layer is the output, \( Y \in \mathbb{R}^m \). Every unit before the last layer has a linear model:

\[
z_{i+1} = w_i^T a_i + b_i, \tag{3.37}
\]

that maps the outputs from each unit, \( a_i \), to the units in the next layer. The first layers units can be represented as \( a_0 \). Every unit then transforms the result of this linear model through an activation function, \( f(z) \):

\[
a_{i+1} = f(z_{i+1}). \tag{3.38}
\]

Two activation functions of interest are the rectified linear unit (ReLU) and the sigmoid logistic function. ReLU is defined by:

\[
f(z) = \max\{0, z\} \tag{3.39}
\]

ReLU is recommended as the activation function for the units preceding the last layer. It replaces negative inputs with a value of 0 and keeps values that are non-negative. The sigmoid function is defined by:

\[
\sigma(z) = \frac{1}{1 + e^{-z}} \tag{3.40}
\]

Sigmoid functions are recommended as the activation function of the output units for binary classification, as it bounds the output in the interval of \([0, 1]\). As the goal of binary classification is to limit our model to two outputs, 0 and 1, typically if the output of the sigmoid is greater than or
equal to 0.5, it is assigned 1, and 0 otherwise:

\[
f(z) = \begin{cases} 
1, & \text{if } \sigma(z) \geq 0.5. \\
0, & \text{otherwise.} 
\end{cases}
\]  

(3.41)

Through this we can limit the output to 0 and 1 for binary classification.

The parameters for the linear models \(w\) and \(b\) are trained through the minimization of a cost function, \(J(w, b)\), which in this study is the binary cross-entropy []:

\[
J(w, b) = -\frac{1}{m} \sum_{i=1}^{m} [y_i \ln(a_{\text{output},i}) + (1 - y_i) \ln(1 - a_{\text{output},i})],
\]  

(3.42)

where \(y\) represents the true labels, and \(a_{\text{output}}\) represents the output of the NN.

The parameters are optimized through a process called gradient descent, where gradient of the cost function with respect to the parameters is used to adjust the parameters through an iterative process, with the goal of minimizing the cost function.


3.3.5 Evaluation

The performance of the classification algorithms is assessed here using two metrics, 1) classification accuracy, and 2) the F1 score. The classification accuracy was described in equation 3.16, but here it is described using different notation which will facilitate comparison with the F1 score. The F1 score is the harmonic mean of both precision and recall [35]. The F1 score is a metric that measures how well we are predicting both classes, as only using the classification accuracy may
not be sufficient if one class is more heavily represented than the other.

First we will say that classification labels of 1 (HLV or INS) are positive, and labels of 0 (LLV or EXP) are negative. Then we define four metrics, true positive (TP), true negative (TN), false positive (FP), and false negative (FN):

\[
TP = \text{Correctly predicted label of 1} \tag{3.43}
\]

\[
TN = \text{Correctly predicted label of 0} \tag{3.44}
\]

\[
FP = \text{Incorrectly predicted label of 1} \tag{3.45}
\]

\[
FN = \text{Incorrectly predicted label of 0} \tag{3.46}
\]

These four metrics represent the total correct and incorrect predictions. The accuracy can then simply be defined as:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.47}
\]

The F1 score is the harmonic mean of both precision and recall, which are defined as:

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{3.48}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{3.50}
\]

Precision represents the proportion of correctly predicted positive labels to the total amount of predicted positive labels. Recall represents the proportion of correctly predicted positive labels to
the total amount of actual positive labels. The F1 score is defined as:

\[ F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \]  \hspace{1cm} (3.51)

3.3.6 Model Selection

The goal of machine learning is to maximize the performance for a predictive task. Assessing the performance of the model is not as simple as evaluating the classification accuracy on the training set. The model may be subject to over fitting and perform poorly on a new set of data. To mitigate this issue, the data can be split into three sets, a training set, a validation set, and a test set [6]. The model is trained on the training set, the hyperparameters are chosen to obtain the highest accuracy on the validation set, and the model is assessed on the test set. The test set helps us evaluate our model for its ability to predict new data. The validation set helps us not choose the hyperparameters in such a way that the model is optimized for the test data, effectively removing its role as serving as a new set of data for testing. Dividing the data into three sets lessens the amount of data available for training. One answer to this problem is to use K-fold cross-validation.

K-fold cross-validation [20] divides the training set into k equal-sized folds, trains on k-1 folds, tests on the remaining fold, and repeats k times. The average classification accuracy across the k iterations is the performance measure. Using k-fold cross-validation allows us to partition the data into only a training set and test set; the hyperparameters are chosen as to maximize the average accuracy outputted from the k-fold cross-validation process. The value for k is generally 5 or 10 [14], and for this thesis, 10-fold cross-validation is used for model selection.
3.3.7 Training Scenarios

Two different evaluation scenarios are employed, subject-specific (SS) and leave-one-subject-out (LOSO) [55]. SS is a scenario that assesses the model’s performance when confined to training and testing on a specific subject. For this thesis, SS was performed by using 80% of a subject’s data for training/validation, and 20% for testing. LOSO assesses the model’s performance when tested with a new subject’s data. LOSO was performed by using 9/10 subjects’ data to perform training/validation, and the remaining subject for testing. The SS classification accuracy was assessed for all 10 subjects, as well as the LOSO classification accuracy for leaving each of the 10 subjects out of training.

