MOBILE MACHINE LEARNING FOR MONITORING HEART DISEASE

MOBILE MACHINE LEARNING FOR REAL-TIME PREDICTIVE MONITORING OF CARDIOVASCULAR DISEASE

BY

OMAR BOURSALIE, B.Eng.

A THESIS

SUBMITTED TO THE SCHOOL OF BIOMEDICAL ENGINEERING

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

Master of Applied Science (2016) McMaster University

(School of Biomedical Engineering) Hamilton, Ontario, Canada

TITLE:	Mobile Machine Learning for Real-time Predictive Monitoring of Cardiovascular Disease
AUTHOR:	Omar Boursalie B.Eng., (Electrical & Biomedical Engineering) McMaster University, Hamilton, Canada
SUPERVISOR:	Dr. Thomas Doyle & Dr. Reza Samavi
NUMBER OF PAGES:	xviii, 105

Lay Abstract

In this thesis, a novel mobile system for monitoring cardiovascular (CVD) disease is presented. The system allows for the continuous monitoring of both physiological sensors, data from a patient's health record and analysis of the data directly on the mobile device using machine learning algorithms (MLA) to predict an individual's CVD severity level. The system successfully demonstrated that a mobile device can act as a complete monitoring system without requiring constant communication with a remote server. A comparative analysis between the support vector machine (SVM) and multilayer perceptron (MLP) to explore the effectiveness of each MLA for monitoring CVD is also discussed. Both models were able to classify CVD severity with the SVM achieving the highest accuracy (63%) and specificity (76%). Finally, the resource requirements for each component in the system were evaluated. The results show that the MLAs complexity was not a barrier to deployment on a mobile device.

Abstract

Chronic cardiovascular disease (CVD) is increasingly becoming a burden for global healthcare systems. This burden can be attributed in part to traditional methods of managing CVD in an aging population that involves periodic meetings between the patient and their healthcare provider. There is growing interest in developing continuous monitoring systems to assist in the management of CVD. Monitoring systems can utilize advances in wearable devices and health records, which provides minimally invasive methods to monitor a patient's health. Despite these advances, the algorithms deployed to automatically analyze the wearable sensor and health data is considered too computationally expensive to run on the mobile device. Instead, current mobile devices continuously transmit the collected data to a server for analysis at great computational and data transmission expense.

In this thesis a novel mobile system designed for monitoring CVD is presented. Unlike existing systems, the proposed system allows for the continuous monitoring of physiological sensors, data from a patient's health record and analysis of the data directly on the mobile device using machine learning algorithms (MLA) to predict an individual's CVD severity level. The system successfully demonstrated that a mobile device can act as a complete monitoring system without requiring constant communication with a server. A comparative analysis between the support vector

machine (SVM) and multilayer perceptron (MLP) to explore the effectiveness of each algorithm for monitoring CVD is also discussed. Both models were able to classify CVD risk with the SVM achieving the highest accuracy (63%) and specificity (76%). Finally, unlike current systems the resource requirements for each component in the system was evaluated. The MLP was found to be more efficient when running on the mobile device compared to the SVM. The results of thesis also show that the MLAs complexity was not a barrier to deployment on a mobile device.

Acknowledgements

I would like to sincerely thank my supervisors, Dr. Thomas Doyle for his confidence and trust which allowed me to start on this journey; and Dr. Reza Samavi for his patience, guidance and support throughout my thesis. It has been an honour and privilege to have the opportunity to learn so much from the both of you.

I would also like to thank my colleagues and friends Devin Packer, George Su, Krystien Wolf, Warren Pawlikowski and Taralynn Shwering for their technical expertise. A special thanks to Micheal Wirtzfeld, Geneva Smith and Avery Chakravarty for their support and writing advice. You have all enriched my graduate experience.

Outside of McMaster, I could always count on the support of my family (Mom, Dad, Suz) and friends (Robert Cekan, Jessica Jolkowski and Calvin Bouwman). Without your never ending support I could not have finished this thesis.

Lastly, I must acknowledge McMaster University for funding me and giving me the opportunity to pursue graduate studies.

To my father, mother and sister

Contents

List of Figures

List of Tables

Notation and Abbreviations

AUC Area Under the Curve

BP Blood Pressure

CPU Central Processing Unit

DRGs Diagnosis Related Group

eHealth Electronic Health

GSR Galvanic Skin Response

HF High Frequency

HR Heart Rate

HRV Heart Rate Variability

ICD-9 International Classification of Diseases, Ninth Revision

ICU Intensive Care Unit

LF Low Frequency

M4CVD Mobile Machine learning Model for Monitoring Cardiovascular Disease

MIMIC Multiparameter Intelligent Monitoring in Intensive Care II Database

MLA Machine Learning Algorithm

MLP Multilayer Perceptron

N-N Normal-to-Normal Interval

NN Neural Network

NN50 Number of successive N-N interval greater than 50 ms

PDA Personal Digital Assistant

pNN50 NN50 divided by the total number of NN intervals

PSD Power Spectral Density

PTB Physikalisch-Technische Bundesanstalt Diagnostic ECG Database

RASPI Raspberry Pi 2 Model B

RMSSD Square root of the mean squared differences of successive N-N intervals

RnF Random Forest

ROC Receiver Operator Curve

RPM Remote Patient Monitoring

SAPS Simplified Acute Physiology Score I

SDNN Standard Deviation of the NN interval

SDR Spectral Distribution Ratio

SHAREE Smart Health for Assessing the Risk of Events via ECG

SQL Structured Query Language

Declaration of Academic Achievement

The following is a declaration that the research described in this thesis was completed by Mr. Omar Boursalie and recognizes the contributions of Dr. Thomas Doyle and Dr. Reza Samavi. Omar Boursalie contributed to the inception of the study, the study's design and was responsible for the experimental testing protocols, design and development of the system, data collection, data analysis and the writing of the manuscript. Dr. Thomas Doyle and Dr. Reza Samavi contributed to the inception of the study, the design of the study and the review of the manuscript.

Chapter 1

Introduction

The growing availability of wearable devices and health records provides new opportunities for the management of chronic diseases through predictive monitoring. In this thesis a new system for the remote monitoring of cardiovascular disease using machine learning algorithms on a mobile device is proposed. The system uses both wearable sensors to monitor physiological signals and health records to provide the patient's clinical data. In the following sections the background and motivations for the proposed system is introduced, the main contributions are summarized and the thesis outline is presented.

1.1 Problem Statement

Cardiovascular Disease (CVD), which includes heart attack, stroke and hypertension, is a chronic disease where the treatment aim is to maintain and improve the patient's current quality of life. CVD is managed through medication, lifestyle changes to diet and exercise that are tracked manually in a daily log. The Canadian Cardiovascular Society Guidelines also recommends periodic appointments in order to perform electrocardiograms (ECG) and blood pressure (BP) measurements (Mancini *et al.*, 2014) as patient deterioration is

preceded by changes in patient's physiological signals (Buist *et al.*, 1999). However, patients fail to maintain their logs effectively (Stone *et al.*, 2003). In addition, periodic appointments which can be weeks or months apart fail to recognize patient deterioration in time. The long-term prognosis of patients with chronic heart disease is low with one study showing 40% of patients dying within two years after being initially hospitalized after a heart attack (Goldberg *et al.*, 2007). CVD continues to increase due to an aging population and rising obesity which is intensifying the effects of chronic diseases (World Health Organization, 2010). CVD is an increasing burden on the healthcare system with 1.6 million Canadians reported having heart disease in 2009 (Dai *et al.*, 2009).

Remote patient monitoring (RPM) is becoming an attractive solution for the management of CVD. RPM uses wearable devices to continuously monitor a patient's physiological signals such as ECG (Pandian *et al.*, 2008), BP (Anliker *et al.*, 2004), stress levels (Sun *et al.*, 2012) and physical activity (Mattila *et al.*, 2009) in real time. The quantity of data acquired each day is large and manual analysis by health professionals is tedious and time consuming (Raghavendra *et al.*, 2011).

In this research the application of machine learning algorithms (MLA) for the automated predication of an individual's CVD severity level is investigated. Machine learning algorithms, a subset of artificial intelligence, are algorithms that are constructed to classify new data by learning trends and relationships from an existing dataset. In the medical field MLAs have been used for screening, diagnosis, treatment, prognosis, monitoring and disease management (Esfandiari *et al.*, 2014). The health data of patients with heart disease is used to "train" the MLAs to identify CVD severity levels. Once trained the MLAs can then be deployed to automatically classify new patients. Machine learning has been shown to increase prediction accuracy with less strict assumptions compared to statistical methods (Luo, 2015).

There are existing RPM systems but they suffer from a number of limitations. First,

existing RPM systems deploy the analysis stage on more powerful remote servers because of the high computational complexity of the MLAs. As a result, current RPM systems must be continuously transmitting the collected health data to the server for analysis. The continuous data transmission is expensive as physiological data from ECG alone can be 2.8 GB in one day (Sufi *et al.*, 2009). In addition, the RPM system cannot be used in poor coverage zones in rural areas or low-income countries. As a result, existing RPM systems are confined to controlled clinical settings. Second, the patient's clinical data is increasingly accessible through their health records containing valuable contextual information on the patient such as age, gender, weight and medication history. However, the literature review on the existing proposals for CVD monitoring devices shows that current systems do not automatically leverage the patient's clinical data when analyzing the physiological data. Instead, with current systems the health professional must implicitly consider the patient's clinical data when interpreting the existing RPM system's results. Third, current RPM systems using MLAs have focused only on reporting the system's accuracy. The existing systems have not been evaluated by their resource requirements such as execution time and current consumption.

1.2 Proposed System

To address the limitations in current devices a new system for the remote monitoring of CVD called M4CVD: Mobile Machine learning Model for Monitoring Cardiovascular Disease is proposed. Fig. 1.1 shows an overview of the proposed system architecture. M4CVD obtains input data from two sources: 1) physiological signals measured using wearable sensors and 2) clinical data from health records. The system improves on previous devices that have been limited to one input source. The system uses wearable sensors and external devices to collect observable trends from the ECG and BP vital signs. Health records provide context

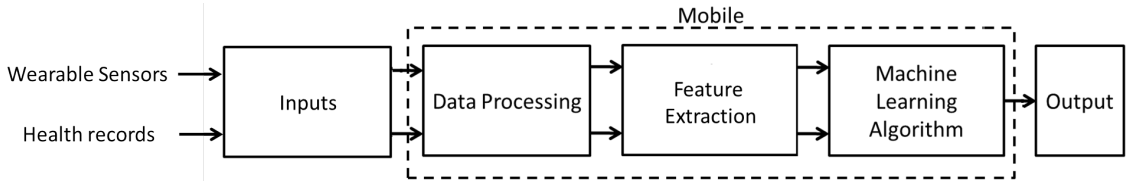


Figure 1.1: A generic overview of M4CVD's system architecture

to the physiological signals by including information such as age, gender and body mass index (BMI). Next, the raw signals were cleaned up by removing the noise and missing data corrupting the signal. In addition, the quality of the recorded ECG was calculated so corrupted signals were not analyzed. The hybrid input data then underwent data and signal processing to extract clinically useful features for heart disease (e.g., heart rate, BMI). The third stage was to analyze the hybrid of collected data. Instead of transferring the calculated features to a remote server M4CVD performs local analysis by feeding the hybrid features to a MLA. Two MLAs were investigated in this thesis: 1) the support vector machine (SVM) and 2) the multilayer perceptron (MLP). Both the SVM and MLP monitored the extracted features to classify a patient's CVD severity as "low" or "high". Finally, while not addressed in this thesis, the system could display and transmit the results to patients, caregivers and health professionals.

A training dataset containing clinical data (real-time physiological recordings and health data) from 518 patients was used to train the SVM and MLP. The SVM achieved an accuracy of $62.50 \pm 3.64\%$ with a maximum accuracy of 71.30%. The MLP overall had lower accuracy ($58.60\% \pm 6.61\%$) but had a maximum accuracy of 82.00%. The results were encouraging with a successful deployment of a binary classifier on a mobile device that analyzed both wearable and clinical data. Unlike existing work the computational complexity of the system was evaluated by examining M4CVD's execution time and current requirements. The MLP was more efficient on a mobile device with an execution time of 1.65 ms compared to the SVM which ran in 71 ms.

1.3 Thesis Contributions

In this thesis the first contribution was presenting a new model for a CVD remote monitoring system. The model offers a new approach to automatic remote monitoring by analyzing multi-source data from ECG and BP sensors as well as from health records. In addition, the M4CVD system was deployed entirely on a low resource mobile device (Raspberry Pi 2). Unlike current RPM systems, M4CVD does not require constant communication with a remote server saving computational and data transmission expenses.

Second, a comparative analysis of the SVM versus the MLP is presented to explore the effectiveness of each algorithm for monitoring CVD when deployed on a mobile device. The SVM and the MLP are compared because each has different properties and complexities for addressing the classification problem. The results show that both algorithms achieved similar classifier performance but the MLP was more efficient in terms of execution time and current consumption. This thesis has demonstrated the importance of considering both accuracy and system computational complexity when selecting classifiers for mobile applications.

The third contribution was to evaluate the resource requirements of an entire RPM system when deployed on a mobile device. The computational results for each stage in Fig. 1.1 can be used as benchmarks to future researchers. Surprisingly the results indicated that the MLA's complexity was not a barrier to mobile deployment. In fact, the analysis stage was less computationally expensive compared to the data acquisition and processing stages of the M4CVD system.

1.4 Organization of Thesis

The thesis is structured as follows:

Chapter 2 presents the literature review and gap analysis on the relevant research. Chapter 3 gives an overview of M4CVD's algorithm and describes the theoretical background. Chapter 4 describes the model implementation and outlines the experimental procedure for evaluating M4CVD. Chapter 5 presents the experimental validation and discusses the potential of the interpreted results. Chapter 6 presents conclusions and discusses directions for future work.

The author brings to the reader's attention that parts of this thesis was published in (Boursalie *et al.*, 2015). This publication is made by the author of this thesis, as the lead author, in collaboration with his supervisors at McMaster University. The systematic review of the related literature and identifying the gaps in the areas of remote monitoring in Chapter 2 are the contributions that have only been published in this thesis. The experimental implementation and discussion in Chapter 4 and 5 are contributions that have only been published in this thesis.

Chapter 2

Literature Review

In this chapter, a literature review on current RPM systems will be presented following the guidelines for a systematic literature review in software engineering (Keele, 2007; Okoli, 2015). The systematic literature review tries to answer the question- are machine learning algorithms being deployed on mobile devices for monitoring diseases? The study is summarized in three stages: 1) planning 2) conducting and 3) reporting the review. These three stages are examined in Sections 2.1, 2.2 and 2.3 respectively.

2.1 Planning the Review

First, the procedure is explained for selecting the keywords. Next, the methodology used for conducting the review is discussed. Last, the exclusion criteria of the review is presented.

2.1.1 Search Terms and Research Databases

In this thesis the search was for academic papers that present devices, systems or algorithms for monitoring CVD. The following terms are considered: data mining, wearable plus synonyms such as wearable systems, mobile plus synonyms like smartphone, wireless,

CVD or cardiovascular disease, patient monitoring, ECG or electrocardiogram, healthcare, telemedicine, SVM or support vector machine, biomedical and machine learning. These keywords were used to create the advanced search queries. The following academic databases were examined: Google Scholar, IEEE Explore, PubMed, Science Direct and Springer Link.

2.1.2 Exclusion Criteria

Papers that presented devices not monitoring cardiovascular disease (such as diabetes) were not examined in this review.

2.2 Conducting the Review

The final search queries based on the search terms in Section 2.1.1 and the number of returned results are summarized in Table 2.1. The other databases were still considered in the analysis and the full results are summarized in Appendix A.

Table 2.1: Advanced Search Queries

Query	**Database**	**#Results**	**Selected**
(data mining) AND (wearable OR wearable system) AND (mobile OR smartphone) AND (CVD OR cardiovascular disease) AND (patient monitoring OR monitoring) AND (ECG OR electrocardiogram) AND (healthcare) AND (telemedicine) AND (SVM OR support vector machine) AND (biomedical) AND (machine learning)	Google Scholar	136	18
"data mining" AND "wearable" OR "wearable system" AND "mobile" OR "smartphone" AND "cvd" OR "cardiovascular disease" AND "monitoring" AND "ecg" AND "telemedicine" AND "svm" OR "support vector machine" AND "machine learning" AND "biomedical"	IEEE Xplore	488	33

Initial review of the search results was done by reviewing the title and abstracts to determine if the paper meets the criteria described in Section 2.1.1. Papers selected were then filtered through the exclusion criteria. 67 papers were selected for data extraction. The papers were categorized into three categories cover the research domain of this thesis:

1. CAT1- Monitoring system with server analysis

2. CAT2- Monitoring System with local analysis
3. CAT3- Mobile phones

A quality appraisal in terms of the paper's relevancy to this thesis was conducted in the following areas:

1. Q1- System uses a MLA
2. Q2- System uses a mobile phone
3. Q3- System uses multiple sensors
4. Q4- System analyzes clinical data

The relevancy of each paper was scored from 0 - 2 where 2 was very relevant. The total score assigned over all four categories describes how relevant the research is to the thesis. A subset of the result of the extraction process is summarized in Table 2.2 while the full results are in Appendix A.