3.4 Clustering

3.4.1 Introduction

Clustering is about separating data into a set of groups, or clusters, which share similar features [52]. Clustering creates groups of data such that data within one cluster are more similar to each other than to data within other clusters. Labels for the input data are not given, therefore the clusters are generated from only information found within the input data [6]. Clustering may help us find the natural structure of the data, helping us validate the decision for classifying the signals as according to their respiration or lung volume phase. If the signals are being separated into clusters resembling HLV/LLV or INS/EXP, we can objectively assess the validity for using one over the other. A common clustering algorithm is the K-means algorithm, which will be used in this thesis. Similarly with classification, an intuition of clustering can be given by looking at living organisms.
The objects of the world appear to have a structural correlation [39]. By observing the attributes of objects, such as animals, we can readily begin to see the similarities and differences among them. The fact that this particular group of objects, animals, has a name implies unity and diversity. Unity in that those objects called animals have unifying characteristics that gave rise to the name; and diversity in that those objects not called animals have different characteristics that disqualify them from being called animal. This observed structure of the world, specifically living organisms, has led to the formation of the science of Taxonomy.

Taxonomy is the science of “naming, describing, and classifying organisms” [1]; it is a hierarchical categorization of living organisms. The taxonomic categories to which organisms are classified are defined according to a set of shared characteristics. For example, animals belonging to the class of mammals share characteristics that are different than those belonging to the class of reptiles. The origins of these biological taxonomical categories have their root in clustering.

The beginning of the taxonomical process is to separate organisms into groups with similar characteristics [1]. This process is analogous to clustering. For example, in antiquity Aristotle divided animals into those with blood, and those without [25]; today, we similarly distinguish animals as vertebrates and invertebrates. Separation can continue for the remaining divisions to obtain sub-categories, developing names for each along the way. Conceptually, we can see that before classification, there was clustering; that there was a separation of organisms into clusters, and subsequently the naming of each cluster.

Computationally, we can cluster the SCG signals into two clusters and seek to understand what physiological phenomenon these clusters describe. With a given metric for similarity, the clustering algorithm can divide the signals into two groups that maximize this objective measurement for both groups. With knowledge of the physiological phenomenon occurring during the clustered SCG events, such as HLV/LLV or INS/EXP, the level to which the clusters represent
these physiological phenomena can be assessed.

3.4.2 K-Means

The K-means clustering algorithm [6] divides input data, \( X \in \mathbb{R}^{m \times n} \), into \( k \) discrete clusters, where the value of \( k \) is chosen by the user. Every cluster has a centroid and each data point belongs to the cluster whose centroid it is closest to. A measurement for distance from a data point in \( X \), \( x \in \mathbb{R}^n \), to a centroid, \( c \in \mathbb{R}^n \), can be calculated by the Euclidean distance, which is defined as:

\[
\text{Distance}(c, x) = \sqrt{\sum_{i=0}^{n} (c_i - x_i)^2}
\] (3.52)

Implementation of the algorithm is a simple iterative process, and the optimization occurs in a greedy fashion. The algorithm begins by randomly assigning locations for \( K \) centroids in the \( n \) dimensional space. It then assigns the data points, \( x \), from the input data, \( X \) (\( x \in X \)) to the closest centroid. The mean of the data points assigned to each cluster is calculated, and the locations of the centroids are reassigned to the mean location. This process repeats until the centroid locations remain unchanged. This is summarized as follows:

1. Select \( K \) centroid locations
2. Repeat until centroid location does not change:
   
   (a) Assign each data point to its closest centroid
   
   (b) Reassign location of each centroid to the mean of its cluster
3. Output the cluster labels for each data point
3.4.3 Evaluation

For the case of two clusters, we can define labels for one cluster to be 0, and the other as 1. We can further assess which cluster most represents either HLV and INS, or LLV and EXP. Each SCG event has an associated lung volume and respiration flow rate value, which was used to originally label them. These values allow the calculation of the mean lung volume and respiration rates for each cluster. The cluster which has a mean lung volume (or inspiration) greater than or equal to zero is assigned the label of 1, and 0 otherwise. With the clusters appropriately labeled, we can evaluate them. This thesis employs two evaluation metrics for clustering: 1) purity, and 2) adjusted rand index (ARI).

Purity evaluates how each cluster represents a specific class [26]. This is done by calculating the accuracy after the clusters have been assigned their appropriate labels as described above. It can then be defined as:

\[
Purity = \text{accuracy(Cluster Labels, Actual Labels)},
\]

Using equation 3.47 to calculate the accuracy. A purity value of 1 represents perfect clustering, while a value close to 0 represents bad clustering.

The adjusted rand index (ARI) is a measurement of how well the clusters agree with an external criteria (such as our known labels for HLV/LLV and INS/EXP). It is a common and recommended [26] metric used to evaluate clustering results. To explain the ARI, specifically for having two known classes and two clusters, the following contingency table is introduced.

The contingency table is shown in table 3.1.

The classes, \(U_1\) and \(U_2\), can be represented as HLV/LLV and INS/EXP. The clusters \(c_1\) and \(c_2\) can
Table 3.1: Contingency Table

<table>
<thead>
<tr>
<th>Class</th>
<th>Cluster</th>
<th>$c_1$</th>
<th>$c_2$</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_1$</td>
<td></td>
<td>$n_{1,1}$</td>
<td>$n_{1,2}$</td>
<td>$a_1$</td>
</tr>
<tr>
<td>$U_2$</td>
<td></td>
<td>$n_{2,1}$</td>
<td>$n_{2,2}$</td>
<td>$a_2$</td>
</tr>
<tr>
<td>sum</td>
<td></td>
<td>$b_1$</td>
<td>$b_2$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

be represented as the two clusters from K-means ($k = 2$). The values $n_{i,j}$ are the number of times the given class appears in the corresponding cluster. The sums across the columns and rows, $a_i$ and $b_j$, and total size of the input data, $N$, are also used to calculate the ARI from the following expression:

$$ARI = \frac{Index - Expected\ Index}{Maximum\ Index - Expected\ Index},$$

(3.54)

where

$$Index = \sum_{i,j} \left( \frac{n_{i,j}}{2} \right),$$

(3.55)

$$Expected\ Index = \frac{\sum_i \left( \frac{a_i}{2} \right) \sum_j \left( \frac{b_j}{2} \right)}{\binom{N}{2}},$$

(3.56)

$$Maximum\ Index = \frac{\sum_i \left( \frac{a_i}{2} \right) + \sum_j \left( \frac{b_j}{2} \right)}{2},$$

(3.57)

The ARI ranges from -1 to 1, where values close to 0 represent random clustering, hence little agreement with external criteria, and closer to 1 represents closer to perfect matching.
CHAPTER 4: Classification Results

4.1 Introduction

This chapter evaluates respiration and lung volume phase classification results, and also determines a suitable classification pipeline for SCG signals. This entails comparing the classification results for using different sets of features, such as time-domain and feature-domain, as well as the classification results when using the three different classification algorithms (SVM, RF, NN).