Table 2.2: Subset of Papers for Data Extraction

Category	**Paper**	**Q_1**	**Q_2**	**Q_3**	**Q_4**	**Total**
CAT2	(Bellos *et al.*, 2013, 2014)	2	2	2	2	8
CAT2	(Bellos *et al.*, 2010a, 2011a,c, 2012)	2	2	2	0	6
CAT3	(Krause *et al.*, 2006)	2	1	2	0	5
CAT2	(Clifton *et al.*, 2011)	2	0	2	0	4
CAT1	(Bellos *et al.*, 2010b)	2	0	0	2	4
CAT1	(Villalba *et al.*, 2009)	0	2	2	0	4
CAT1	(Takata *et al.*, 2008)	0	2	2	0	4
CAT3	(Melillo *et al.*, 2015a)	2	0	2	0	4
CAT1	(Baron *et al.*, 2011)	0	0	2	0	2
CAT1	(Patel *et al.*, 2012)	2	0	0	0	2

2.3 Reporting the Review

In this section a summary of the contributions from the selected papers is provided according to the four research areas of interest outlined in Section 2.2. First, systems using MLAs are discussed in Section 2.3.1. Next, RPM systems using mobile devices are examined in Section 2.3.2. Finally the wearable sensors and clinical data currently used in RPM systems are investigated in 2.3.3 and 2.3.4 respectively. The properties of the most relevant systems are summarized in Table 2.3. Over half the systems reviewed used MLAs to analyze the collected data and most systems deploy the MLAs on remote servers. The review also indicates that health records are not being leveraged automatically by RPM systems.

Table 2.3: A comparison between M4CVD and other monitoring systems

Paper	Number of Sensors	ECG?	Conduct Local Analysis?	Conduct Server Analysis?	MLA Analysis?	Mobile?	Database?	Clinical Data?
M4CVD (2016)	2	Y	Y	N	Y	Y	Y	Y
M4CVD (2015) (Boursalie *et al.*, 2015)	3	Y	Y	N	Y	Y	Y	Y
CHRONIOUS (Bellos *et al.*, 2011b)	8	1	Y	Y	Y	Y	Y	Y
(Solar *et al.*, 2013)	5	Y	Y	Y	Y	Y	Y	N
(Gultepe *et al.*, 2014)	7	N	N	Y	Y	N	Y	Y
(Guidi *et al.*, 2014a)	0	Y	N	Y	Y	N	Y	Y
(Clifton *et al.*, 2014)	4	Y	N	Y	Y	Y	Y	N
(Alshurafa *et al.*, 2014)	3	Y	N	Y	Y	Y	N	N
(Krause *et al.*, 2006)	3+	N	Y	N	Y	Y	Y	N
(Oresko *et al.*, 2010)	1	Y	Y	N	Y	Y	-	N
(Liu *et al.*, 2011)	9	Y	N	Y	Y	N	Y	N
(Chen *et al.*, 2005)	3	Y	N	Y	N	Y	N	N
(Torres-Huitzil and Nuno-Maganda, 2015)	1	N	Y	N	Y	Y	N	N
(Luo, 2015)	0	N	N	Y	Y	N	N	Y
(Juen *et al.*, 2015)	1	N	Y	N	Y	Y	Y	N
(Kunnath *et al.*, 2013)	2	Y	Y	Y	N	Y	N	N
(Raghavendra *et al.*, 2011)	1	Y	Y	N	Y	Y	Y	N
(Leite *et al.*, 2011)	5	Y	N	Y	Y	N	Y	N
(Katsaras *et al.*, 2011)	4	Y	N	N	N	Y	N	N
(Clifton *et al.*, 2011)	4	Y	N	Y	Y	N	Y	N
(Villalba *et al.*, 2009)	5+	Y	N	Y	?	Y	N	N
(Takata *et al.*, 2008)	3+	Y	N	Y	?	Y	Y	N
(Krause *et al.*, 2005)	1	N	Y	N	Y	Y	N	N
(Melillo *et al.*, 2015a)	2	Y	N	Y	Y	Y	Y	N
(Shih *et al.*, 2010)	2	Y	N	Y	Y	Y	Y	N
(Pandian *et al.*, 2008)	6	Y	N	Y	N	Y	N	N
(Kailanto *et al.*, 2008a)	1	Y	N	Y	N	Y	N	N
(Anliker *et al.*, 2004)	1+	Y	N	Y	N	Y	-	N
(Prabhakara and Kulkarni, 2014)	1	N	N	Y	-	Y	N	N
(Depari *et al.*, 2014)	1	Y	N	N	N	Y	N	N
(Jung *et al.*, 2014)	0	N	N	Y	N	Y	N	Y
(Gao *et al.*, 2013)	1	Y	Y	N	N	Y	N	N
(Mayton *et al.*, 2012)	3+	N	N	Y	N	Y	N	N
(Baron *et al.*, 2011)	5	Y	N	Y	-	Y	-	N
(Patel *et al.*, 2010)	1	N	N	Y	Y	N	Y	N
(Zhu *et al.*, 2007)	0	N	N	Y	Y	N	Y	N
(Özkaraca *et al.*, 2011)	1	Y	N	N	N	Y	N	N

2.3.1 Systems using Machine Learning Algorithms

MLAs are increasingly being used in the medical field for screening, diagnosis, treatment, prognosis, monitoring and disease management (Esfandiari *et al.*, 2014). The results of the literature review indicates that MLAs are being used in RPM systems for two main applications: 1) novelty detection and 2) severity classification.

Novelty detection is the identification of abnormal signals from long term vital sign monitoring. ECG signals are a good candidate for automated abnormal detection (e.g., arrhythmia detection) as manual interpretation of the long term data is very time consuming. Novelty systems have previously identified abnormal signals by comparing extracted ECG features such as heart rate (HR) against feature thresholds that trigger an alarm when it is exceeded. However, health professionals ignore the system's alarms due to high false positive rates (Tsien and Fackler, 1997). MLAs offer improved methods for novelty detection as the algorithms consider the cumulative effect of all the features in the signal. Shih *et al.* used an embedded system to alert the hospital when it detected abnormal beats achieving a classification accuracy of 90.8-97.8% (Shih *et al.*, 2010). Another study describes the implementation of a classifier of ECG beats with a naive bayes classifier with the aim of being used for real-time detection on a mobile environment (Raghavendra *et al.*, 2011).

Monitoring systems have also applied MLAs to abnormality detection when monitoring multiple physiological signals simultaneously. Clifton *et al.* used intensive care unit (ICU) monitors to analyze patient respiratory rate, HR and BP (Clifton *et al.*, 2011). The system was able to detect periods of normality and abnormality with an accuracy of 95% using an SVM. Each physiological signal was assigned its own MLA threshold above which an alert was generated. A RPM was then developed that used wearable sensors (ECG, HR, SpO2) for ambulatory monitoring allowing the patient to move freely within the hospital (Clifton *et al.*, 2014). The system demonstrated that predictive monitoring with mobile sensors is feasible with the SVM achieving an accuracy of 94%. However the system only works within

the ICU as all the collected data are continuously transmitted to servers for analysis. The system only considered each vital sign individually without considering their cumulative effect. MLAs are well suited for novelty detection tasks as these systems can be trained on a small training set which contains a high ratio of normal-to-abnormal training examples. In addition such systems can also be easily individualized by training them with the patients own signals while they are in the healthcare setting. However, novelty systems are one-class classifiers (normal or not) that cannot be used for assessing different risk levels (e.g., low or high).

When sufficient training examples of each target risk level (low/high) is available a two-class severity model can be constructed. Melillo *et al.* monitored 139 patients with a Holter ECG sensor and performed follow up appointments 12 months later. Using heart rate variability (HRV) features they were able to classify patients severity level as low or high for a vascular event using a random forest (RnF) classifier achieving an accuracy of 91.8% (Melillo *et al.*, 2015a). While this was a follow up study the trained system could be deployed with new patients to evaluate their heart disease severity level. Similarly, Liu *et al.* demonstrated that HRV values can be used with a machine learning system to predict mortality rates of critically ill patients with an accuracy of 70-78% using a SVM (Liu *et al.*, 2011). Leite *et al.* developed an automated intelligent system that monitors BP, HR, RR and temperature using fuzzy models and neural networks (NN) (Leite *et al.*, 2011). The system had four alert levels (none, low, medium and high) achieving an accuracy of 94%. The proposed RPM system was not mobile, did not use wearable sensors and was designed to be used only within the ICU.

Another interesting finding from the literature review is that MLA systems in the medical domain use small training datasets compared to other MLA fields. Medical systems with MLAs use average training set sizes of only a few dozen to a few hundred patients. In contrast MLA systems in other disciplines (such as image recognition) use training sets

with hundreds of thousands of training examples. Clinical databases containing physiological and clinical data are becoming more widely available. However, most studies reviewed involved their own internal data collection stage in which to collect physiological and clinical data from human subjects. The small training sizes thus reflects the difficulty and expense in recruiting large number of appropriate patients for longitudinal studies as well as the increased privacy concerns compared to other MLA fields. To overcome this it is important the training set used is a good representation of the user population. The small datasets used may explain the popularity of the SVM in RPM systems since SVMs generalize well on small datasets.

The review demonstrates that MLAs can be used in monitoring applications with high accuracy. However the MLA systems reviewed have been primarily server based despite the opportunities for mobile applications. One explanation for the popularity of desktop/ server based systems is the higher computational power available on desktop systems compared to mobile systems. However another limitation is the significant computing expertise that makes machine learning inaccessible to many healthcare researchers without a background with MLAs (Luo, 2015). In the last decade MLAs have become more accessible with MLAs tools in programs such as MATLAB, R (R Core Team, 2013) and WEKA (Hall *et al.*, 2009). But these programs are all desktop based encouraging the continued use of server or cloud services. RPM systems are limited due the requirement of constant communication with the remote server which is not feasible outside of controlled settings. M4CVD goes beyond novelty detection of single physiological signals by considering multiple vital signs. The proposed system can be deployed outside the laboratory due to the analysis stage being located on the mobile device.

2.3.2 Systems using Mobile Devices

Mobile devices provide a convenient platform in remote monitoring services due to their ubiquity and patient familiarity. In the literature, mobile devices have been primarily used in RPM systems as the acquisition and transmission module. As the computational power of mobile devices has grown systems have begun starting to use them for signal processing and local analysis. As a result, the resource requirements of these monitoring systems has become a growing area of study.

The mobile phone has long been used as a replacement to a stand-alone communication module in RPM systems. Chen *et al.* used a mobile phone with a wearable sensor unit containing ECG, temperature and accelerometer. The system integrated with a health record management layer using decision rules to provide feedback to the patient and health professionals (Chen *et al.*, 2005). Mobiles phones also allow patients to interact with the system. Wanda-CVD is a smartphone-based monitor that transmits sensor and questionnaire data to a remote server for predicting positive outcomes (Alshurafa *et al.*, 2014). A random forest classifier with 100 trees was found to have highest accuracy with an area-under-the-curve (AUC) of 92.4%. 90 patients (African American women aged 25-45) used Wanda-CVD for six months with the intervention group (45 patients) showing a 55-60% decrease in cholesterol levels demonstrating the potential of RPM to assist patients. The MyHeart Project used external sensors embedded in clothes, bed, weight scale and blood pressure monitors to manage heart failure (Villalba *et al.*, 2009). A PDA was used to receive data and transmit it to a remote server for further analysis. In addition MyHeart used a PDA to collect questionnaire data and to guide the patient to take accurate measurements with the sensors. Health professionals can view all patient's data through the web portal. The system was tested on 37 patients with no usability or acceptability problems reported. MyHeart demonstrated how a mobile system allows patients to interact and become proactive in managing their health. Unlike other systems MyHeart was not passively recording ECG data but actively

instructed patients on proper recording techniques to acquire suitable ECG recordings for assessment.

Increasing data processing power has allowed mobile systems to become capable of conducting signal processing locally. Kalianto *et al.* developed an ECG platform that records 2-channel ECG and transmits the data to a mobile phone (Nokia 6681) for local analysis to compute and display the patient's heart rate (Kailanto *et al.*, 2008b). The authors found that HR calculations were completed quickly. However, saving the physiological data to the phone took most of the processing time. Depari *et al.* developed a smartphone application that integrates with two ECG sensor bands worn by the patient. The system conducts time and frequency analysis using modulation techniques to acquire the ECG signal. Basic threshold detection was used to evaluate the heart beat with the ECG data being stored in the smartphone memory for offline analysis (Depari *et al.*, 2014). The smartphone was also used to power the sensor unit via the audio output. However, the battery life of the system was not investigated.

MLAs can be also be deployed directly on mobile devices due to their increasing computational power. An early example of a MLA on a mobile system was the SenSay system that used a backpack laptop and PDA to identify the optimal time to interact with the user (Krause *et al.*, 2006). SenSay used sensors (tracking location, activity, physiology, schedule and ambient context), baysian networks and K-means clustering to create a user context model. SenSay demonstrated that a mobile platform can be ubiquitously used to measure patient context for an extended period of time. Similarly, Takata *et al.* created LifeLog to analyze heart rate, step meter, GPS and video recorder directly on a laptop. LifeLog was successfully able to provide context to recorded data to determines the users current action (Takata *et al.*, 2008). Oresko *et al.* presented HeartToGo, a mobile phone that performed data acquisition, feature extraction and classification of ECG signals (Oresko

et al., 2010). The system used MLAs achieving an accuracy of 90% in real-time CVD detection but monitored a single wearable sensor. M4CVD will improve on HeartToGo by measuring data from multiple sensors and medical records. Solar *et al.* developed a wearable platform to monitor congestive heart failure patients. The wearable platform contained sensors for ECG, sweat and accelerometers with the latter being used for activity recognition (80.3% accuracy). The MLA was used to estimate energy expenditure based on activity recognition and heart rate (Solar *et al.*, 2013). The system performed primary processing on a server however the system also deployed a "lite" version on the PDA in the case of loss of communication with the remote server. However the exact specifications of the "lite" version are not sufficiently explained. The training set used was small with 4 healthy volunteers enrolled to test the activity recognition and energy expenditure components of the system. The mobile device of the RPM system remains primarily an acquisition and processing platform with the real risk assignment being made by the health professional.

The closest system to M4CVD found in the systematic literature was the CHRONIOUS system (Bellos *et al.*, 2010b). CHRONIOUS is an integrated platform for the management and real-time assessment of the health status of patients suffering from chronic obstructive pulmonary disease and chronic kidney disease. The system was composed of a smart shirt of wearable sensors (Bellos *et al.*, 2010b) and a PDA containing MLA classifiers for severity assessment (Bellos *et al.*, 2011c). The health data is transmitted to a remote clinical server for additional analysis and monitoring by health professionals (Bellos *et al.*, 2011a). CHRONIOUS uses a hybrid severity system (SVM, random forest and rule based) to classify the patients current health status into five levels of severity achieving a characterization accuracy of 95% (Bellos *et al.*, 2011c, 2014). However, CHRONIOUS does not utilize clinical data from health records and the system's power consumption requirements were not investigated as is done in this thesis.

A key difference between mobile systems and remote servers is the limited computational

resources available. One resource is battery life which is a major constraint on wearable platforms. Understanding the resource requirements is a key metric to assess the systems overall usability and to identify areas for system improvement. However, despite the importance of energy characterization few studies have investigated these system requirements (Bhargava *et al.*, 2003; Comito and Talia, 2015). One that has investigated the energy requirements is Krause *et al.* who investigated the trade-off between power consumption and prediction accuracy of their eWatch wearable platform (Krause *et al.*, 2005). eWatch used a MLA to analyze microphone, light and temperature sensors to determine user context. A range of sampling frequencies and classification configurations were investigated. Krause *et al.* determined that while accuracy does increase with higher sampling there is a threshold at which accuracy improvement is small but comes at the continued expense of battery life. These results would enable the device to choose different configuration modes to balance power consumption and system accuracy. Comito *et al.* discuss how the next generation of mobile systems should be designed to minimize energy consumption (Comito and Talia, 2015). They developed a smartphone application to test three MLAs: J48, K-means and association rules. Comito *et al.* examined CPU, memory and energy requirements. While MLAs were found to be computationally intensive it was determined that appropriate tuning of parameters can lead to improved performance. Mobile systems such as CHRONIOUS focus on reporting system accuracy which enables comparison of system performance with other remote-server based systems. However the existing system's resource requirements has not been investigated which limits the assessment of the overall usability of the CHRONIOUS systems. In this thesis the resource requirements of M4CVD will be investigated to identify limitations in the proposed system.