4.2 Feature Evaluation

4.2.1 Introduction

Feature evaluation was implemented by determining which features resulted in the highest classification accuracy. A SVM is used as the classification algorithm to evaluate the performance, as the results are universally consistent [43]. To generalize the results, data from all patients was considered, meaning the SVM is trained and tested on a dataset containing all the subjects’ SCG events. Of the three axial components of the SCG signal, only the z- component is used for feature evaluation of time- and frequency-domains. A random split of 50% of the data is used for training and 50% for testing (this is not to be confused with the 80/20 split used to evaluate SS classification results) during feature evaluation. The evaluation metric is the mean classification accuracy across 10 iterations. This process is summarized as follows:

1. For $i = 1$ to $10$

   (a) Random 50/50 split of data to create a training set ($X_{train}$ and $Y_{train}$) and test set ($X_{test}$ and $Y_{test}$)
(b) Train the SVM with the training set

(c) Output the predicted labels, $Y_{pred}$, of $X_{test}$

(d) Calculate the classification accuracy, $accuracy_i = accuracy(Y_{test}, Y_{pred})$

2. Output mean accuracy, $accuracy_{mean} = \frac{\sum_{i=1}^{10} accuracy_i}{10}$

The default SVM hyperparameter settings from the sci-kit learn SVM library are used ($C = 1.0$, $\gamma = \frac{1}{n}$ where $n = number of features$).

### 4.2.2 Time-Domain

The time-domain features (mean, median, and standard deviation) were each individually assessed, and the concatenation of all three was also assessed. Each had 32 features (extracted from the 32 bins of the SCG signal), therefore the concatenation of all three had a total of 96 features. The classification accuracies are presented in table 4.1. For both respiration and lung volume, using all three time-domain features resulted in the highest classification accuracy of 87%. Individually, for respiration using only the mean resulted in the highest (83%), and for lung volume only using the standard deviation resulted in the highest (88%).

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Feature</th>
<th>mean</th>
<th>median</th>
<th>standard deviation</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung Volume</td>
<td>mean</td>
<td>0.85</td>
<td>0.86</td>
<td>0.88</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>median</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>standard deviation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respiration</td>
<td>mean</td>
<td>0.83</td>
<td>0.81</td>
<td>0.79</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>median</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>standard deviation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>mean</td>
<td>0.84</td>
<td>0.83</td>
<td>0.83</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>median</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>standard deviation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.2.3 Frequency-Domain

The frequency-domain features (median, average power, and frequency amplitudes) were each individually assessed, and the concatenation of all three was also assessed. The median and average power features each had 16 features, and the frequency amplitude features were a total of 32. Together after concatenation there was a total of 64 features. The classification accuracies are presented in table 4.2. For both respiration and lung volume, the individual feature that resulted in the highest accuracy were the frequency amplitudes (88% and 92% respectively). Using all three feature types did not significantly change the performance over using only the frequency amplitudes.

Table 4.2: Frequency-Domain Feature Evaluation for Classification

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Feature</th>
<th>median</th>
<th>average power</th>
<th>frequency amplitudes</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung Volume</td>
<td>median</td>
<td>0.79</td>
<td>0.82</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>average power</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>frequency amplitudes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>All</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respiration</td>
<td>median</td>
<td>0.73</td>
<td>0.74</td>
<td>0.88</td>
<td>0.87</td>
</tr>
<tr>
<td>Average</td>
<td>median</td>
<td>0.76</td>
<td>0.78</td>
<td>0.90</td>
<td>0.90</td>
</tr>
</tbody>
</table>

4.2.4 Time- with Frequency-Domains

Because both the time- and frequency-domain features resulted in or equaled the highest performance when combining all, the combination of all time- and frequency-domain features will be assessed. Combining both the time and frequency features resulted in a total of 160 features. The classification results are shown in table 4.3. Using all features did not the affect the performance significantly compared to when using only the frequency features.
Table 4.3: Time- with Frequency-Domain Feature Evaluation for Classification

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Feature</th>
<th>Time and Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung Volume</td>
<td></td>
<td>0.93</td>
</tr>
<tr>
<td>Respiration</td>
<td></td>
<td>0.89</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>0.91</td>
</tr>
</tbody>
</table>

4.2.5 Three-Dimensional SCG

The results of using features extracted from both the time- and frequency-domain features for $SCG_x$, $SCG_y$, $SCG_z$, $|SCG|$ (magnitude of 3D signal), $\alpha$, and $\beta$ are evaluated, as well as the results of concatenating them. $SCG_{xyz}$ is concatenation of the three axial components, and $SCG_{3D}$ is the concatenation of $\alpha$, $\beta$, and $|SCG|$. The results are shown in Table 4.4. The highest accuracy occurred with $SCG_{xyz}$ at an accuracy of 96% for lung volume classification and 92% for respiration cycle classification. Of the individual SCG signals, $SCG_y$ and $SCG_z$ had the highest accuracy. The additionally derived components, such as the orientation and magnitude of the three-dimensional SCG signal, performed similarly to when using the original SCG signals.

Table 4.4: Three-Dimensional Feature Evaluation for Classification

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Feature 1</th>
<th>Feature 2</th>
<th>Feature 3</th>
<th>Feature 4</th>
<th>Feature 5</th>
<th>Feature 6</th>
<th>Feature 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung Volume</td>
<td>0.88</td>
<td>0.94</td>
<td>0.93</td>
<td>0.96</td>
<td>0.80</td>
<td>0.89</td>
<td>0.90</td>
</tr>
<tr>
<td>Respiration</td>
<td>0.86</td>
<td>0.88</td>
<td>0.89</td>
<td>0.92</td>
<td>0.81</td>
<td>0.85</td>
<td>0.86</td>
</tr>
<tr>
<td>Average</td>
<td>0.87</td>
<td>0.91</td>
<td>0.91</td>
<td>0.94</td>
<td>0.81</td>
<td>0.87</td>
<td>0.88</td>
</tr>
</tbody>
</table>
4.3 Classification Algorithm Evaluation

4.3.1 Introduction

The classification results for identifying the respiration and lung volume phase are presented here. A SVM, RF, and NN are used for this purpose in two training scenarios: 1) subject specific (SS) and 2) leave-one-subject-out (LOSO) (see section 3.3.7). The SS results are obtained for all 10 subjects, and the LOSO results are presented similarly for when that subject when they are left out of training. For example, a SS accuracy is reported for subject 1, and a LOSO accuracy is reported for leaving out subject 1 from training. Both time- and feature-domain features are used, and the concatenation of these features extracted from the three axial components ($SCG_{xyz}$) represents the feature vector used for classification.

4.3.2 Support Vector Machine (SVM)

The classification results for using a SVM during the SS and LOSO scenarios are listed in tables 4.5 and 4.6 respectively. The average SS classification accuracy for both lung volume and respiration across all 10 subjects was 94%, and the average F1 score was 94% and 93% respectively. The LOSO classification performance was worse, with an average classification accuracy/F1 score for lung volume and respiration of 61%/60% and 68%/54% respectively.

4.3.3 Random Forest (RF)

The classification results for using a RF during the SS and LOSO scenarios are listed in tables 4.7 and 4.8 respectively. The average SS classification accuracy/F1 score for lung volume and respiration was 95%/95%, and 90%/88% respectively. The LOSO classification performance was
Table 4.5: SVM SS Classification Results

<table>
<thead>
<tr>
<th>Subject</th>
<th>Lung Volume</th>
<th></th>
<th></th>
<th>Respiration</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>F1 Score</td>
<td>Accuracy</td>
<td>F1 Score</td>
<td>Accuracy</td>
<td>F1 Score</td>
<td>Accuracy</td>
<td>F1 Score</td>
</tr>
<tr>
<td>1</td>
<td>0.97</td>
<td>0.96</td>
<td>0.87</td>
<td>0.85</td>
<td>0.94</td>
<td>0.92</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>2</td>
<td>0.98</td>
<td>0.98</td>
<td>0.96</td>
<td>0.95</td>
<td>0.96</td>
<td>0.95</td>
<td>0.96</td>
<td>0.95</td>
</tr>
<tr>
<td>3</td>
<td>0.96</td>
<td>0.96</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>4</td>
<td>0.95</td>
<td>0.95</td>
<td>0.90</td>
<td>0.90</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>5</td>
<td>0.94</td>
<td>0.94</td>
<td>0.98</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>6</td>
<td>0.90</td>
<td>0.89</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
</tr>
<tr>
<td>7</td>
<td>0.93</td>
<td>0.93</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>8</td>
<td>0.93</td>
<td>0.93</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>9</td>
<td>0.96</td>
<td>0.96</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>10</td>
<td>0.95</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
</tr>
<tr>
<td>Average</td>
<td>0.95</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Table 4.6: SVM LOSO Classification Results

<table>
<thead>
<tr>
<th>Subject</th>
<th>Lung Volume</th>
<th></th>
<th></th>
<th>Respiration</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>F1 Score</td>
<td>Accuracy</td>
<td>F1 Score</td>
<td>Accuracy</td>
<td>F1 Score</td>
<td>Accuracy</td>
<td>F1 Score</td>
</tr>
<tr>
<td>1</td>
<td>0.73</td>
<td>0.68</td>
<td>0.83</td>
<td>0.82</td>
<td>0.94</td>
<td>0.92</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>2</td>
<td>0.60</td>
<td>0.53</td>
<td>0.75</td>
<td>0.67</td>
<td>0.93</td>
<td>0.93</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>3</td>
<td>0.75</td>
<td>0.75</td>
<td>0.56</td>
<td>0.66</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>4</td>
<td>0.61</td>
<td>0.55</td>
<td>0.66</td>
<td>0.62</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>5</td>
<td>0.42</td>
<td>0.59</td>
<td>0.61</td>
<td>0.61</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>6</td>
<td>0.59</td>
<td>0.71</td>
<td>0.74</td>
<td>0.47</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>7</td>
<td>0.60</td>
<td>0.39</td>
<td>0.59</td>
<td>0.43</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>8</td>
<td>0.74</td>
<td>0.70</td>
<td>0.79</td>
<td>0.81</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>9</td>
<td>0.51</td>
<td>0.67</td>
<td>0.60</td>
<td>0.07</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>10</td>
<td>0.60</td>
<td>0.45</td>
<td>0.67</td>
<td>0.23</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>Average</td>
<td>0.61</td>
<td>0.60</td>
<td>0.68</td>
<td>0.54</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
</tr>
</tbody>
</table>

also worse, with an average classification accuracy/F1 score for lung volume and respiration at 55%/51% and 60%/47% respectively.
### Table 4.7: RF SS Classification Results