2.3.3 Systems Analyzing Sensor Data

In this section strict acquisition systems reviewed in the literature are briefly discussed. Multiple devices record ECG for extended periods of time such as CardioNet, a commercial wearable ECG platform that transmits abnormal ECG patterns to a remote monitoring station for manual analysis (Chan *et al.*, 2012). CardioNet can monitor patients for up to 30 days, an improvement over the Holter monitor (24 hours) but only measures a single physiological signal with no contextual data and no local processing. A further improvement on these single monitoring devices are multi-parameter systems such as SmartVest (Pandian, 2008). SmartVest is a washable shirt that contains sensors for ECG, temperature, BP and galvanic skin resonse (GSR), which are transferred to a remote server to extract HR, R-R interval and GSR duration. However the systems relies on strict threshold detection to identify when one or more signals are out of range. An improvement on SmartVest is AMON, a wrist device monitoring BP, SpO2, ECG and accelerometers (Anliker, 2004). Unlike previous devices AMON looks at combination of the overall features. Only when multiple features are out of range was an alert sent to the server for analysis. The software used for analysis in AMON is proprietary and could not be studied. AMON goes beyond data acquisition to assist the physician in monitoring the patients health by only displaying abnormal recording. In addition the patient's clinical data (age, weight and height) are also displayed. However a medical professional was still required to analyze all of the recorded data. Unlike SmartVest and AMON, the HealthWear system uses the patient's health record for the storage of the recorded data. HealthWear is a chronic patient monitoring system that used a garment and external sources (such as oxygen saturation) with a mobile connection to connect the data to a patient's health record (Katsaras *et al.*, 2011). The system underwent a clinical trial with 48 male patients. The intervention group showed a reduced number of hospital visits (2 vs 32 in control group) and an overall hospital stay that was half of the control group (3.6 vs 6.8 days). Most users (20/24) found the garment acceptable

but the system had a 13% failure rate in transmitting to the remote server. HealthWear demonstrated that remote monitoring systems can be used by patients outside of clinical settings but highlights the limitations of constant cellular communication. Wearable sensors are becoming smaller, more ubiquitous and getting wider acceptability with patients and the public. These acquisition must constantly transmit to remote servers which is expensive and not always viable outside the clinical setting. The device proposed in this thesis will improve upon existing CVD monitoring systems by increasing the monitored input data and by using an MLA to perform data analysis on a mobile device without the need for constant server communication.

2.3.4 Systems Analyzing Clinical Data

Systems that exclusively analyze clinical data are not considered continuous monitoring systems and were not part of the systematic review. However, several papers did use clinical data and physiological parameters (e.g., HR) that were manually entered into the clinical database. Guidi *et al.* developed a hospital based system designed to identify patients at short and long term risk for heart failure (Guidi *et al.*, 2014b). The clinical based decision support system investigates the use of a MLA to evaluate heart failure severity and predict heart failure type. The RnF algorithm was found to have the best performance with an accuracy of 83.3% and 86.6% in severity and type prediction respectively. Guidi *et al.* system highlights the challenges for short and long term monitoring using a medical record only system without using wearable sensors. Gultepe *et al.* developed a decision support system to identify patients at high risk of sepsis using a SVM (Gultepe *et al.*, 2014). Similar to CVD management, compliance with sepsis management guidelines can improve patient outcomes and the current burden is with monitoring large groups of patients. The training database of sepsis-only patients used by Gultepe *et al.* is similar to the training set of CVD-only patients used in this thesis. The SVM was reported to have the best accuracy with 72.6%

but the system only analyzed vital signs (temperature, respiratory rate and lactate levels) and not clinical data. An advantage of systems analyzing only health record databases is the large number of records available for training. The sepsis training set used by Gultepe *et al.* contained 741 adult patients. However the sepsis decision support system was only designed for the clinical setting as all readings are entered into the system while the patient is in the hospital.

An interesting finding from the literature review is that most of the systems that collected clinical data only displayed it to the end user. The clinical data was not included in the automated analysis stage. One explanation was proposed by Raghavendra *et al.* who states that the physiological signal from a single patient will inherently register relevant clinical data such as age and gender (Raghavendra *et al.*, 2011). However, the authors disagree since this contradicts medical training that considers clinical data when health professionals review ECG data manually. The authors believe that including medical data along with physiological signals will provide important insights compared to classifying physiological signals alone.

Analyzing medical record databases is a growing research area. Health databases are also a promising platform to store physiological data recorded by wearable sensors. However, clinical data is only manually updated by healthcare professionals during a patient's visit and so is updated less frequently compared to wearable systems. Unlike training sets for wearable sensors, health record databases contains large patient populations for the training of MLA. By combining wearable sensors and data from health records, M4CVD will be able to rapidly interpret physiological measurements in the context of a patient's medical history to calculate their CVD severity level.

2.4 Literature Review Summary

Based on the papers reviewed, current monitoring systems face a number of limitations. The findings show that: 1) MLAs are being used to analyze health data however in monitoring applications this is mostly conducted on remote servers. 2) Despite mobile phones increasing in computational power they are still primarily used as simple signal acquisition devices despite their ability to run MLAs locally. The resource requirements (such as battery life) of the mobile systems is also not sufficiently investigated. 3) While more monitoring systems are collecting data from multiple sensors they are still analyzing them against individual thresholds rather than considering the overall cumulative effect. 4) Finally, clinical data is underused in monitoring systems which depend on manual consideration when the health professional reviews the system's results. M4CVD will improve upon existing CVD monitoring systems by analyzing data from wearable and health record databases. The analysis will be completed on the mobile device and does not require constant communication with remote servers. The resource requirements of the M4CVD system on a mobile device are also investigated. While previous work has proven the effectiveness of the individual components of the proposed device, no work has integrated these characteristics.

Chapter 3

Architecture Overview and Theoretical Background

In this chapter the system architecture and background material from which this thesis draws is presented. Fig. 3.1 is referred to throughout the thesis and shows the system architecture for M4CVD. The components of the system are Input (Section 3.1), Preprocessing (Section 3.2), Feature Extraction (Section 3.3), Discretization and Normalization (Section 3.4), Machine Learning (Section 3.5) and System Output (Section 3.6). Preprocessing, Data Processing and Machine Learning are all performed on the mobile device.

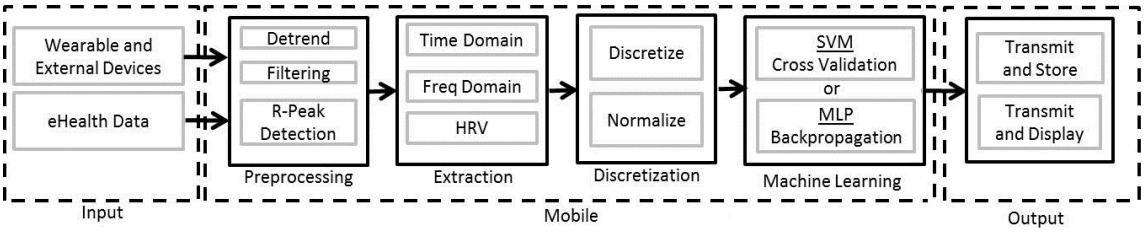


Figure 3.1: The proposed system architecture for M4CVD

3.1 Inputs

Table 3.1 lists the risk factors for heart disease from The Canadian Heart and Stroke Foundation (Heart and Stroke, 2013). The proposed system obtains input data from two sources to monitor these CVD risk factors: 1) physiological signals measured using wearable sensors/ external devices and 2) clinical data from health records. Each input type has its advantages and disadvantages. By combining the input data M4CVD will be able to rapidly interpret physiological measurements in the context of a patient's medical history to calculate the patient's CVD severity level.

Table 3.1: Monitoring cardiovascular disease risk factors

Feature	Wearable Sensor?	Clinical Database?
Age	N	Y
Gender	N	Y
Body Mass Index	N	Y
Systolic Blood Pressure	Y	Y
Diastolic Blood Presure	Y	Y
Cholesterol	N	Y
ECG	Y	N

3.1.1 Wearable Sensors and External Devices

Patient deterioration is always preceded by noticeable changes in several physiological parameters which can be monitored using wearable sensors (Solar *et al.*, 2013). Wearable sensors are worn by the patient and the signals are recorded by an embedded device such as a data acquisition microcontroller. Wearable sensors offer several advantages: they are unobtrusive in daily life and continue to decrease in cost while increasing in processing power (Zheng *et al.*, 2014). This thesis focuses on ECG and BP monitors because they are important risk factors for CVD and both types of sensors are increasingly accepted

by the public. An ECG wearable sensor was used to record the electrical activity of the heart (Köhler *et al.*, 2002). A 3-lead ECG sensor was used because it achieves a balance between recording the heart's activity and the patient's comfort when using the system. The ECG acquisition procedure followed the current gold standard guidelines outlined by the European and North American Cardiology Society Task Force (Camm *et al.*, 1996). The ECG signal was sampled at 250 Hz for a total of 5 minutes allowing for short term time and frequency domain feature analysis. For BP monitoring an external monitor was used because a viable nonintrusive and continuous blood pressure wearable sensor remains an active area of research. Wearable sensor data, which are not typically stored in clinical databases due to their large file size, allows for the examination of observable long-term trends but lacks contextual information.

3.1.2 Clinical Database

Clinical context is an important consideration when analyzing physiological signals. For example, what is considered a normal range of a patient's heart rate variability has been shown to be dependent on a patient's age (O'Brien *et al.*, 1986). When health professionals analyze a patient's physiological signal it is done within the context of the patient's clinical information such as their age, gender, BMI and medication history. However, despite the importance of clinical data most RPM systems only display clinical data. The clinical data is not included as part of the automated analysis. In addition, the clinical data must be updated manually by the patient which can result in error-prone and incomplete records. Instead of manual input, M4CVD could automatically retrieve the information from a patient's health record to provide clinical context. Health records are updated by healthcare professionals in a clinical setting so mistakes from self-measurement and manual input are minimized (Jung *et al.*, 2014). However, compared to wearable systems, health records are updated less frequently and so any clinical trends from the health record data

alone will take longer to be detected.

3.2 Preprocessing

In this section the preprocessing steps used to improve the quality of the physiological signals and clinical datasets will be discussed. Raw physiological signals suffer from noise, motion artifact and missing data which corrupts the quality of any features extracted (Banaee *et al.*, 2013) while data from clinical databases suffers from incomplete records with missing data.

3.2.1 Physiological Signal Preprocessing

The ECG signal undergoes three preprocessing steps: 1) filtering, 2) detrending and 3) QRS detection as shown in Figure 3.2. First, the ECG signal was filtered using a 4^{th} order Butterworth bandpass filter (0.05-50 Hz) to remove noise, motion artifact and other physioelectric signals such as muscle activity that are not currently under observation (Ellis *et al.*, 2015). Next, the ECG signal baseline wandering caused by the patient's breathing pattern was removed by detrending the signal. Detrending the signal enables the use of beat detection and feature extraction algorithms from a common baseline. Finally, most ECG feature extraction techniques require the location of the QRS signals within an ECG recording. Automatic beat detection is an active area of research with multiple methods such as Pan-Tompkin (Pan and Tompkins, 1985), wavelet, neural networks and adaptive filter have been proposed. A full review of automatic beat detection techniques is outside the scope of this thesis; interested readers are referred to Köhler *et al.* (2002) for a comprehensive review. A computationally simple method is to apply a threshold to isolate the tallest component of the signal which is assumed to be the QRS complex. Another example is the Pan-Tompkins Algorithm which bandpass filters and differentiates the signals before applying a variable peak threshold to identify the QRS complex (Pan and Tompkins,

1985). Multiple QRS detection techniques were investigate for this thesis and the results are discussed in Chapter 4.

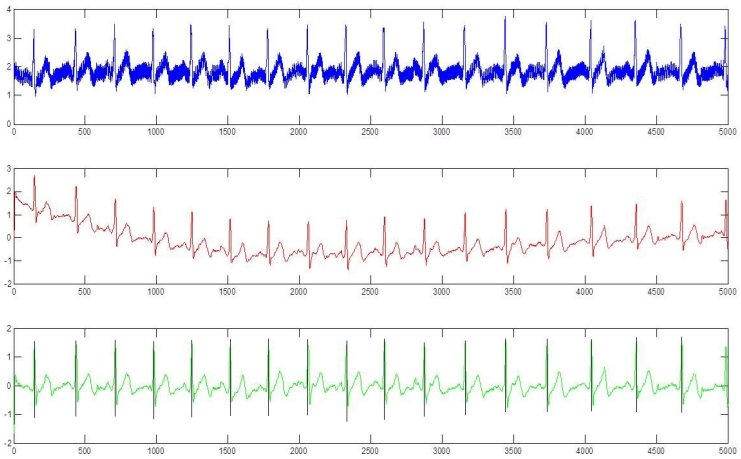


Figure 3.2: The raw ECG signal (top) was filtered (middle) and then detrended (bottom) to enable clean feature extraction. The location of each QRS complex within the ECG signal was then determined (bottom, black lines)

However, despite the preprocessing steps, in some cases the noise and motion artifact was too large to be correctly removed and the ECG signal was not suitable for feature extraction. To prevent low quality signals from being analyzed each ECG signal was also evaluated for overall signal quality using the metrics described by Li *et al.* (2008). Specifically two metrics were used to assess signal quality: 1) Spectral distribution ratio (SDR) and 2) Kurtosis.

Spectral distribution ratio (SDR), Eq. 3.1, is the ratio of the sum of power, P, within the QRS band (f = 5-14 Hz) to the power in the overall signal (f = 5-50 Hz) (Murthy *et al.*, 1978). A good ECG signal has been shown to have a SDR between 0.5 and 0.8 (Li *et al.*, 2008) while a low SDR indicates high frequency noise contamination such as from muscle artifact. On the other hand, a high SDR is an indication of electrode motion which introduces motion artifacts.

$$SDR(k) = \int_{f=5}^{f=14} P(k)df / \int_{f=5}^{f=50} P(k)df \tag{3.1}$$

Kurtosis, $\hat{K}$, is a metric of how uniform a signal is by measuring the relative peakness of a distribution with respect to a Gaussian distribution. Kurtosis is defined in Eq. 3.2 with ECG signal (x_i), mean (μ_x) and standard deviation ($\hat{\sigma}$) over M samples (Li *et al.*, 2008). A clean ECG signals has a kurtosis greater than 5 (He *et al.*, 2006) while muscle artifact and baseline wander corrupts the signal resulting in a kurtosis value less than 5 (Clifford *et al.*, 2006).

$$\hat{K} = \frac{1}{M} \sum_{i=1}^{M} \left[\frac{x_i - \hat{\mu_x}}{\hat{\sigma}} \right]^4 \tag{3.2}$$

In this thesis an overall ECG quality metric combining the SDR and kurtosis metrics was used. The overall metric was modified from Li *et al.* (2008) and is shown in Eq. 3.3.

$$\text{Overall ECG Quality} = \begin{cases} 1, & \text{if } 0.8 \geq \text{SDR(k)} \geq 0.5 \text{ and } \hat{K} \geq 5 \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

An ECG signal with a combined metric of 0 indicates a poor signal quality (either from poor SDR or kurtosis) and was not considered for analysis during the training or deployment of the MLA. In this case, M4CVD would take a new ECG sample to record a cleaner signal.

3.2.2 Clinical Data Preprocessing

Missing and incomplete data is a common challenge when working with health records (Rubin, 1976; Wells *et al.*, 2013). In this thesis, the clinical database (introduced in Chapter 4) suffered from missing data for patient height (33%), both weight and age (6.4%) and blood

pressure (4.4%). These incomplete records cannot be used for training the proposed MLAs. One approach to address incomplete data is list-wise deletion where any patient record with missing values is deleted. List-wise deletion is simple to execute but can greatly reduce the total number of training examples available. List-wise deletion has also been shown to introduce bias in cases when the missing data is not randomly distributed (Little and Rubin, 2014). In this thesis list-wise deletion was used when a data point (e.g., weight, age and BP) was missing in the record and the impact of the deleted records was not significant on the training set. However, deleting all records that were missing height resulted in the training set being decreased by 30% which was considered significant due to its impact on computing patients' BMI. Therefore, other statistical methods were investigated to substitute patients' missing height values as described below.

Mean imputation, K-nearest neighbours and regression imputation were investigated to replace the missing height feature with substituted values (Gelman and Hill, 2006). Mean imputation replaces the missing value by the mean value of all the complete patient examples maintaining the sample size but reducing the variability in the data (Eekhout *et al.*, 2012). Instead of using the mean value of all of the complete records, k-nearest neighbours first finds the most similar complete records to the target record (Batista *et al.*, 2002). The imputation value for the missing record was then only calculated from the mean of the K-nearest neighbours instead of the whole database. K-nearest method preserves the distribution but it does underestimate the variability in the dataset (Roth, 1994). Unlike mean and k-nearest neighbours, regression imputation imputes height based on other variables in the database. Patients with known age, weight and height was used to construct a 2^{nd} order regression polynomial model (Fig. 3.3) that was then used to impute the height in missing patient records. Age and weight are used in the regression model because there is a known relationship between these features (Wagstaff *et al.*, 2009). Regression imputation maintains the variability in the database but predicted values are all fitted to the regression model

with no consideration for patients who deviate from the model (Enders, 2010).