<table>
<thead>
<tr>
<th>Subject</th>
<th>Lung Volume</th>
<th>Respiration</th>
<th>Lung Volume</th>
<th>Respiration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>F1 Score</td>
<td>Accuracy</td>
<td>F1 Score</td>
</tr>
<tr>
<td>1</td>
<td>0.97</td>
<td>0.96</td>
<td>0.87</td>
<td>0.84</td>
</tr>
<tr>
<td>2</td>
<td>0.98</td>
<td>0.98</td>
<td>0.96</td>
<td>0.95</td>
</tr>
<tr>
<td>3</td>
<td>0.95</td>
<td>0.95</td>
<td>0.94</td>
<td>0.93</td>
</tr>
<tr>
<td>4</td>
<td>0.98</td>
<td>0.98</td>
<td>0.86</td>
<td>0.82</td>
</tr>
<tr>
<td>5</td>
<td>0.96</td>
<td>0.95</td>
<td>0.86</td>
<td>0.85</td>
</tr>
<tr>
<td>6</td>
<td>0.96</td>
<td>0.96</td>
<td>0.93</td>
<td>0.89</td>
</tr>
<tr>
<td>7</td>
<td>0.90</td>
<td>0.89</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>8</td>
<td>0.96</td>
<td>0.96</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>9</td>
<td>0.89</td>
<td>0.89</td>
<td>0.89</td>
<td>0.85</td>
</tr>
<tr>
<td>10</td>
<td>0.95</td>
<td>0.95</td>
<td>0.92</td>
<td>0.88</td>
</tr>
<tr>
<td>Average</td>
<td>0.95</td>
<td>0.95</td>
<td>0.90</td>
<td>0.88</td>
</tr>
</tbody>
</table>

### Table 4.8: RF LOSO Classification Results

<table>
<thead>
<tr>
<th>Subject</th>
<th>Lung Volume</th>
<th>Respiration</th>
<th>Lung Volume</th>
<th>Respiration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>F1 Score</td>
<td>Accuracy</td>
<td>F1 Score</td>
</tr>
<tr>
<td>1</td>
<td>0.59</td>
<td>0.58</td>
<td>0.72</td>
<td>0.66</td>
</tr>
<tr>
<td>2</td>
<td>0.46</td>
<td>0.55</td>
<td>0.66</td>
<td>0.54</td>
</tr>
<tr>
<td>3</td>
<td>0.60</td>
<td>0.65</td>
<td>0.58</td>
<td>0.27</td>
</tr>
<tr>
<td>4</td>
<td>0.56</td>
<td>0.62</td>
<td>0.59</td>
<td>0.42</td>
</tr>
<tr>
<td>5</td>
<td>0.49</td>
<td>0.56</td>
<td>0.39</td>
<td>0.40</td>
</tr>
<tr>
<td>6</td>
<td>0.73</td>
<td>0.73</td>
<td>0.65</td>
<td>0.52</td>
</tr>
<tr>
<td>7</td>
<td>0.53</td>
<td>0.07</td>
<td>0.58</td>
<td>0.30</td>
</tr>
<tr>
<td>8</td>
<td>0.55</td>
<td>0.60</td>
<td>0.60</td>
<td>0.63</td>
</tr>
<tr>
<td>9</td>
<td>0.49</td>
<td>0.65</td>
<td>0.54</td>
<td>0.51</td>
</tr>
<tr>
<td>10</td>
<td>0.53</td>
<td>0.11</td>
<td>0.67</td>
<td>0.42</td>
</tr>
<tr>
<td>Average</td>
<td>0.55</td>
<td>0.51</td>
<td>0.60</td>
<td>0.47</td>
</tr>
</tbody>
</table>

### 4.3.4 Neural Network (NN)

The classification results for using a NN during the SS and LOSO scenarios are listed in tables 4.9 and 4.10 respectively. The average SS classification accuracy/F1 score for lung volume
and respiration was 95%/95%, and 94%/93% respectively. Similarly with the SVM and RF, the LOSO classification performance was worse, with an average classification accuracy/F1 score for lung volume and respiration at 55%/51% and 60%/47% respectively.

Table 4.9: NN SS Classification Results

| Subject | Lung Volume | | | Respiration | | | Accuracy | F1 Score | Accuracy | F1 Score |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 0.98 | 0.97 | | 0.90 | 0.89 | | 2 | 0.97 | 0.96 | | 0.96 | 0.95 | | 3 | 0.96 | 0.96 | | 0.98 | 0.98 | | 4 | 0.97 | 0.97 | | 0.93 | 0.92 | | 5 | 0.96 | 0.95 | | 0.90 | 0.90 | | 6 | 0.95 | 0.95 | | 0.98 | 0.97 | | 7 | 0.90 | 0.89 | | 0.95 | 0.95 | | 8 | 0.95 | 0.94 | | 0.89 | 0.89 | | 9 | 0.93 | 0.93 | | 0.95 | 0.93 | | 10 | 0.95 | 0.95 | | 0.95 | 0.92 | | Average | 0.95 | 0.95 | | 0.94 | 0.93 | | 48 | | | Table 4.10: NN LOSO Classification Results

| Subject | Lung Volume | | | Respiration | | | Accuracy | F1 Score | Accuracy | F1 Score |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 0.71 | 0.65 | | 0.80 | 0.74 | | 2 | 0.59 | 0.44 | | 0.68 | 0.59 | | 3 | 0.73 | 0.71 | | 0.48 | 0.59 | | 4 | 0.63 | 0.54 | | 0.68 | 0.48 | | 5 | 0.42 | 0.57 | | 0.60 | 0.65 | | 6 | 0.72 | 0.77 | | 0.74 | 0.63 | | 7 | 0.54 | 0.09 | | 0.55 | 0.38 | | 8 | 0.63 | 0.56 | | 0.68 | 0.74 | | 9 | 0.50 | 0.67 | | 0.63 | 0.35 | | 10 | 0.58 | 0.35 | | 0.77 | 0.57 | | Average | 0.60 | 0.53 | | 0.66 | 0.57 | | 48
4.4 Classification Discussion

4.4.1 Feature Evaluation

Determining which features to use for feature extraction in a classification pipeline is one necessary step. Comparing the two main types of features used, time-domain and frequency-domain features, using the frequency domain features, specifically the amplitudes of the frequency spectrum, led to the highest classification accuracy. Combining the time-domain features with the frequency-domain features did not significantly affect the performance. Because one consideration is computational efficiency, and a reduction in the number of features results in faster executions, using only the amplitudes of the frequency spectrum from the FFT of an SCG signal may be a reasonable decision. These results are contrasted with an earlier study, that mentioned using only time-domain features, as opposed to only frequency-domain and the combination of both, resulted in the highest performance for classifying the respiration cycle of SCG signals [54].