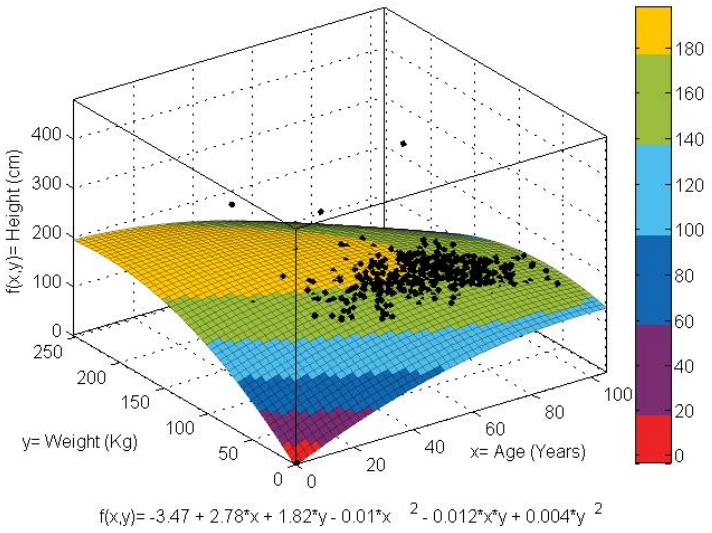


Figure 3.3: 2^{nd} order polynomial regression model (R^2= 0.983) fitted to patients of known weight, age and height (black points, n=536). The model is used to impute the missing height of patients in their health record

.

Imputation maintains the sample size but each method introduces its own level of uncertainty. To determine the best imputation method all three models (mean, k-nearest neighbor and regression) were used to impute the missing heights for the same set of patients with missing values. The distribution for the imputed height values from the three models was then compared to the height distribution of complete patient recordings as shown in Table 3.2. The regression model was chosen for this thesis since it best modeled the real patient population height mean and distribution. It is important to note that this preprocessing stage was only completed during the initial training of M4CVD. Once deployed M4CVD would prompt users to enter any missing data values.

Table 3.2: The regression model best matched the height distribution of the training data (bolded) when comparing the three imputation methods

Real Database Height Average	**Mean Imputation Height Average**	**K-nearest Neighbor (k=3) Mean Height Average**	**Regression Imputation Height Average**
170± 16 cm	170 cm	76 ± 11 cm	**168 ± 8 cm**

3.3 Feature Extraction

The next aspect of data processing was feature extraction to extract the time, heart rate variability and frequency features shown in Table 3.3. Feature extraction allows for the conversion of continuous physiological signals (e.g., ECG) into discrete values (e.g., heart rate) for classification. The health record data already contains the features of interest so additional feature extraction was unnecessary. Similarly, BP features do not require further feature extraction since the external BP monitor already records the systolic and diastolic BPs of interest. The focus of the remainder of this section is on ECG feature extraction techniques.

Table 3.3: Features extracted from wearable sensor physiological data

Sensor	**Time Domain**	**Frequency Domain**
ECG	P-QRS-T wave amplitude, PR/ QRS/ TQ Interval (Hampton, 2013) Heart Rate Variability, Heart Rate, mean R-R interval (Camm *et al.*, 1996)	Spectral Energy, Power spectral energy varianceNN, Low Frequency Power High Frequency Power, Power Ratio (Camm *et al.*, 1996)
BP	Diastolic Blood Pressure Systolic Blood Pressure	

3.3.1 Time Domain Feature Extraction

The electrocardiogram signal (Fig. 3.4) in the time domain depicts the cardiac activity of the heart. The key ECG waves are the P, Q, R, S and T waves representing the depolarization of the heart as it pumps blood throughout the body. The key ECG wave intervals are the

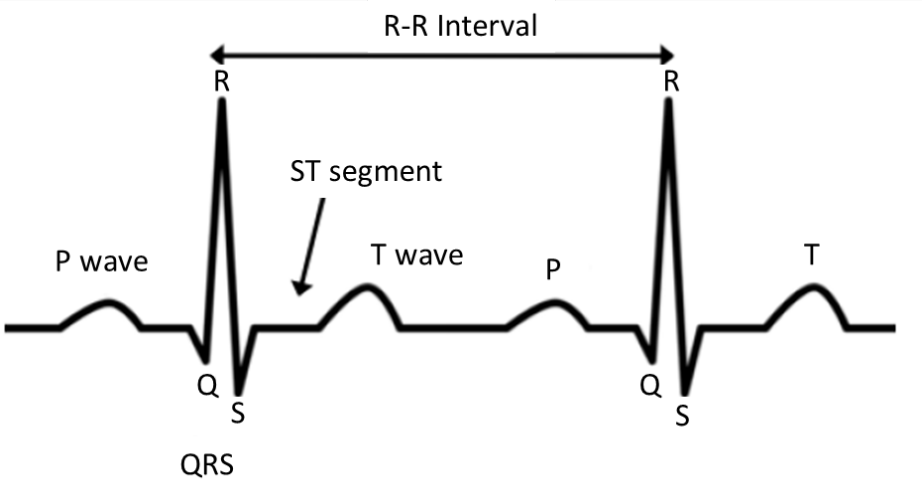


Figure 3.4: A labeled electrocardiogram sample (Sun *et al.*, 2012)

R-R, P-R, Q-T, S-T and T-P intervals. Both features have well known amplitudes and durations (Section 3.4). Any deviations outside the healthy ECG ranges are then used as various anatomical and clinical indicators. The ECG is a common and useful diagnostic tool used by health professionals who manually inspect ECG recordings. Automatic time feature extractions is a popular research area with numerous automatic detection methods proposed in the literature (Martis *et al.*, 2014). Figure 3.5 outlines the feature extraction stage used in the thesis. Time domain features provides a simple analysis tool to evaluate the heart function.

3.3.2 Heart Rate Variability

Heart Rate Variability encompasses the time domain markers that are related to the variations between consecutive QRS peaks. The interval between QRS peaks is known as the R-R or normal-to-normal (N-N) intervals that result from sinus node depolarization (Camm *et al.*, 1996). Changes in HRV have been shown to demonstrate patient deterioration before appreciable changes in heart rate occur (Horn and Lee, 1965). HRV has also been shown to

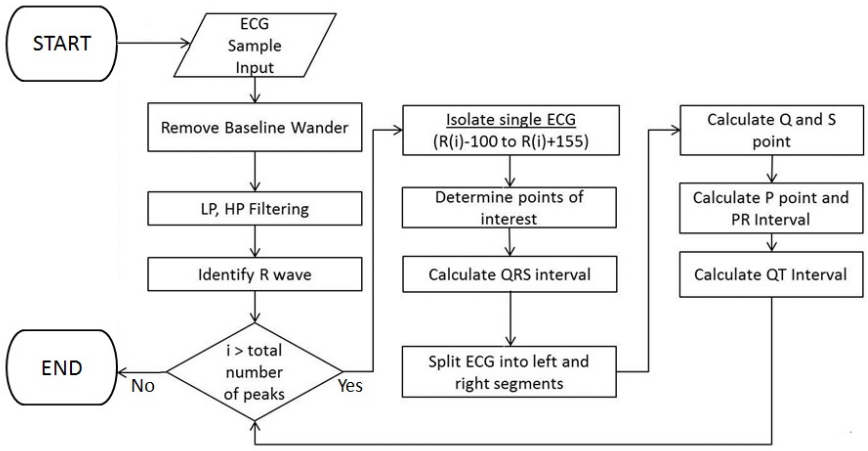


Figure 3.5: ECG time feature extraction procedure

be a strong predictor of mortality following an acute myocardial infarction (Kleiger *et al.*, 1987). A reduction of HRV has been reported as a result of multiple types of cardiological diseases (Camm *et al.*, 1996). HRV has the potential to provide additional insights into physiological and pathological conditions to enhance risk assessment. HRV features are computationally efficient to calculate since they are based on the R peak which is the easiest wave to detect from an ECG recording. Additional features can be calculated from the N-N intervals such as the SDNN (standard deviation of the N-N interval), rMSSD (square root of the mean squared differences of successive NN intervals), NN50 (Number of successive N-N interval greater than 50 ms) and pNN50 (percentage of successive NN interval greater than 50 ms). HRV features can be calculated from either 5 minute or 24 hour long recordings (Camm *et al.*, 1996).

3.3.3 Frequency Feature Extraction

The time domain monitoring of the P-QRS-T complex does not provide a complete picture on the heart's function. An ECG signal can also be converted into the frequency domain

for further analysis. The Fourier Transform represents the ECG as a sum of sine waves of different frequencies. A frequency spectrum was then used to show the ECG distribution over the frequency range. Frequency analysis has been shown to provide insights regarding the autonomic modulations of the patient's heart. However, frequency domain features are more computationally expensive to calculate compared to time domain features. Fig. 3.6 shows the normalized frequency domain of an ECG signal with the low frequency (LF) spectrum (0.04-0.15) and high frequency (HF) spectrum (0.15-0.4) highlighted in green and red respectively (Camm *et al.*, 1996). These frequency spectra can be extracted from recordings of 2-5 minutes in length (Saykrs, 1973). Additional features such as the power of the LF and HF bands as well as the ratio of LF/HF powers can also be calculated. The distribution of the power and central frequencies of the LF and HF may vary in relation to changes in autonomic modulations of the heart period (Malliani *et al.*, 1991).

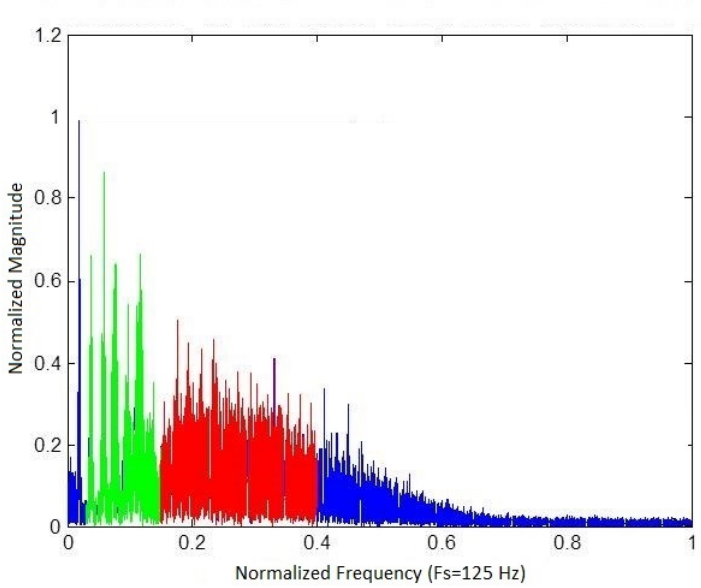


Figure 3.6: Frequency plot of ECG signal (blue) with DC removed. Low frequency (green) and high frequency (red) are highlighted. ECG was sampled at 125 Hz

3.4 Feature Normalization and Discretization

Each feature extracted from wearable sensors and health records have their own numeric ranges. The different numeric ranges of each feature presents a challenge when training MLAs. Features with larger physiological ranges may be assigned more weight over smaller feature ranges regardless of the importance of the feature to classification accuracy (Graf *et al.*, 2003). The first method to eliminate this range bias was to normalize all the features to a range of (0,1). This thesis also investigates the effect of two other feature inputs formats on classifier accuracy: Discretization and One Hot Coding.

The input features can be discretized into categories (e.g., 1,2,3,4) corresponding to each feature's range of healthy and unhealthy values. For example, resting heart beat has a wide physiological range between 60-100 beats/min. Heart rate can be discretized by making a category where the patient's HR is either within (HR = 1) or outside (HR = 0) the healthy range. On the other hand, BMI has a smaller range of 16-50 kg/m^2. BMI can also be discretized corresponding to low, medium, high and highest risk (e.g., 1,2,3,4). Features with known healthy and unhealthy ranges set by The Canadian Heart and Stroke Foundation (Heart and Stroke, 2013) are used to discretize the input features as shown in Table 3.4. These discrete categories provides the opportunity to incorporate medical knowledge directly into the proposed model. These discretized features may be a more useful indicator towards calculating overall heart disease severity. Features without known healthy and unhealthy ranges (e.g., age and HRV) were normalized.

Another method to represent the discretized inputs was using "one hot codes". One hot coding is a method used in machine learning for representing discrete categories by assigning each category a unique binary code (e.g., BMI is now 001, 010, 011, 100). One hot coding allows the MLA to be more specific when evaluating each features importance to the classification problem. For example, one hot codes allow for the evaluation of how the

low, medium, high and highest BMI category individually contribute to severity estimation. With one hot coding this thesis can report on whether having medium BMI is an important feature for CVD severity analysis instead of just BMI in general. However, each digit of the one hot code is considered a separate input to the MLA which increases the classifier's higher dimensional feature space compared to the normalization and discretize methods.

Table 3.4: Discretization and normalization of the extracted features

Feature	Wearable Sensor?	Clinical Database?	Physiological Range	MLA Normalization
Age	N	Y	Continuous	Continuous
Gender	N	Y	Male Female	0 1
Body Mass Index	N	Y	Normal <24 kg/m^2 Overweight 25-29.9 kg/m^2 Obese I 30-39.9 kg/m^2 Obese II >40 kg/m^2	1 2 3 4
Systolic Blood Pressure	Y	Y	Low Risk <120 mmHg Medium Risk 121-139 mmHg High Risk >140 mmHg	0 1 2
Diastolic Blood Pressure	Y	Y	Low Risk <80 mmHg Medium Risk 80-89 mmHg High Risk >90 mmHg	0 1 2
Heart Rate	Y	Y	Normal 60-100 beats/min Abnormal Other	0 1
QRS Inteval	Y	N	Normal <0.12 s Abnormal Other	0 1
PR Interval	Y	N	Normal, 0.12-0.2 s Abnormal Other	0 1
Q Wave Interval	Y	N	Normal <0.04 s Abnormal >0.04 s	0 1
P wave amplitude	Y	N	Normal <3 mV Abnormal >3 mV	0 1
R-R Interval	Y	N	Normal 0.4-1.5 s Abnormal >1.5s	0 1
Heart Rate Varability	Y	N	Continuous	Continuous
ECG Low and High Freq Power	Y	N	Continuous	Continuous

3.5 Machine Learning Algorithms

The next step as shown in Fig. 3.1 was the design and training of the machine learning algorithm for the computing of a decision regarding patient risk. In this thesis the monitoring system was presented with a classification problem to classify a patient's CVD severity level as low or high. A MLA offers several advantages in this classification problem compared to traditional programming methods. First, it is difficult to directly program the algorithm that can assemble the collected data into a coherent assessment of patient risk. Second, it's impossible to directly code a solution since the programmer cannot implicitly program all possible scenarios. Finally, a traditional program is static once complete so the programmer would have to completely rewrite and debug the program to incorporate any new rules in the future. On the other hand, a MLA is dynamic so programmers can easily retrain the MLA to incorporate new training examples without having to rewrite the entire program.

Two MLAs that are used for classification problems are 1) the support vector machine and 2) the multi-layer perceptron which are discussed in Sections 3.5.1 and 3.5.2 respectively. The SVM and MLP were investigated in this thesis because they both attempt to solve the classification problem through the calculation of a decision boundary that separates the two classes of data. However, they approach the same classification problem using two different architectures, training procedures and parameters. As a result, the SVM and MLP have different accuracy and complexity performances. In this thesis a comparative analysis was conducted between the SVM and MLP for deployment in a RPM system on a mobile device.

3.5.1 Support Vector Machines

The support vector machine (Cortes and Vapnik, 1995), defined in Eq. 3.4, creates a classification boundary (hypersurface) to separate input vector $\mathbf{x}_i$ into one of two classes $\mathbf{y}_i \ \epsilon \ \{0, 1\}$. The optimal SVM finds the $\mathbf{w}$ and b that minimize the empirical error (risk) and

the complexity of the classifier, a trade-off achieved using the structural risk minimization (SRM) principle to minimize the upper bound of the generalization error (Vapnik and Chervonenkis, 1974). In the training stage, the SVM is optimized using a set of input vectors $\mathbf{x}_i$ with known class labels $\mathbf{y}_i$. In the testing stage, new data vectors with unknown classes are then classified.

$$y_i(\mathbf{w}^T\phi(x_i)+b) \geq 1-\xi_i \tag{3.4}$$
$$\xi_i \geq 0$$

$$\text{such that } \underset{\mathbf{w},b,\xi}{min}\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}\xi \tag{3.5}$$

$$k(\mathbf{x},\mathbf{x_i}) = \phi(\mathbf{x_i}^T)\phi(\mathbf{x_j}) = \exp(-\frac{1}{2\sigma^2}\|\mathbf{x}-\mathbf{x_i}\|^2) \tag{3.6}$$

The Vapnik–Chervonenkis (VC) theory (Vapnik and Chervonenkis, 1974) allows for the control of the classifier complexity in Eq. 3.5 by minimizing the classifier margin,$\|\mathbf{w}\|$, between the class boundary and the dividing hypersurface with bias b. The slack variable ξ and the penalty constant C allows for some feature points to appear on the wrong side of the soft margin of the hypersurface. As $C \to 0$ more points are allowed to fall on the wrong side of the hypersurface which increases generalization at the expense of training accuracy. When two classes are not linearly separable in 2D space a kernel ϕ, defined in Eq. 3.6, is used to map $\mathbf{x}_i$ to a higher dimension features space $\phi(x_i)$ as shown in Fig. 3.7. A non-linear separating hypersurface is then constructed in the feature space. A popular kernel for the SVM is the radial basis function (RBF) (Haykin, 2009; Tao, 1993) and is used when the number of training examples (i) is greater than the number of features ($\mathbf{x}$) as is the case with the data in this thesis. The SVM and the kernel have two parameters that require optimization by the user: the penalty constant C, and the kernel width σ^2 (or $\gamma = 1/(\sigma^2)$). The parameter selection is a balance between the SVM model's training accuracy and generalization. K-Fold cross validation is a procedure to determine the optimized C and

σ^2 by dividing the input data into K folds, training the SVM on each K-1 fold and testing on K^{th} fold. A grid search of each unique combination of C and σ^2 is then conducted to identify the parameters that provide the best cross-validated classification accuracy. Once the SVM has been trained it can be applied to classify new data.

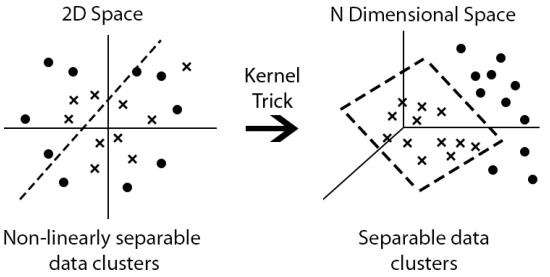


Figure 3.7: SVM Kernel Trick allows for the classification of non-linearly separable data points by mapping the data to a higher dimensional space

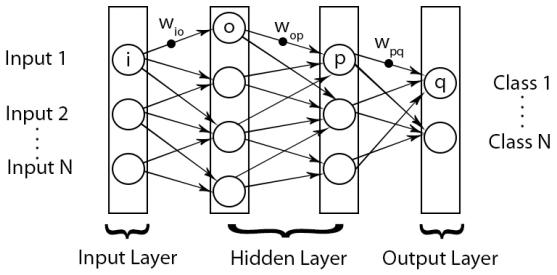


Figure 3.8: Multilayer Perceptron Architecture

3.5.2 Multilayer perceptron

The MLP is one of the most popular MLAs used for decision support systems in literature (Andreu-Perez *et al.*, 2015). Similar to the SVM, the optimal MLP focuses on determining optimal weights and bias but instead follows the empirical risk minimization (ERM) principle which minimizes error on the training data (Vapnik and Vapnik, 1998). The base unit in a MLP is the perception which is a mathematical model inspired by human neurons (Rosenblatt, 1961). Each perception has an input vector $\mathbf{x}_i$, weights $\mathbf{w}$ and bias, b, fed into a summing junction. The summing junction feeds into a transfer function which implements an all-or-nothing (0,1) response similar to biological neurons. When training a perception the objective is to adjust the weights and bias to generate the correct output class labels $\mathbf{y}_i$ when a certain input is applied. The weights and bias are initially chosen randomly. Similar to the SVM the perception is a supervised learning technique since the training data with known inputs and classes is used when adjusting the weights/ bias. Fig. 3.8 shows the architecture of the MLP where perceptrons (circles) are combined into

a layer. Multiple layers are then combined to form the MLP. Each perceptron in one layer is connected to each perceptron in the next layer via a weight. A MLP is composed of one input layer for the input features and one output layer consisting of N classes. The N layers between the input and output layer are called hidden layers which creates a feature space allowing non-linear classifiers which greatly increases the interpretative capability to enable the MLP to solve non-linear problems.

When training a MLP an input vector $\mathbf{x}_i$ is applied to the first layer and propagates through the network to generate an output vector $\mathbf{y}_i$. The error function E, defined in Eq. 3.7, between the MLP estimated output $\mathbf{y}_i$ and the desired output $\mathbf{d}_i$ is calculated per propagation step k. Back-propagation training (Rumelhart *et al.*, 1985) is a process to update the weights such that the update minimizes the network's overall classification error on each step. The process calculates the gradient vector p for each weight that indicates how to modify the weight per step. The gradients are calculated by working backwards from the output layer towards the input layer. The weights are adjusted on each step k using Eq. 3.8 where γ is the learning rate (how large the change on each step is) and $p(k)$ is the gradient vector for step k. Other methods for calculating the weight gradient are used in literature (Haykin, 2009) and in this thesis the Levenberg-Marquardt method (Hagan and Menhaj, 1994) was found to have the best results.

$$E(w) = \frac{1}{2}\sum_{i=1}^{N} \|y(x_i, w) - d_i\|^2 \qquad (3.7) \qquad\qquad w(k+1) = w(k) + \gamma p(k) \qquad (3.8)$$

In addition to optimizing the weights there are architecture parameters and hyper-parameters that need to be configured by the user. Architecture parameters include the number of perceptions per layer, number of hidden layers and training function. Hyper-parameters are function specific parameters such as the learning rate or momentum values (Haykin, 2009). Once these parameters have been selected the dataset is divided into training, validation and test sets. The training set is used to optimize the weights of the

MLP using back-propagation. To prevent over-fitting (memorization) the accuracy of the MLP when classifying data from the validation set is tracked rather than the training set accuracy. Training stops when error on the validation set fails to improve over a certain number of steps. The test set measures the system accuracy in labeling new data on the best trained MLP. Unlike the SVM, the MLP requires a more trial and error approach to optimize the architecture and hyper-parameters. Once the MLP has been trained it can be classified on new data. Interested readers are referred to (Müller *et al.*, 2001) and (LeCun *et al.*, 2015) for a more detailed explanation on the SVM and MLP respectively.

3.5.3 Comparison between SVM and MLP

Both the SVM and MLP are popular algorithms for classification problems due to their ability to map features to higher dimensional space. The SVM maps to higher dimensional space through the kernel function while the MLP uses hidden layers (Andreu-Perez *et al.*, 2015). Both algorithms are examples of supervised learning since the training data has labeled classes that is used when adjusting the weights/ bias. Similar to the SVM, the optimal MLP focuses on determining optimal weights and bias but instead follows the empirical risk minimization principle which minimizes error on the training data (Vapnik and Vapnik, 1998).

One key difference of the MLP is the optimization solution may have multiple local minima which can prevent locating the optimal classifier. In contrast, the SVM optimization solution is global and unique (Zanaty, 2012). The SVM has a simple geometric interpretation and gives a sparse solution and was chosen because of its ability to construct a robust classifier from only a small data set. In addition the SVM has good generalization properties allowing it to classify new data. However, the MLP can be more easily implement multi-classifications by modifying the number of units in the output layer.

The input data from the wearable devices and health records are fed into the deployed

mobile MLA to classify the patient's severity level as low or high. The SVM performs classification by creating a higher dimensional hyperspace and comparing the location of the new input data to the previously trained hypersurface via the support vectors. In contrast, the neural network applies the weights and bias calculated during training on the new input data.

3.6 System Output

The last stage of M4CVD was the communication of the system's results to the patient and/or the healthcare professional. The output stage was not addressed in this thesis however some recommendations for this module are discussed here. It is important that both the classification result and the collected health data is presented to the users for them to make their own assessments. Unlike current systems that transmits continuously, M4CVD can provide different levels of communication to the patient and healthcare professional depending on the calculated severity level. For example, a classification of a low severity level will only result in notification to the patient providing continuous feedback between appointments. On the other hand, a high risk level classification could result in M4CVD sending an alert to a medical professional. A high risk severity level could also cause M4CVD to enter a confirmation stage that reevaluates the patient's risk level at a high sampling rate. A high severity classification result would be more computationally expensive compared to low severity classification. However, the system still uses less resources compared to current RPM systems that continuously transmit to a remote server.

Chapter 4

System Implementation

This chapter outlines the implementation of the algorithm components. The implementation environment is described in Section 4.1. Section 4.2 introduces and validates the ECG and BP sensors used in the deployed mobile system on healthy patients. In Section 4.3 the medical database used for the training and testing set of M4CVD is presented. Data processing and machine learning implementation on the training database are discussed in Section 4.4 and 4.5 respectively. Finally, the experimental procedure is outlined in Section 4.6 for evaluating M4CVD's accuracy and hardware performance.

4.1 Implementation Environment

In this thesis the development and deployment of M4CVD was done on different target hardware as shown in Fig 4.1. The cross-validation training of the MLAs was computationally expensive but was not included in the deployed MLA. As a result, system development was completed on a more powerful computer before deploying the MLA on the mobile device. System development was done on a 64-bit Windows 7 laptop with a 2.2 GHz Intel i7 CPU and 12 GB RAM. MATLAB 2014a was used for the training and testing of M4CVD.

MATLAB has the tools for the rapid development of machine learning models and data visualization. The best M4CVD models were then tested on a mobile device. The system was deployed in C++ on a Linux Raspberry Pi 2 Model B (RASPI) single board computer with a quad-core ARMv7 900 MHz processor, 1 GB RAM and a Broadcom VideoCore IV GPU. Table 4.1 shows that the performance of the Raspberry Pi 2 was similar to two popular smartphones: the Apple iPhone 5S and the low-cost Motorola Moto G. The Raspberry Pi was chosen because it can host many interchangeable circuits (called shields) which allows for rapid prototyping of systems. All algorithm blocks of M4CVD were implemented on both the PC and RASPI platforms.

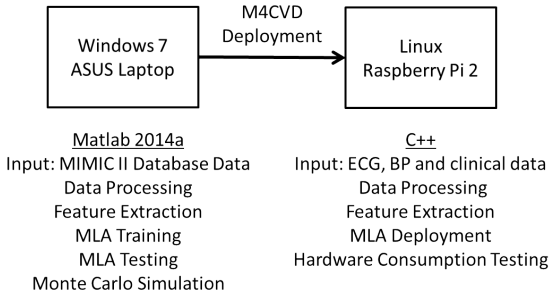


Figure 4.1: M4CVD was trained and tested on a laptop to assess classifier accuracy and deployed on a Raspberry Pi 2 to evaluate the system's resource requirements

Table 4.1: Comparing the Raspberry Pi 2 to two smartphones

	Raspberry Pi 2 (2015)	**Apple iPhone 5S (2013)**	**Motorola Moto G (2014)**
CPU	ARM Cortex-A7	Apple A7	ARM Cortex- A8
Processor Speed	900 MHz	1.3 GHz	1.2 GHz
Cores	4	2	4
RAM	1 GB	1 GB	1 GB

4.2 Sensors

Wearable devices such as the Fitbit have great potential for RPM but face many challenges regarding clinical accuracy, battery life and comfort levels (Patel *et al.*, 2010). Consumer wearable sensors (e.g., Apple Watch) are not yet sufficient for vital sign monitoring and in this thesis a research grade acquisition system was used for testing M4CVD. Specifically

the Libelium electronic health (eHealth) sensor platform was used for ECG and BP acquisition. The signal acquisition shield can interface with the RASPI to record many different physiological signals such as ECG, BP, GSR and temperature providing a large number of sensors to study.

4.2.1 ECG Validation

A standard three-lead (Lead I, II, III) configuration was used to record ECG signals (Fig. 4.2). ECG signals were acquired, filtered and amplified by the conditioning circuit shown in Appendix B.1. The ECG was then stored on the RASPI for further processing and analysis. The RASPI also powers the ECG acquisition shield as shown in Appendix B.2. Measuring the current consumption of the Pi allows for the evaluation of the current consumption of the Pi as well as the ECG shield.

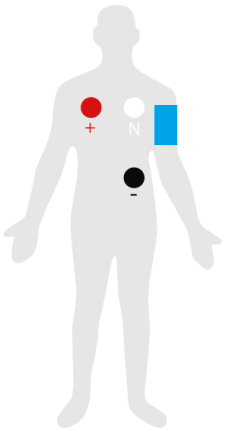


Figure 4.2: M4CVD sensor configuration for ECG (+,-,N) and BP (blue)

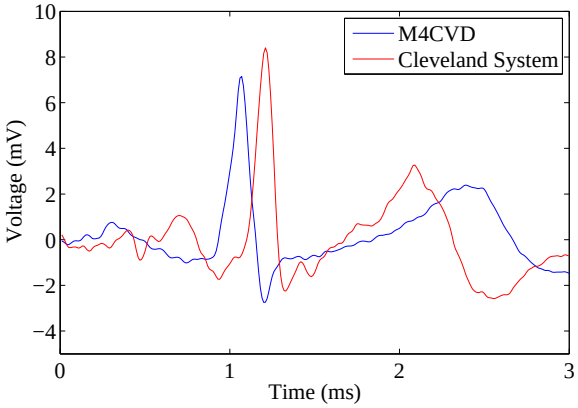


Figure 4.3: An ECG recording from the same patient recorded by the Libelium sensor shield and Cleveland BioRadio

A CleveLabs BioRadio, a medical grade ECG recorder, was used to validate the Libelium ECG sensor used in this thesis. A 24 year old male volunteer had his ECG recorded by both systems with the same electrodes, configuration and sample frequency (Fs=600 Hz).

The ECG signals were then preprocessed using the procedure discussed in Chapter 3 using a butterworth bandpass filter (0.05-50 Hz). A visual inspection of the two ECG recordings (Fig. 4.3) shows similar P, QRS and T waves with a time offset between the signals. Overall, the signal properties from the two systems match and validates the ECG sensor.

4.2.2 Blood Pressure Validation

The Libelium eHealth sensor platform includes a Kodea KD-202F BP monitor that was used in this thesis to measure BP. The pressure cuff was placed around the patients bicep (Fig. 4.2) and measures systolic and diastolic blood pressure and pulse. The monitor can store up to 80 measurements including their timestamps. The Kodea monitor was chosen because the BP values can be automatically uploaded to the RASPI for further analysis by the Libelium eHealth sensor platform (Appendix B.3).

A second commercial grade BP monitor (Life Source UA-787) was used to validate the Kodeo system. A 59 year old volunteer had his BP recorded by both systems. BP was recorded with the volunteer in a stationary sitting position who rested 5 minutes before taking the recordings. Table 4.2 shows the recorded values by both systems and they are consistent with each other (within ± 5 units) with five readings taking over three days.

Table 4.2: Comparison of Kodea and Life Source BP monitor

Date **Y/M/D/T**	**Kodea** **Sys-Dia-Pulse** **mmHg-mmHg-/min**	**Life Source** **Sys-Dia-Pulse** **mmHg-mmHg-/min**
15/10/22 12:55 PM	129-79-67	124-78-66
15/10/23 07:24 PM	123-75-70	111-73-66
15/10/25 04:47 PM	132-88-59	144-92-60
15/10/25 07:23 PM	111-77-66	115-71-63
15/10/25 08:05 PM	120-72-64	102-70-65
Mean	**123-78-65**	**119-76-64**
Standard Deviation	**8.2-6-4**	**15-9-2.5**

4.3 Database and Patient Selection

The requirements for the training database are: 1) Contain patients with heart disease 2) Contain both clinical and physiological data 3) Both low and high risk classes are represented and 4) Sufficient population size. PhysioNet's (Goldberger *et al.*, 2000) directory of medical databases had three databases that met the criteria (Table 4.3). The MIMIC II (Multiparameter Intelligent Monitoring in Intensive Care II) database was chosen due to its large population and the detailed records available for each patient (Saeed *et al.*, 2011).

Table 4.3: A comparison between three medical databases considered for M4CVD

	MIMIC II: Multiparameter Intelligent Monitoring in Intensive Care Database II (Saeed *et al.*, 2011)	**SHAREE: Smart Health for Assessing the Risk of Events via ECG (Melillo *et al.*, 2015b)**	**PTB: Physikalisch-Technische Bundesanstalt Diagnostic ECG Database (Bousseljot *et al.*, 1995)**
Number of Patients	33,000 (1000 used)	139	290
Database Breakdown	-ECG -BP -Clinical Data -ICU Records	-ECG -BP -Clinical Data	-ECG -BP -Respiration -Clinical Data
Pros	-Most detailed patient records -Only database with continuous BP -Potential for the largest training set	-Contains "low risk" and "high risks" labels -All patients have the same hypertensive disease	-Large Number of CVD Patients -Includes healthy patients -Variety of different CVD represented
Cons	-Patients have other conditions besides CVD -Few pure CVD patients -No healthy patients -Manual Labelling required	-Small database size -Small number of "high risk" patients	-Diagnostic Database -No "low risk" vs "high risk"

MIMIC II is a publicly available database containing both a waveform and clinical database of deidentified ICU patients collected from 2001-2008 from three hospitals in Boston. The waveform database contains physiological recordings on ECG, BP and respiration while the clinical database contains detailed patient records. The physiological recordings were obtained from bed-side monitors with each patient record typically containing hundreds of individual files. The clinical records were retrieved from hospital databases and digitized nurses notes. MIMIC II can be accessed online after completing a license agreement using a structured queried language (SQL).

Patients with heart disease were identified in the database as those whose primary International Classification of Diseases (ICD-9) code was between 390-459 (WHO *et al.*,

1978). Of the 1000 patients identified in MIMIC II who have CVD, 839 of them had physiological recordings matched to their corresponding clinical records and served as the initial dataset in this thesis. The statistical breakdown of the patients used in the final training set is discussed in Chapter 5.

4.3.1 Class Labeling

In supervised learning, the MLA requires the dataset to have prelabeled classes (e.g., low or high risk) before the classifier can be trained. The MIMIC II database is a general purpose repository of physiological and patient records that does not include a label regarding a patient's CVD risk. So before training M4CVD each patient's CVD risk level must be labeled as low or high. 839 patients was too time consuming for a health professional to label manually so two automated methods were investigated for classifying a patient's risk level: 1) Simplified Acute Physiology Score I (SAPS) and 2) Diagnosis Related Group (DRG) weight. The two metrics are calculated at the hospital by health professionals and are included in the MIMIC II database.