Additionally, incorporating the x and y axial components of the SCG signal may also be worthwhile. From the three axial components, the orientation and magnitude of the three-dimensional vector can be obtained, and identical features could be extracted from them. However, using them as features for classification did not result in an increase in classification performance. What did lead to an increase in performance was concatenating the features extracted from the three axial components. The average improvement in classification accuracy was 3%, which may justify this approach at the expense of execution time, as the number of features increases by a factor of 3.
4.4.2 Classification Algorithm Evaluation

The results of this study indicated that there was not a significant difference in classification performance between using lung volume or respiration cycles as the classification labels. The SS classification performance was high for all three classification algorithms for both lung volume and respiration, with the highest classification accuracy/F1 score for lung volume and respiration being 95%/95% (RF and NN) and 94%/93% (SVM and NN) respectively, and the lowest being 95%/94% (SVM) and 90%/88% (RF) respectively. This indicates training a classification algorithm to identify the respiration or lung volume phases of SCG signals can successfully be performed, given that it is unique to the individual, as is pointed out by the results from LOSO.

The LOSO results also indicate that there is not a significant difference in performance when using either lung volume or respiration cycles as the classification labels. Although the average respiration classification accuracy (65%) was 6% greater than the average for lung volume (59%) for LOSO across the three classification algorithms, the average F1 score for respiration was 53% or 2% less than for lung volume (55%). The low performance of LOSO is contrasted with a prior study [54], which achieved an average LOSO classification accuracy of 88.1% when identifying the of respiration cycle of SCG events. If similar conditions could be met, then the ability to identify the lung volume phase of SCG events could be more properly assessed. One explanation for low LOSO performance may be that the inter-subject variability is significant. The changes in the SCG signal due to lung volume and respiration may manifest themselves differently for different people. This may also be the result of over-fitting. The classification algorithms not properly generalizing the effects of respiration or lung volume and therefore not able to effectively predict signals from a new subject. A solution to over fitting can be to collect data from a large number of subjects, specifically focusing on the diversity of subjects (diversity of attributes among the subjects that may affect the SCG signal).
4.4.3 Classification Pipeline for SCG signals

One of the objectives for this thesis was to find an appropriate classification pipeline for SCG signals. This includes a procedure for feature extraction and classification. As mentioned earlier, the amplitudes of the frequency spectrum resulted in the highest classification accuracy among the assessed types, and the concatenation of these features from the x, y, and z components of the SCG signal thereafter improved the accuracy by 3%. If a lack of computational resources indicate a necessity of reducing the number of features, then the additional axial components could be ignored, if not, then the inclusion of them could increase the performance. Of the classification algorithms, there is not clear winner among the three. The SVM and NN were marginally better than the RF for LOSO. They could be used interchangeably, however if a relatively larger number of SCG events are collected, then the NN may outperform both the SVM and RF [13]. Another consideration is computational resources. The RF executed faster than the SVM, and the NN has the longest execution time.
CHAPTER 5: Clustering Results

5.1 Introduction

This chapter evaluates respiration and lung volume phase clustering results for using the following sets of features: time-domain, feature-domain, and both. The K-means clustering algorithm is used, and clustering is executed on each subject individually. The set of features which resulted in the highest clustering agreement with the given labels of INS/EXP and HLV/LLV is presented. The level to which respiration or lung volume is demonstrated in clustering is also presented.

5.2 K-Means Clustering Evaluation

5.2.1 Introduction

The K-means clustering algorithm is used to cluster the data for each subject into two clusters (k = 2). The clusters are assigned to the classes for which they are most populated by. A sample plot from one subject’s SCG events on their lung volume and respiration signals are displayed to demonstrate how the clusters were formed with respect to INS/EXP and HLV/LLV (the plots for all subjects can be found in appendix B).

5.2.2 Time-Domain

The clustering results for using time-domain features are shown in table 5.1. The Average purity/ARI for both lung volume and respiration were 0.55/0.05. These results indicate that the clusters are not being formed with respect to INS/EXP or HLV/LLV. This is visually conclusive by
figures 5.1 and 5.2, which shows the SCG events (represented as points on the signal) on a sample subject’s lung volume and respiration signal.

Table 5.1: K-means Clustering Results with Time-Domain Features

<table>
<thead>
<tr>
<th>Subject</th>
<th>Metric</th>
<th>Lung Volume</th>
<th>Respiration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Purity</td>
<td>ARI</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>0.50</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.82</td>
<td>0.42</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0.52</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.57</td>
<td>0.04</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>0.49</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>0.50</td>
<td>0.00</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>0.52</td>
<td>0.00</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>0.54</td>
<td>0.00</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>0.49</td>
<td>0.00</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>0.50</td>
<td>0.00</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>0.55</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Figure 5.1: K-Means Produced Clusters using Time-Domain Features According to Lung Volume
The clustering results for using frequency-domain features are shown in table 5.2. The average purity/ARI for lung volume was 0.81/0.45 and 0.64/0.31 for respiration. These results indicate that the clusters are sufficiently being formed with respect to HLV/LLV and to a lesser degree INS/EXP. This is visually conclusive by figures 5.3 and 5.4, which similarly shows the SCG events (represented as points on the signal) on a sample subject’s lung volume and respiration signal.
Table 5.2: K-means Clustering Results with Frequency-Domain Features