The DRG weight is a payment classification system used in the United States that measures the relative amount of resources that the hospital used to treat the patient (Averill *et al.*, 2003). A single DRG code was calculated for a patient's stay in the hospital. The DRG code was based on the following: the patient's ICD-9 diagnosis code, procedure code, gender and age. The DRG code was calculated by a coding expert at the hospital and a weight of 1 indicates the patient used an average amount of hospital resources expected for their condition. The DRG code is used as an indicator for a patient's illness severity since a more ill patient will consume more hospital resources.

SAPS is an ICU scoring system designed to measure the severity of a disease (Le Gall *et al.*, 1984). A patient's SAPS score corresponds to predicted mortality between 0-100% and was calculated based on 17 physiological signs recorded in the first 24 hours of the

patient's admittance to the ICU. A patient's SAPS score was calculated on every admittance to the ICU and a patient may receive multiple SAPS scores within a single hospital visit.

The method for the automatic prioritisation of patients for treatment in the ICU proposed by Gattinoni *et al.* (2004) and Iapichino *et al.* (2006) was used in this thesis. First, both the median DRG and median SAPS score were calculated for the patients in training dataset. The high risk patients were then defined as those whose DRG/SAPS score was above the calculated median value for each label. Table 4.4 shows the patient distribution per class and M4CVD was trained using both labeling systems. It is important to note that neither metric will be used as a input feature (only output classification) since these two metrics are not available outside of clinical settings.

Table 4.4: Class breakdown for DRG and SAPS I using a two class labeling system of the final 545 patients used for training M4CVD

Metric	**Low Risk**	**High Risk**
DRG	266 DRG <=3.11	279 DRG>3.11
SAPS I	243 SAPS<=13	286 SAPS>13

4.3.2 Database Limitations

The MIMIC II database has several advantages as it includes a large patient database and a combination of physiological and medical records. However, using a third party database not designed specifically for this study introduces limitations regarding the features that can be studied. This study was restricted to features that can be extracted from the physiological recordings and clinical data found in the MIMIC II database. For example, GSR was initially investigated as a potential sensor input to M4CVD to monitor patient's stress levels. Since the MIMIC II database does not include GSR recordings this feature

was subsequently removed from consideration. Similarly the database medical records did not consistently include information on patient habits (e.g. smoking and exercise) and thus could not be examined in this thesis.

The main technical restriction of the MIMIC II database is the quality of the stored ECG recordings which limits the type of ECG features that can be extracted. The MIMIC II documentation (Saeed *et al.*, 2011) outlines how ECG monitors with different recording settings and sampling rates were used to monitor patients. As a result, all ECG signals stored in MIMIC II were standardized by the MIMIC II designers to Fs=125 Hz using peak-picking techniques to limit the size of the ECG recordings in the database. All reconstructed ECG assume an average constant interval of 8 ms between each sample when the interval can be between 2-14 ms. The resulting decimated ECG had reduced time and amplitude resolution and can be interpreted visually. However, the frequency-domain features (such power ratio) were compromised and so these features may not be suitable for inclusion in the final design. In the next section the impact the standardization has on the P-QRS-T peak detection and quality assessment methods is investigated.

4.4 Data Processing Implementation

In this section the ECG preprocessing and feature extraction blocks introduced in Section 3.2.1 and 3.3 are implemented and validated. The data processing for the BP recordings are also discussed.

4.4.1 ECG Preprocessing Validation

In the MIMIC II database a patient may have multiple ECG recordings taken throughout their stay in the ICU which range in length from a few seconds to several hours. For each patient the longest ECG recording was selected for feature extraction with the average

recording length being 17 ± 10 hours. The current gold standard for ECG recording (Camm *et al.*, 1996) recommends restricting analysis to recording lengths of 5 mins or 24 hours. This thesis focused on analyzing 5 minute recordings since few patients had continuous 24 hour long recordings. Each patient's longest ECG recording was then subdivided into 5 minute segments for signal quality assessment and ECG feature extraction.

In preprocessing the first step was to design a 4^{th} order Butterworth bandpass filter (0.05-50 Hz) using the MATLAB signal processing toolbox to calculate the co-efficients used by the filter when deployed on the PC and the RASPI. Second, a low-order polynomial was fitted to each 5 minute ECG sample using MATLAB's polyfit function to remove baseline wandering. The clean signal was obtained by subtracting the calculated polynomial from the original signal. The filter and detrended was tested successfully on a random ECG signal as shown in Fig. 4.4.

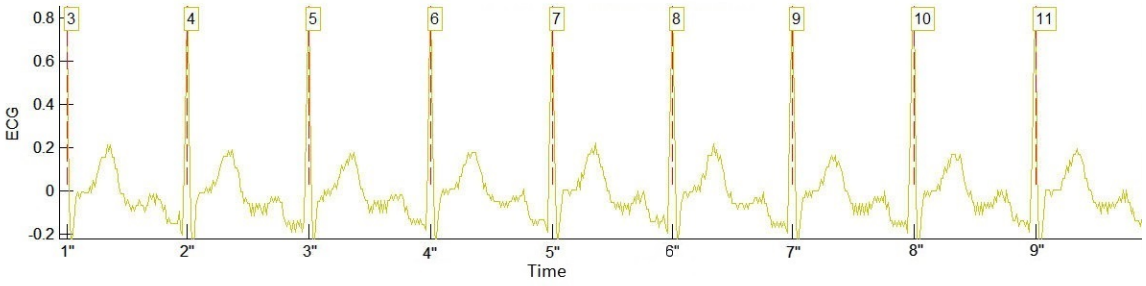


Figure 4.4: R peak detection using the pantom package (Pan and Tompkins, 1985)

Next, the signal quality metrics were evaluated on a random subset of 20 (12 good, 8 poor) ECG signals from the MIMIC II database. The kurtosis, spectral distribution ratio and the combined quality metric (kurtosis + SDR) were calculated. Each ECG signal then underwent automated R peak detection. The R peaks and ECG quality assessment were then visually validated with the results showing that the combined metric performed best overall with a success rate of 75% identifying good (9) and poor quality signals (6). The combined quality metric was also successful in identifying the three worst ECG signals

where automated R peak detection failed completely. While the combined metric had a high false negative rate (10%) this was considered acceptable to ensure only good quality signals were used for feature extraction. Each 5 minute ECG segment in the patient's recording underwent preprocessing and was assigned a quality label based on the combined metric. 677 patients had at least one 5 min ECG signal labeled as good quality. The first 5 minute segment of good quality per patient was selected for ECG feature extraction.

The last step in ECG preprocessing was to identify all the R peaks within a patient's 5 minute recording for use in feature extraction in the next section. The following open-source R peak detection libraries (all available from PhysioNet) were investigated: pantom (Pan and Tompkins, 1985), sqrs (Engelse and Zeelenberg, 1979), gqrs and wqrs. A random subset of 16 good quality ECG signals were selected for R peak detection using the four libraries. Visual inspection determined the best R peak detector was the pantom package (Fig. 4.4) which had a success rate of 93% and was used for R peak detection in M4CVD.

4.4.2 Blood Pressure Implementation

The MIMIC II database also contains the patient's numeric blood pressure readings that were taken simultaneously with the ECG signal. The MIMIC II databases includes both invasive continuous arterial blood pressure (sampled at once per second) and non-invasive blood pressure recorded by a cuff once per hour. Similar to the ECG recordings the BP readings cover a time period of seconds to several hours. In this thesis the highest (worst) non-invasive blood pressure that was recorded within the 24 hour window of the patient's 5 minute ECG sample was used for training M4CVD. No further preprocessing was required once the systolic and diastolic readings were extracted.

4.4.3 Feature Extraction Implementation

Both continuous and discrete features are extracted from sensors (ECG and BP) and health records. Continuous values include HRV features from the ECG sensor and patient's age from the clinical records. Discrete values are extracted from the BP monitor and BMI from the clinical records.

The automatic identification of the P-QRS-T waves used the open source ecgpuwave package which locates the onsets and ends of the P, QRS and T waves around each preidentified R peak (Laguna *et al.*, 1994). The ecgpuwave package was successfully validated by identifying the P-QRS-T waves on an ECG recording taken from the gold standard MIT-BIH arrhythmia database (Moody and Mark, 2001) as shown in Fig. 4.5. However, the ecgpuwave failed when tested on an ECG recording from the MIMIC II database (Fig. 4.6) with too many false wave identifications. Subsequently the calculation of the ECG wave intervals was unsuccessful. The failure of the ecgpuwave library may be due to more noise found in the ICU recordings. Most likely the library's failure was a result of the decimation of the MIMIC II ECG signals. It is outside the scope of this thesis to improve the automatic peak detection methods. As a result the time domain ECG features (expect R peak detection) were not included in the final feature set used for training the MLAs.

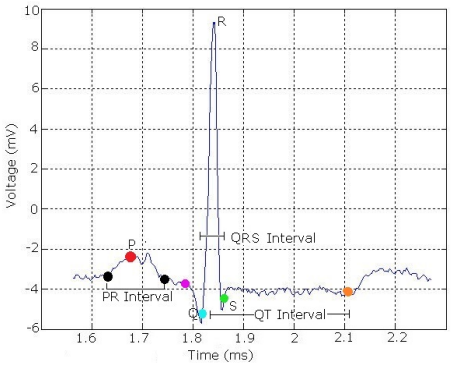


Figure 4.5: Wave peak detection successfully worked on MIT-BIH ECG

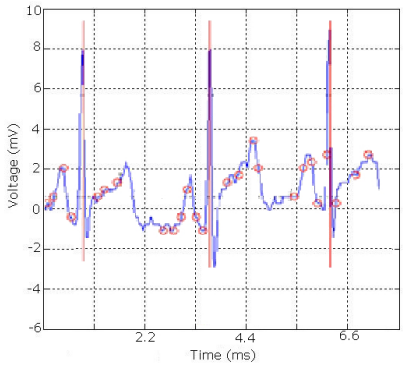


Figure 4.6: MIMIC II ECG peak detection had too high a rate of false positives

Since R peak detection was validated successfully the next step was to determine the patient's HR ($RR_{total}/5$) in beats/min and HRV features. Next, the intervals between each successive R peaks ($RR_{interval}$) and the patient's mean R-R interval (RR_{mean}) over the 5 minute recording were calculated. Eq. 4.1 - 4.2 shows the calculation of the SDNN and rMSSD features respectively. Finally, the patients HRV=$RR_{mean}/5$ and pNN50 (number of successive $RR_{interval} > 50$ ms) are determined. A manual inspection of the dataset patient's HR found 159 (23%) patients with HR less than 40 beats/ min which is not possible. Further study determined it was due to incorrect automatic R peak detection and these patients were also excluded from the study. In total, 518 patients labeled by their DRG value (502 patients with SAPS) were used for training M4CVD.

$$SDNN = \sqrt{\left(\frac{1}{a}\right) * \sum_{j=1}^{a} \left(RR_{Interval}(j+1) - RR_{mean}\right)^2} \tag{4.1}$$

$$rMSSD = \sqrt{\left(\frac{1}{a-1}\right) * \sum_{j=1}^{a} \left(RR_{Interval}(j+1) - RR_{Interval}(j)\right)^2} \tag{4.2}$$

a= total number of R-R intervals

4.5 Machine Learning Implementation

Twenty-four input features for monitoring CVD were identified from the literature (Appendix C). Eleven of these features (Table 4.5) were successfully validated for further study. These 11 features were converted into normalized, discretized and one-hot coding formats before being used as inputs for training the SVM and MLP.

All 518 DRG labeled patients (502 patients labeled with SAPS) with 11 features each were used to train and test the SVM and MLP. The LibSVM machine learning library (Chang and Lin, 2011) was used to implement the SVM. LibSVM was used because the library is

Table 4.5: The 11 Features from ECG and BP sensors and health records used for training the MLA. C=continuous features, D=discrete feature

Clinical Data	**Blood Pressure Sensor**	**ECG Sensor**
1. Gender (D)	4. Systolic Blood Pressure (D)	6. Heart Rate (D)
2. Age (C)	5. Diastolic Blood Pressure (D)	7. Mean R-R Interval (C)
3. BMI (D)		8. Heart Rate Variability (C)
		9. Standard Deviation of R-R (C)
		10. Square Root of Mean Difference of R-R (C)
		11. Percentage of R-R interval greater than 50 ms (C)

written in C++ allowing the SVM to be trained on the PC using MATLAB and then deployed on the RASPI for evaluation. Similarly, the MATLAB neural network toolbox was used to implement the MLP.

The SVM and MLP internal parameters (weights and bias) are optimized automatically during the training process. However, the best hyper-parameters for each algorithm must still be selected manually in order to produce the best classifier. For the SVM, the kernel and its associated parameters for mapping input data into higher dimensional space must be selected manually. For the MLP the learning function, number of hidden layers and number of neurons per layer are also manually selected. Table 4.6 shows the SVM kernel and MLP learning functions examined in this thesis.

Table 4.6: List of Investigated SVM Kernel and MLP Learning Functions

SVM Kernel	**MLP Learning Function**
Linear	Scaled Conjugate Gradient Descent
Polynomial (degree 2-5)	Levenberg-Marquardt Backpropagation
Radial Basis Function	Gradient Descent with Momentum
Sigmoid	

To determine the best classifier configuration 10-fold cross-validation was used to train multiple SVM and MLP under different configurations. For each classifier configuration, ten MLAs were generated using a random subset of patients as the testing set (30%) and the remaining patients as the training set (70%) to reduce bias. For the MLP the dataset was divided into 80% training and 20% testing sets with 25% of the training data was used as the validation set. The average MLA's accuracy was obtained by taking the average accuracy over each unique MLA for a given classifier combination. A broad cross-validation over several magnitudes was run first followed by a fine grid search focusing on the most promising configuration. The pair that provided the best average classification accuracy was chosen.

Fig. 4.7 shows a SVM cross-validation over a coarse grid search of C and gamma parameters. The best DRG SVM had an accuracy of 61.29%. Increasing the SVM parameter C initially improves generalization as it increases the system's misclassification penalty. However, increasing C too much over penalizes the system and reduces the systems accuracy. The graph in Fig. 4.7 shows the trade-off between optimizing accuracy and generalization. Another interesting property in Fig. 4.7 is the ridge of different (C, gamma) combinations that result in the same accuracy. The ridge of (C, gamma) values is commonly found in cross validation when the change in C are offset by a subsequent change in gamma that maintains the algorithm's overall accuracy. Compared to the SVM, the MLP architecture optimization is more heuristic compared to the SVM and alternative procedures have been proposed to improve parameter selection such as grid and random search (Bergstra and Bengio, 2012). Fig. 4.7 shows the CV for selecting the best number of neurons in a single hidden layer neural network. As the number of neurons increases the accuracy of the network overall slightly improves with the best DRG MLP achieving a 64% accuracy. However as the number of neurons increases the average training time for the network increases

dramatically unlike the SVM which has consistent training time between classifier configurations. It is interesting that networks with fewer neurons (less training time) achieved similar accuracy to larger networks indicating that more neurons were not helping the MLP differentiate between the two patient classes.

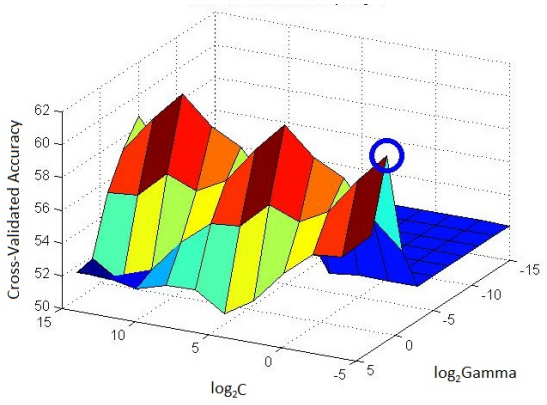


Figure 4.7: SVM Cross-Validation of Polynomial (deg=2) Kernel. Best accuracy circled

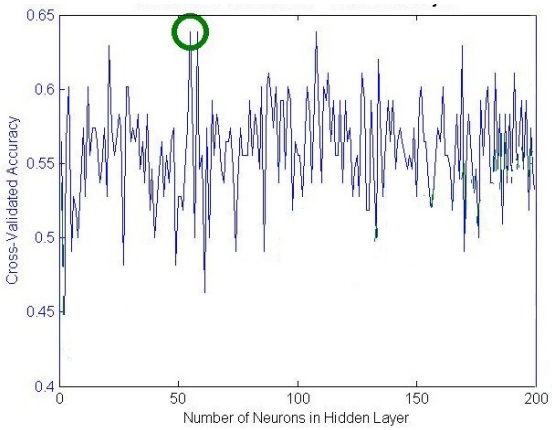


Figure 4.8: MLP Cross-Validation to determine the best number of neurons in the hidden layer. Best accuracy circled

4.6 Procedure for Evaluating System Performance

In this section the experimental procedure used to evaluate M4CVD is described. First, the training database population distribution is described. CV training is then conducted to identify the best MLA configurations. The best SVM and MLP algorithms were then investigated to determine their average performance by running a Monte Carlo Simulation. Finally, the developed system was deployed onto the RASPI to evaluate M4CVD's hardware performance and resource consumption.

A Monte Carlo simulation tests a MLA with the same parameters and hyper-parameters multiple times using different subsets of patients for the training and testing sets in each run. No patient record was used in both the training and testing set during the same

run. Running a Monte Carlo simulation helps limit bias in the results caused by a random selection of a training/ testing sample set that disproportionately produces a high result. The best SVM and MLP were trained 1000 times and the behavior of each classifier was described using average accuracy, sensitivity and specificity. This thesis also compared the SVM and MLP's receiver operator curve (ROC) which is a plot representing the performance of a binary classifier system. The ROC plots sensitivity vs (1-specificity) that allows for the comparison of the overall performance of the SVM and MLP classifiers.

The hardware requirements for the acquisition, preprocessing, feature extraction and classifier modules when deployed on the RASPI were also investigated. In this thesis the mobile device resources such as average CPU, RAM, file size and execution time were monitored. The RASPI Linux kernel collects these system wide statistics which can be accessed with the top or htop libraries. The average resource consumption results were reported with the RASPI connected to a mouse, keyboard, monitor, Ethernet and the acquisition circuit.

M4CVD's current consumption was also measured to provide future researchers benchmark results they can directly apply to their own systems regardless of the battery used. However, the RASPI does not have a internal system to measure energy consumption. The RASPI was powered using a 5V lab bench supply and current consumption was monitored using an external over-the-top current sense circuit shown in Fig. 4.9 (Regan, 2005). The over-the-top circuit measures the voltage across a shunt resistor (100 mΩ) located between the lab bench and M4CVD. The shunt resistor itself has minimal impact on the current that M4CVD draws and subsequently the voltage across the shunt resistor was very low. The over-the-top circuit includes an amplifier circuit that outputted a voltage proportional to the current that was passed through the shunt resistor. An Ohmmeter was used to validate the over-the-top circuit (Table 4.6) by comparing the current output from both systems at the same time while the RASPI was being used. Both systems show similar current readings

for the assigned RASPI tasks.

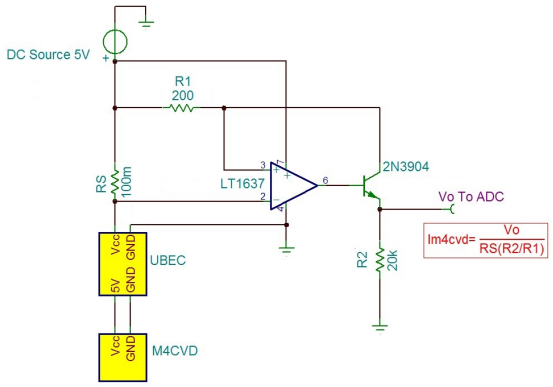


Figure 4.9: The current sense circuit outputs a voltage proportional to the current passing through RS

Table 4.6: The current sense circuit was validated by comparing the circuit output with an Ohmmeter

Raspberry Pi 2 Task	**Ohmmeter**	**Sense Circuit (Recorded Voltage/ Calculated Amperage)**
Booting	0.21 amps	2.8 V (0.28 amps)
Fractal Demo	0.37 amps	3.5 V (0.35 amps)
Video Demo	0.35 amps	3.2 V (0.32 amps)

Chapter 5

Experimental Validation

Health monitoring systems are beginning to incorporate multi-sensor data with other health sources to create pervasive monitoring platforms. However, current monitoring systems require expensive and continuous data transmissions to remote servers to analyze the health data. The objective in this thesis was to develop a CVD monitoring system where all the analysis was performed on a mobile device incorporating data from both physiological sensors and patient's health records. Since M4CVD was deployed on a low resource device the resource requirements for each processing block was also evaluated to identify limitations in a mobile monitoring system.

In this chapter the experimental results of M4CVD are presented and discussed. The chapter is structured as follows: Section 5.1 provides a breakdown of the training set used for optimizing the MLAs. In Section 5.2 the results of the cross-validation training is presented and the best SVM and MLP classifier's performance are discussed in Section 5.3. In Section 5.4, M4CVD's hardware performance is presented as deployed on a Raspberry Pi 2. Finally, the overall research findings and system limitations are discussed in Section 5.5 and 5.6 respectively.

5.1 Training Data Distribution

This study was performed on a dataset of 518 patients containing both physiological and clinical records formed from a subset of the MIMIC II database. 518 patients had been assigned DRG metrics and of those 502 patients also had SAPS metrics. The clinical characteristics of the training database is shown in Table 5.1 examining the baseline features of the low, high and total population groups. A two sample t-test was used to compare continuous variables (e.g., age) while a chi-square test was to compare categorical variables (e.g., gender) between the low and high risk groups with a p value less than $\alpha = 0.05$ deemed significant. With the DRG labeling system, the significance test showed that weight and systolic and diastolic blood pressure was different between the low and high groups. For the SAPS labeling system the significance test showed that age and diastolic blood pressure was different between the risk groups. These results suggest there is a difference between the low and high risk groups identified when labeled by the DRG and SAPS metrics. Subsequently, the machine learning algorithms may be able to separate the classes by constructing a hypersurface. MLAs are capable of identifying non-linear relationships unlike the basic statical tests used in this section. However, the statistical results suggest that the DRG labeling may yield a better classifier since more baseline features are statistically different implying greater separation between low and high groups

Table 5.1: Patient baseline characteristics using DRG and SAPS I severity labels. Bold p-values are clinically significant ($\alpha < 0.05$)

Clinical Features	**Labeling System**	**Low Risk (LR)**	**High Risk (HR)**	**p-value between LR and HR**	**Total Patient Population**
Age	DRG	68.7 ± 16.0 (n=267)	67.1 ± 14.0 (n=251)	0.238	67.9 ± 15.0 (n=518)
	SAPS I	65.0 ± 15.6 (n=273)	70.1 ± 15.6 (n=229)	**0.00001**	67.7 ± 15.0 (n=502)
Gender (Female)	DRG	40.4% (n=267)	32.3% (n=251)	0.053	36.5% (n=518)
	SAPS I	34.4 % (n=229)	38.9% (n=229)	0.304	36.5% (n=502)
Weight (kg)	DRG	81.5 ± 21.5 (n=267)	85.5 ± 18.3 (n-251)	**0.021**	83.4 ± 20.1 (n=518)
	SAPS I	84.4 ± 21.2 (n=273)	82.8 ± 18.9 (n-229)	0.36	82.2 ± 19.2 (n=502)
Systolic Blood Pressure (mmHg)	DRG	148.5 ± 30.8 (n=267)	137.2 ± 32.1 (n=251)	**0.00005**	143.0 ± 31.9 (n=518)
	SAPS I	144.7 ± 30.0 (n=273)	141.2 ± 33.9 (n=229)	0.22	143.1 ± 31.6 (n=502)
Diastolic Blood Pressure (mmHg)	DRG	92.2 ± 22.8 (n=267)	81.7 ± 25.3 (n=251)	**0.000001**	87.1 ± 24.6 (n=518)
	SAPS I	89.3 ± 23.1 (n=273)	84.1 ± 25.8 (n=229)	**0.02**	86.9 ± 24.5 (n=502)

It is important to note that all the patients included in the dataset were admitted to the ICU for heart disease. As a result, the mean averages of the baseline characteristics (e.g., age, BP) are higher compared to the overall population. When the training set mean values in Table 5.1 are compared to the Heart and Stroke Foundation's physiological limits in Table 3.4. Both the low and high severity groups on average would be classified as high risk for systolic blood pressure ($>$ 140 mmHg) and medium risk for diastolic blood pressure (80-90 mmHg). The training set input feature on average are much higher than the overall population. When the inputs are categorized using thresholds determined by the overall population and not from those who already have heart disease. The discretization stage may now be limiting the resolution of the system as certain feature sets will be classified as high risk. Another unexpected results was that low risk patients in the training set on average had higher blood pressure (148 mmHg) compared to high risk patients (127 mmHg). Lower average blood pressure may be attributed to the additional medication and attention that higher risk patients receive in the ICU which suggests that the automatic labels was correctly identifying high risk patients.

5.2 Classifier Performance

In this section cross-validation training was performed to identify the best training configurations that maximize accuracy. In addition to identifying the optimal hyper-parameters the thesis also investigated: 1) two machine learning algorithms (SVM/MLP) 2) two class labeling systems (DRG/SAPS) and 3) three input feature formats (normalized, discretized or one hot format). Seven popular SVM kernels and three MLP learning functions as previously discussed in Section 4.5 were also investigated. In total 60 unique algorithm combinations were trained, tested and evaluated using 10-fold cross-validation.

Table 5.2 shows the cross-validation accuracy for the SVM kernels investigated in this

thesis. The kernels tested had different complexity levels ranging from low (linear) to high (RBF) but they all achieved similar CV accuracy. Higher classifier complexity did not always improve accuracy and indicates that the mapping to higher dimensional space was not successfully separating the data. However the choice of kernel did impact training time from a few minutes (linear) to a hours (polynomial). In fact, the polynomial kernel took the longest to train yet resulted in no improvement in accuracy. Polynomial kernels are known to have only intermediate levels of discrimination and subsequently are less widely used (Lee *et al.*, 2011). The RBF kernel is a popular kernel in the literature due to its ability to handle non-linear decision boundaries. The results in this thesis agree showing the RBF kernel performing best over all three feature input types. The RBF kernel achieved the highest SVM accuracy of 65.12%. These results demonstrate the importance of testing numerous kernels with different complexity levels since the SVM designer may accept a decrease in accuracy if it can be achieved using a less computationally complex classifier.

The MLP cross-validation results are shown in Table 5.3 and shows similar accuracy regardless of the learning function tested. The learning function is the method by which the weights and bias values are updated during backpropagation as described previously in Section 3.5.2. The simplest learning function investigated was the gradient descent with momentum which had similar performance to the other two more complex functions. The Levenberg-Marquardt back-propagation method achieved the highest accuracy (79%) but has higher memory requirements compared to other learning functions. However, the learning function was only used during the training stage and was not deployed on the device. Unlike the SVM, the MLP failed to converge towards an optimal solution during CV training. The MLP was expected to show consistent accuracy improvement before plateauing (or slightly decreasing) as more processing neurons are included. Instead, as shown in Fig. 4.7, the MLP achieved inconsistent accuracy as more neurons were added. For example, some MLP networks achieved 69% accuracy with a single hidden layer containing

5 neurons while another network with 63 neuron's in the hidden layer achieved only 67% indicating that the MLP was not proportionally improving in the classification task. Despite this, the MLP outperformed the SVM by achieving the highest cross-validation accuracy of 79% compared to the SVM maximum of 65.12%. The MLP also did show improved classification accuracy when testing different class labels and different feature input formats. The improved results may reflect how the MLP updates its internal weights and bias during training. The MLP updates each weight and bias individually through back-propagation while the SVM maps the input data into higher dimensional space. It appears that the MLP was better positioned to consider the relative importance of each input feature individually when trying to optimize its weights/bias. As a result the MLP does respond to the different input feature type formats.

Both the DRG resource and SAPS severity metric are included in clinical records in the U.S.A and can potentially be used as automatic severity labels to enable patient records to be analyzed using MLAs. The results show that the SVM and MLP were successfully able to distinguish between low and high risk patients when using either the DRG and SAPS metrics. The MLP was more sensitive to the choice of labeling systems showing higher accuracy with the SAPs labels indicating better separation between the classes compared to the DRG. The SVM showed similar performance regardless of the labeling method used and was not as sensitive. In order to understand these results the method for calculating the SAPS and DRG metrics needs to be considered. The SAPS score was calculated based on 17 physiological characteristics which includes age, BP and ECG overlapping with 9/11 (81%) features used as inputs to M4CVD. On the other hand the DRG metric was calculated by an expert based how many resources the patient uses during there stay in the hospital. The only metrics shared between DRG and M4CVD is age and gender (2/11) so it appears the MLAs had more difficulty separating the two classes. The results indicate that the SAPS metric is a better tool for the automatic risk labeling of patient records compared to

Table 5.2: Accuracy measurement of 10-fold cross-validation for the support vector machine for DRG and SAPS I labeling

Classifier	Kernel	Feature Input Format	DRG CV Accuracy	SAPS I CV Accuracy
SVM	Radial Basis Function	Normalized	64.46%	62.95%
		Discretize (0,1,2)*	64.39%	**65.12%***
		One Hot (00,01,10)	64.07%	63.35%
	Polynomial (2nd order)	Normalized	62.35%	63.97%
		Discretize (0,1,2)*	61.41%	**64.35%**
		One Hot (00,01,10)	64.20%	64.20%
	Polynomial (3rd order)	Normalized	62.16%	62.77%
		Discretize (0,1,2)	61.87%	61.87%
		One Hot (00,01,10)*	**63.14%**	61.35%
	Polynomial (4th order)	Normalized	62.33%	62.38%
		Discretize (0,1,2)*	61.78%	61.78%
		One Hot (00,01,10)	**64.11%**	61.16%
	Polynomial (5th order)	Normalized	62.35%	61.17%
		Discretize (0,1,2)*	62.32%	**62.32%**
		One Hot (00,01,10)	62.56%	61.56%
	Linear	Normalized	53.10%	63.60%
		Discretize (0,1,2)*	63.71%	63.70%
		One Hot (00,01,10)	**64.00%**	61.20%
	Sigmoid	Normalized	63.88%	64.00%
		Discretize (0,1,2)*	**63.50%**	61.98%
		One Hot (00,01,10)	64.49%	60.16%

Table 5.3: Accuracy measurement of 10-fold cross-validation for the MLP for DRG and SAPS I labeling. The number of neurons in the hidden layer is shown in brackets

Classifier	Learning Function	Feature Input Format	DRG CV Accuracy	SAPS I CV Accuracy
MLP	LM	**Normalized***	68.63% (2)	**79.00%* (77)**
		Discretize (0,1,2)	69.96% (5)	77.00% (72)
		One Hot (00,01,10)	67.00% (63)	74.00% (68)
	SCG	Normalized	65.69% (5)	**78.00% (31)**
		Discretize (0,1,2)	70.50% (10)	75.00% (18)
		One Hot (00,01,10)	70.60% (13)	73.00% (45)
	GD	Normalized	67.65% (38)	**72.00% (95)**
		Discretize (0,1,2)	68.63% (32)	71.00% (63)
		One Hot (00,01,10)	70.59% (74)	64.00% (74)

LM: Levenberg-Marquardt Backpropgation SCG: Scaled Conjugate Gradient Descent GD: Gradient Descent with Momentum

Best accuracy for each function is bolded. Highest accuracy overall for each MLA denoted with *

the DRG metric. Automatic labeling provides a useful tool to allow researchers to quickly evaluate a public databases suitability for research use. It could also assist researchers in identifying useful subset of patients in a database for follow up manual evaluation. For example, in this thesis automatic labeling was used to identify the most promising subset from the initial 1000 patient cohort.

As previously discussed in Section 3.4 the input features can be fed into the MLA in three different input formats (normalized, discretize and one-hot coding). The expectation was that discretizing the input features corresponding to their unique feature risk level would allow the algorithms to better identify critical features and training examples. Unfortunately the results show only a slight increase in accuracy between the normalized input feature and its equivalent discretized value. In fact the normalized input format outperformed the discretize format in the MLP. The highest accuracy was achieved by a MLP that used a normalized input feature set with no *a priori* medical knowledge. The discretization stage may be injecting designer bias into M4CVD. As previously discussed in Section 5.1, the training population on average has higher feature mean values compared to the overall population. Since the physiological ranges used in Table 3.4 are based on the entire population the discretization stage assigns many patient's features as medium or high risk reducing the MLA's sensitivity. On the other hand, a normalized input feature with no discretization stage enables the MLA to determine for itself what defines high and low risk within this patient population. While a discretization stage may provide slightly improvements to the MLA it seems to be inhibiting the algorithms ability to generalize. In addition, the discretization stage also introduces additional hardware overhand with little increase in accuracy.

5.3 Monte Carlo Simulation Results

The best SVM (RBF kernel, SAPS labeling, discretization features format) and MLP (Levenberg-Marquardt learning function, SAPS labeling, normalized features format) from the cross-validation tests was selected for further study. The performance metrics used to describe the behavior of each classifier are accuracy, sensitivity, specificity and AUC as shown in Table 5.4. The SVM outperformed the MLP in the Monte Carlo tests achieving the best overall performance values, i.e., an average accuracy of 62.5%, AUC of 0.66, specificity of 76.21% and maximum accuracy of 71.3%. Fig. 5.1 shows the mean ROC for the SVM (red) and MLP (blue). The closer the ROC curve is to the upper left corner the higher the overall accuracy of the classifier (Zweig and Campbell, 1993). The results show that the mean SVM ROC was consistently higher compared to the MLP ROC indicating that the SVM was the better classifier for severity classification. The ROC curves demonstrates that both models achieved stable, reusable parameter configurations. Based on classifier performance alone the authors would recommend the SVM for deployment in M4CVD. However, the classifier complexity must also be considered before the best algorithm can be determined. The number of support vectors can be used as a representative of the model's complexity Lee *et al.* (2011). On average, the SVM used 271 patients (out of 351 training examples) as support vectors for constructing the hypersurface indicating that the algorithm was successfully generalizing from the training data. But the best SVM used the most complex kernel studied in this thesis. On the other hand, the best MLP perceptron has 1694 weights and bias between its 11 feature input layer, 77 neuron hidden layer and 2 neuron output layer. The algorithm's hardware performance on the RASPI is discussed in Chapter 5.4.