<table>
<thead>
<tr>
<th>Subject</th>
<th>Metric</th>
<th>Lung Volume</th>
<th>Respiration</th>
<th>Lung Volume</th>
<th>Respiration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Purity</td>
<td>ARI</td>
<td>Purity</td>
<td>ARI</td>
<td>Purity</td>
</tr>
<tr>
<td>1</td>
<td>0.92</td>
<td>0.70</td>
<td>0.67</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.90</td>
<td>0.63</td>
<td>0.74</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.87</td>
<td>0.54</td>
<td>0.71</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.86</td>
<td>0.53</td>
<td>0.65</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.64</td>
<td>0.25</td>
<td>0.57</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.82</td>
<td>0.40</td>
<td>0.67</td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.60</td>
<td>0.09</td>
<td>0.65</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.89</td>
<td>0.61</td>
<td>0.55</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.85</td>
<td>0.49</td>
<td>0.58</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.77</td>
<td>0.30</td>
<td>0.70</td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>0.81</td>
<td>0.45</td>
<td>0.65</td>
<td>0.31</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.4: K-Means Produced Clusters using Frequency-Domain Features According to Respiration

5.2.4 Time- with Frequency-Domains

The clustering results for using both time and frequency-domain features are shown in table 5.3. The Average purity/ARI for lung volume was 0.62/0.15 and 0.57/0.10 for respiration. These results indicate that the clusters are not sufficiently being formed with respect to HLV/LLV or INS/EXP. This is visually conclusive by figures 5.5 and 5.6, which shows the SCG events (repre-
presented as points on the signal) on a sample subject’s lung volume and respiration signal.

Table 5.3: K-means Clustering Results with Time- and Frequency-Domain Features

<table>
<thead>
<tr>
<th>Subject</th>
<th>Metric</th>
<th>Lung Volume</th>
<th>Respiration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Purity</td>
<td>ARI</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>0.50</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.87</td>
<td>0.56</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0.60</td>
<td>0.05</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.73</td>
<td>0.29</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>0.49</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>0.83</td>
<td>0.43</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>0.51</td>
<td>0.00</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>0.57</td>
<td>0.03</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>0.63</td>
<td>0.13</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>0.50</td>
<td>0.00</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>0.62</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Figure 5.5: K-Means Produced Clusters using Time- and Frequency-Domain Features According to Lung Volume

5.3 Clustering Discussion

Clustering, as with classification, can benefit from feature extraction. The K-means clustering results clearly indicate that the frequency features resulted in the highest performance. The
clusters generated using the frequency features agreed with the HLV/LLV labels and agreed to a lesser extent to the INS/EXP labels. The clusters generated when the time-domain features were included did not apparently reveal anything physiologically useful, and clearly did not demonstrate clustering according to lung volume or respiration phases.

The clustering results when using frequency features suggests that separating SCG signals according to their lung volume phase may be superior to separating according to respiration phase. The average purity/ARI values of 0.81/0.45 when comparing the clusters to the lung volume labels was approximately 25%/45% greater than when comparing the clusters to the respiration labels (0.65/0.31). This supports the conclusion made from a prior study at BARL at UCF [44] that grouping SCG signals according to lung volume phase was superior than according to respiration cycle. One possible physiological may be that the intrathoracic pressure changes and heart displacement are significant parameters affecting the SCG morphology, as these are more associated with lung volume than respiration [44].
CHAPTER 6: CONCLUSION

6.1 Summary

SCG events are subject to high intra- and inter-subject variability in part due to physiological parameters such as respiratory activity. Respiration has been shown to cause morphological changes in the SCG signal that affect derivable clinical information such as cardiac time intervals (CTIs). Prior literature has grouped SCG events according to their respiration cycle (inspiration and inspiration) as a way to mitigate variability [55], but a study conducted by BARL has shown that grouping according to high and low lung volume (HLV and LLV) may be superior [44].

This thesis first developed a classification pipeline for SCG signals and then used three classification algorithms (SVM, RF, and NN) to identify the respiratory and lung volume phases SCG events occurred. During the development of the classification pipeline, time-domain and feature-domain features were assessed, with the frequency features (specifically the amplitudes of the frequency domain) resulting in the greatest performance. Two training scenarios were used, subject specific (SS) and leave-one-subject-out (LOSO). The classification accuracy with the SS training scenarios for predicting respiration/lung volume phases when using a SVM, RF, and NN were 95%/94%, 95%/90%, and 95%/94% respectively. The average LOSO classification accuracies were 61%/68%, 55%/60%, and 60%/66% respectively.

For the second part of the thesis, SCG events were clustered using the K-means clustering algorithm, and the clusters were compared to see if they more resembled separation according to lung volume or respiration. Two metrics, purity and adjusted rand index (ARI), were used to assess the comparison. The clustering was performed while using only time-domain features, only frequency-domain features, and both time- and frequency-domain features. The resulting
purity/ARI for using these three sets of features when comparing to respiration was 0.55/0.05, 0.65/0.31, and 0.57/0.10 respectively. When comparing to lung volume they were 0.55/0.05, 0.81/0.45, and 0.62/0.15, respectively.

6.2 Conclusion

A classification pipeline for SCG events was developed for this thesis, along with the assessment of various features. The results of this study suggest that the incorporation of frequency features, specifically the amplitudes of the frequency spectrum, results in higher performance for both classification and clustering. It also suggested that extracting features from additional axial components of the SCG signal resulted in higher classification accuracies. One concern is that the number of features with this method is tripled, potentially leading to undesirous computational run-times. Of the three classification algorithms (SVM, RF, NN), neither of the three were clearly superior, but the NN should be assessed with a larger number of samples before this conclusion is made.

The results of classification did not suggest a clear advantage to using either respiration or lung volume as a separating criterion over the other, but the results of clustering does suggest separating by lung volume is superior. While accurately separating SCG events into their respective groups, whether they be respiration or lung volume phases, is important, it is not as important as the separating criteria itself. The separating criteria should be that the SCG events are separated into groups where similarity is high within a group and dissimilarity is high when comparing one group to another; a minimization of variability due to respiration is sought within a group. The K-means clustering algorithm provided a means to objectively assess the separating criteria by comparing them with the generated clusters. The clusters more resembled separation by lung volume, suggesting this physiological phenomena causes more pronounced changes in the SCG
signal than respiration.

One concern is that the LOSO classification training scenario resulted in poor performance for both separating criteria. Ideally, a classification model could be trained to accurately identify the lung volume or respiration phase of SCG events from a new subject. This would eliminate the need to obtain respiration information for a new subject’s SCG events we wish to separate. However, this would require performing well on LOSO. If only high SS performance could be obtained, then every person would need their own classification model, therefore requiring their respiration information. However, a prior study classifying the respiration phases of SCG events managed to obtain a classification accuracy of 88.1% with a LOSO training scenario [54]. If these results can be reproduced, then it is reasonable to suggest that a classification model can be trained to accurately identify the respiration or lung volume phases of SCG events for new subjects.

6.3 Recommendations for Future Work

Future work could include further investigation of separating criteria for SCG signals. One suggestion is through the use of graph similarity methods. One study measured the similarity of SCG events from inpatient and outpatient heart failure subjects with K-nearest-neighbors graphs [16]. If the similarity between INS and EXP SCG events is higher than between HLV and LLV, then this would be another objective measurement in favor of the separation of SCG signals by lung volume phase.

Feature selection in this thesis was completed heuristically, and future studies could incorporate a more systematic feature selection process. Possible methods could be through the use of Principal Component Analysis (PCA) [6] or recursive feature elimination [55].

The effects of lung volume could also be further investigated. Isolation of the effects of lung
volume can be studied by recording SCG events during breath-holding (end expiration breath-hold for LLV, and end inspiration breath-hold for HLV). By training on SCG events during breath-holding, we can see if a classification algorithm can accurately predict the lung volume phases of SCG events occurring during normal respiratory behavior.

Separation of SCG signals into their respiratory phases is a preliminary step, and ultimately future studies should focus on the application of machine learning to assess cardiac health. Data from various subjects with cardiovascular complications should be collected for this purpose.
APPNEDIX: K-MEANS CLUSTERING FIGURES
Figure 6.1: Time-Domain Features

(1) Subject 1 - Lung Volume First Five Minutes

(2) Subject 1 - Respiration First Five Minutes

(3) Subject 1 - Lung Volume Second Five Minutes

(4) Subject 1 - Respiration Second Five Minutes

(5) Subject 2 - Lung Volume First Five Minutes

(6) Subject 2 - Respiration First Five Minutes

(7) Subject 2 - Lung Volume Second Five Minutes

(8) Subject 2 - Respiration Second Five Minutes

(9) Subject 3 - Lung Volume First Five Minutes

(10) Subject 3 - Respiration First Five Minutes

(11) Subject 3 - Lung Volume Second Five Minutes

(12) Subject 3 - Respiration Second Five Minutes

(13) Subject 4 - Lung Volume First Five Minutes

(14) Subject 4 - Respiration First Five Minutes

(15) Subject 4 - Lung Volume Second Five Minutes

(16) Subject 4 - Respiration Second Five Minutes

(17) Subject 5 - Lung Volume First Five Minutes

(18) Subject 5 - Respiration First Five Minutes

(19) Subject 5 - Lung Volume Second Five Minutes

(20) Subject 5 - Respiration Second Five Minutes
Figure 6.1: Frequency-Domain Features

(1) Subject 1 - Lung Volume First Five Minutes
(2) Subject 1 - Respiration First Five Minutes
(3) Subject 1 - Lung Volume Second Five Minutes
(4) Subject 1 - Respiration Second Five Minutes

(5) Subject 2 - Lung Volume First Five Minutes
(6) Subject 2 - Respiration First Five Minutes
(7) Subject 2 - Lung Volume Second Five Minutes
(8) Subject 2 - Respiration Second Five Minutes

(9) Subject 3 - Lung Volume First Five Minutes
(10) Subject 3 - Respiration First Five Minutes
(11) Subject 3 - Lung Volume Second Five Minutes
(12) Subject 3 - Respiration Second Five Minutes

(13) Subject 4 - Lung Volume First Five Minutes
(14) Subject 4 - Respiration First Five Minutes
(15) Subject 4 - Lung Volume Second Five Minutes
(16) Subject 4 - Respiration Second Five Minutes

(17) Subject 5 - Lung Volume First Five Minutes
(18) Subject 5 - Respiration First Five Minutes
(19) Subject 5 - Lung Volume Second Five Minutes
(20) Subject 5 - Respiration Second Five Minutes
Figure 6.1: Time- and Frequency-Domain Features

(1) Subject 1 - Lung Volume First Five Minutes

(2) Subject 1 - Respiration First Five Minutes

(3) Subject 1 - Lung Volume Second Five Minutes

(4) Subject 1 - Respiration Second Five Minutes

(5) Subject 2 - Lung Volume First Five Minutes

(6) Subject 2 - Respiration First Five Minutes

(7) Subject 2 - Lung Volume Second Five Minutes

(8) Subject 2 - Respiration Second Five Minutes

(9) Subject 3 - Lung Volume First Five Minutes

(10) Subject 3 - Respiration First Five Minutes

(11) Subject 3 - Lung Volume Second Five Minutes

(12) Subject 3 - Respiration Second Five Minutes

(13) Subject 4 - Lung Volume First Five Minutes

(14) Subject 4 - Respiration First Five Minutes

(15) Subject 4 - Lung Volume Second Five Minutes

(16) Subject 4 - Respiration Second Five Minutes

(17) Subject 5 - Lung Volume First Five Minutes

(18) Subject 5 - Respiration First Five Minutes

(19) Subject 5 - Lung Volume Second Five Minutes

(20) Subject 5 - Respiration Second Five Minutes
LIST OF REFERENCES