Table 5.4: M4CVD Performance for SVM and MLP. The mean of 1000 experiments is shown for each performance metric

Classifier	Accuracy Max	Accuracy Min	Accuracy Mean
SVM	71.30%	49.00%	62.5 ± 3.64 %
MLP	82.00%	36.00%	58.9 ± 6.61 %

Classifier	AUC	Sensitivity	Specificity
SVM	0.66 ± 0.03	45.53 ± 07.04%	76.21 ± 06.01%
MLP	0.61 ± 0.08	58.89 ± 11.57%	59.05 ± 11.23%

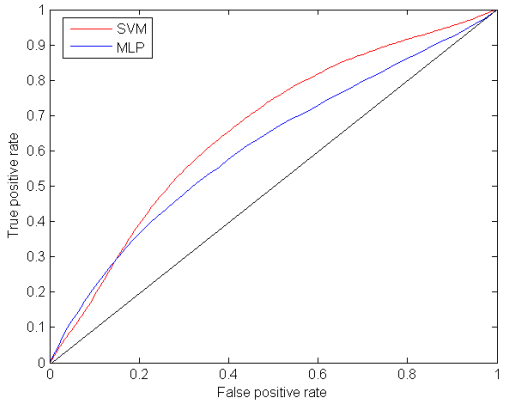


Figure 5.1: ROC curve for severity estimation. The mean of 1000 experiments has been shown for each ROC curve

These results are surprising considering the superior CV accuracy achieved by the MLP (78%) compared to the SVM (65%). One explanation is that the MLP weights and bias are initially randomly selected before training. As discussed in Chapter 3 multiple local minima solutions are possible and the MLP may be getting trapped in different local minima during optimization. The SVM on average preformed better than the MLP which is also reflected in the CV results. The range of CV accuracy for MLP was twice as large (43%-78%) compared to the SVM (52%-65%). The MLP did not converge to a solution as more neurons were added. The strength and weakness of the MLP seems to be its ability to update individual weights and bias through backpropagation. Since the MLP is more sensitive to each individual features contribution it makes it more difficult for the MLP to converge between different training/testing subsets due to the large variability within each feature. The Monte Carlo simulations shows the MLP average accuracy was slightly lower than the SVM at 58.9% but did achieve the highest maximum average of 82% during the Monte Carlo run. These results indicate the great potential of the MLP in this classification problem if the proper dataset is used for training.

Unlike the MLP, the SVM finds only the global minima allowing it to reproduce its

results across different subsets of training and testing data. It is also possible that the high number of support vectors allows the algorithm to remain consistent between different training sets as most training examples are retained by the model. The SVM kernel trick for mapping into higher dimensional space appears to allow it to generalize more consistently. The identification of the support vectors provides information on which patient records should be selected for further study.

The predictive accuracy results in this thesis are somewhat low compared to the previous work done in Boursalie *et al.* (2015) which evaluated a proof-of-concept M4CVD. M4CVD was tested (no preprocessing modules) on a synthetic database of 200 patients using a SVM and achieved an accuracy of 90.5%. The lower results in this thesis can most likely be attributed to the technical limitations of the training dataset as discussed previously in Section 4.3.2. The synthetic database was also more generally distributed between all possible feature values. A distribution based on the general population may have resulted in higher accuracy compared to when the data is from the population who already have CVD (MIMIC II). A synthetic database with clean data would also have a more distinct separation between the two classes making the calculation of the hypersurface easier.

The system results presented in this thesis are promising since they do exceed those of current early-warning systems used by medical emergency teams in hospitals (Hillman *et al.*, 2005). Medical emergency teams are trained health professionals that intervene early due to any quick changes in a patient's cardiac health that matches their calling criteria. The purpose of these teams is to intervene early to improve patient outcomes. Hillman *et al.* investigated 23 hospitals (11 control, 12 interventions) and found that the teams calling criteria had an accuracy of 30% of identifying patients who were subsequently admitted to the ICU. The study also found no substantial impact of these teams interventions. The results in this thesis show that M4CVD achieved an accuracy of 58.9% (SVM)/ 58.9% (MLP) in identifying between low and high risk patients. M4CVD had better specificity

(76%-SVM, 59%-MLP) compared to the manual systems as well.

Similar monitoring systems in the literature such as the Heart2Go (Oresko *et al.*, 2010) and the CHRONIOUS (Bellos *et al.*, 2014) systems achieved results of 90% and 98% respectively. Both systems were deployed on a mobile device further demonstrating the feasibility of the proposed system. HeartToGo was strictly a beat classification system using a single wearable sensor. M4CVD improves on HeartToGo by incorporating multiple sensors and clinical data. CHRONIOUS, a 5-level severity estimator, achieved a high accuracy of 94% using both the SVM and MLP. CHRONIOUS's training data (like M4CVD) was only collected from patients who already have heart disease. Unfortunately the suitability of the two training databases cannot be compared since the details on the CHRONIOUS's training set were not published (e.g., number of each class, baseline characteristic distribution). In addition CHRONIOUS's training set was relatively small with 30 patients compared to the 518 patients used in this thesis. Neither system investigated the hardware consumption requirements of their systems as is done in the next section.

5.4 Hardware Performance

All the acquisition, preprocessing and classification modules of M4CVD were implemented on a Raspberry Pi 2. The total program size was 371 KB which is reasonable for smartphones and other low processing systems. The execution time of M4CVD when deployed on the RASPI is split between the acquisition, preprocessing and analysis stages. M4CVD first records the 5 min ECG signal. The subsequent data processing and classification stages of M4CVD analysis takes approximately six seconds. Overall, M4CVD was very fast compared to a manual human monitoring. For example, analyzing 500 patients would take a human several hours to complete while M4CVD can analyze all the records in less than one hour. M4CVD required 274 mA for the ECG acquisition module and 242.6 mA to run the

processing and analysis blocks. A device's battery life can be calculated using Eq. 5.1 where 0.7 is a constant to allow for external factors that can affect battery life. Assuming a device with a 1200 mAh battery, M4CVD will last about 1.5 hours. However, this not an accurate measure of the systems lifespan as it assumes that M4CVD was continuously running and handling both the signal acquisition and analysis. M4CVD could be interrupt driven to only run at periodic intervals (still more frequent then the health professional) to increase battery life. For the remainder of this chapter the current consumptions requirements for each block will be examined instead of battery life.

$$\text{Battery Life} = (\text{Battery Capacity (mAh)} \,/\, \text{Load Current (mA)}) \ * 0.70 \qquad (5.1)$$

Table 5.5 shows the breakdown for each module in M4CVD examining hardware resources (CPU, memory and execution time) and current consumption. These results show that the classification stage (SVM or MLP) required among the lowest resources in terms of execution time and current consumption. However, the MLP required among the highest levels of CPU. It was surprising that the preprocessing (filter, detrend, quality and R peak detection) and the feature extraction modules in total accounted for 98% of the program execution time not including the 5 minute ECG acquisition stage. The quality assessment was the single longest module to run (2970 ms) in M4CVD as the Pan-Tompkin algorithm uses additional filters for R peak detection and was not optimized for mobile devices. The preprocessing modules all used similar CPU (25%) as they processed the same 5 minute ECG signal. The normalization/ discretization stage was also surprisingly CPU intensive at 41%. Overall the system's memory usage was minimal indicating that these types of system may not require large RAMs once deployed. Despite the MLAs have similar classifier performances during the Monte Carlo simulations they showed very different hardware performances. The SVM took nearly 70x longer to analyze the same processed input data and consumed nearly 2x the current compared to the MLP. The SVM also requires more

Table 5.5: Hardware consumption for acquisition, data processing and classification modules on Raspberry Pi 2. Highest value for each metric is bolded

Metric	Acquisition	Filter	Detrend	Quality Assessor	R Peak Detection	Feature Extraction	Normalization/ Discretization	MLA	
								SVM	MLP
CPU Usage	1.39 ± 0.01%	25.81 ± 5.79%	25.39 ± 4.00%	26.35 ± 4.31%	25.75 ± 5.12%	28.87 ± 11.19%	41.50 ± 18.32%	38.28 ± 6.11%	**54.03 ± 16.44%**
Memory Usage	0.2%	0.4%	**0.59%**	1.45%	0.29%	0.3%	0.2%	0.12%	0.1%
Program Size	57 KB	12 KB	28 KB	28 KB	**130 KB**	16 KB	16 KB	68 KB	20 KB
Execution Time	322.5 secs (5 min)	902.87 ± 9 ms	1660 ± 10 ms	**2970 ± 18 ms**	605 ± 5.86 ms	4.27 ± 0.15 ms	2.67 ± 0.38 ms	71.0 ± 1.72 ms	1.65 ± 0.10 ms
Current Consumption	274 ± 6.6 mA	28.6 ± 6.4 mA	34.3 ± 7.3 mA	33.3 ± 7.5 mA	**103 ± 17 mA**	16.7 ± 14.9 mA	8.3 ± 6.9 mA	11.7 ± 10.7 mA	6.7 ± 7.5 mA

physical space and slightly more memory usage compared to the MLP. The results indicate that the MLP was more efficient for deployment on a low resource device. The superior MLP efficiency seems to result from a fundamental difference between how the SVM and the MLP construct their respective hypersurfaces. When deployed the SVM must constantly map any new input data vector into a higher dimensional space to compare it against the hypersurface constructed from the support vectors. On the other hand, once training was complete the MLP becomes only a series of equations using the optimized weights and bias. The device can then process the MLP equations more quickly compared to the kernel mapping (which includes a dot product) required in the SVM. The hardware consumption results are in direct contrast to the cross-validation training where it took longer to optimize the MLP compared to the SVM since each new neuron increased the MLP's computational complexity and training time. On the other hand, the training time per SVM remains approximately the same as only the hyper-parameter values change within the same kernel. However, while training time is an important factor for the designer it does not impact the final deployment on a mobile device. Despite the SVM achieving higher accuracy and specificity there was a greater computational expense to deploying the SVM on a mobile device. These results indicate that the designer must consider the relative importance of both classifier accuracy and hardware requirements. For example, a designer for a smartphone application may decide to use the SVM despite its hardware requirements. The designer of

an ultra low resource or implantable devices would perhaps accept the slight reduction in accuracy in order to achieve greater hardware efficiency. The importance of considering the computational cost of machine learning algorithms also applies to other mobile applications such as speech and activity recognition.

Just as the SVM and MLP parameters can be modified to improve classifier performance they can also be adjusted to optimize hardware consumption. The CV of the MLP shown in Fig. 4.8 in Chapter 4 allows the designer to select an architecture that achieves similar accuracy to the best case but uses less neurons (less complexity) in the hidden layer. In effect, the designer is making a trade-off in algorithm accuracy to improve efficiency. Similarly, the SVM can have different kernels investigated (such as the linear kernel) that provide similar accuracy performance but may consume less hardware resources. Another method to improve SVM complexity is to reduce the number of support vectors as demonstrated by Lee *et al.* who investigated the energy trade-offs between the number of support vectors and energy consumption (Lee *et al.*, 2011). Other studies have also investigated improving SVM through fixed-point arithmetic (Anguita *et al.*, 2012). Finally both algorithms would benefit from reducing the number of input features fed into each MLAs. Research has shown that reducing the number of input features can result in improved hardware performance without a major reduction in system accuracy (Comito and Talia, 2015).

Studies in the literature have begun to investigate the consumption requirements when evaluating their systems as the availability and use of low resource devices has grown. Kunnath *et al.* WiCard system transmitted accelerometer data and ECG data through a microcontroller via Bluetooth to a remote server for analysis. The current requirements of each device in WiCard was determined and the transmission stage consumed the most current (Kunnath *et al.*, 2013). Similarly, Alsurafa *et al.* also determined that the majority of battery life in their system was spent on processing sensor data and transmitting the data to servers (Alshurafa *et al.*, 2015). However, they did not investigate the consumption

requirements of each module in their system as done in this thesis. The results presented in this thesis demonstrated that the MLA's complexity was not a barrier for adoption on a low resource device. In fact, M4CVD saved computational overhead and communication expenses by completing the analysis locally.

5.5 Research Findings

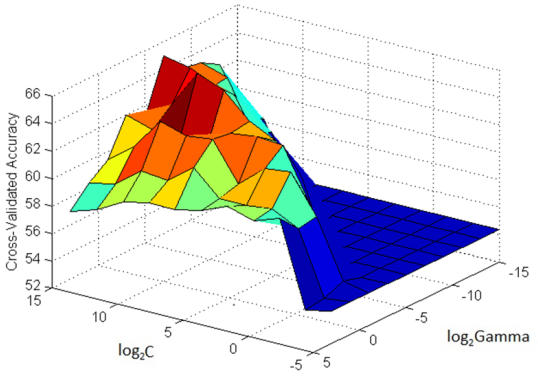
Remote health monitoring has been shown to improve patient's cardiovascular health and reduce hospital readmissions (Clark *et al.*, 2007). Concurrent advances in sensors, health records, mobile devices and health informatics provides new opportunities to move treatment from periodic monitoring towards continuous care (Andreu-Perez *et al.*, 2015). Increasingly patient's health data is being analyzed automatically. Some of the analysis is also beginning to take place directly on the low-resource processing platforms. As a result, an important evaluation criteria of these systems has become their hardware performance and resource consumption.

The common operations used in machine learning and signal processing have a complexity order of approximately $O(n^3)$ (Andreu-Perez *et al.*, 2015). The initial hypothesis in this thesis was that machine learning algorithms are very complex and present a considerable burden for low resource devices. However, this thesis showed that the MLAs can be run successfully on a low resource device. In fact, the MLAs were among the least consuming modules of the system studied. Overall the results of this thesis indicate that the greatest area for improvement lies with creating more mobile efficient preprocessing modules. For example, the ECG preprocessing module in this thesis improved the quality of the entire ECG signal. However, M4CVD only analyzes features based on the R wave. A band-pass filter could be used to extract only the R wave reducing the preprocessing module's computational requirements.

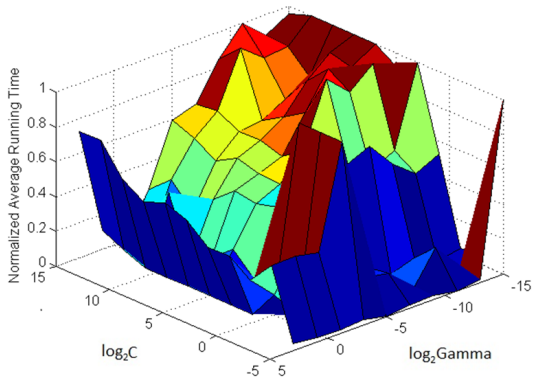
This thesis has demonstrated that a low resource device can act as a monitoring platform analyzing hybrid data from multiple physiological sensors and health record data to assess patient severity. All the signal acquisition, preprocessing, feature extraction and analysis modules were completed on the same low resource device. Many of the systems reviewed in the literature used mobile devices only for signal acquisition and preprocessing before transmitting the data to a remote server. The findings in this thesis suggest that these systems are already handling the more computationally expensive tasks locally and are actually increasing the overhead to their systems by transmitting to remote servers. A system capable of local analysis offers several advantages such as lower server and cellular costs. M4CVD would also continue to function even in poor coverage zones. As wearable devices become more powerful they can begin to take over these signal processing duties which would further lessen the load on the main monitoring device.

The objective of cross-validation training of a MLA was to locate the hyper-parameters that maximize system accuracy. In this thesis, k-fold cross-validation was conducted on both the SVM and MLP. The traditional cross-validation results showed similar average accuracy across the various algorithms, kernels and functions that were tested. However, a very interesting result was that the MLP's weights and bias network was more efficient on a mobile device compared to the SVM's kernel. Doing only a traditional cross-validation, the designer would perhaps choose the SVM over the MLP despite their similar accuracy and the nearly 70x more computational power the SVM required. These findings indicate that future designers of machine learning algorithms should not just measure accuracy but also measure resource consumption during cross-validation training. The optimal classifier is now a balance between acceptable classifier accuracy with reasonable consumption requirements. Fig. 5.2 shows the proposed cross-validation procedure for the SVM examining both the classifier's accuracy (left) and its normalized execution time (right). It is assumed that the classifier with the shortest run time is also the classifier with the lowest resource

requirements on the mobile device. The new CV graph allows for the selection of the best CV model in terms of both accuracy and average execution time. It would be up to the designer and user to decide the acceptable trade-offs for their application. Different algorithms with various accuracies and energy consumptions could be loaded into M4CVD as different energy profiles where the user can decide the power/accuracy balance in the system settings. For example, examining Fig. 5.2 (left) the highest CV SVM accuracy was 65.3% and took 1.1 ms to run. However, the designer may decide to implement a power saver mode resulting in a 5% decrease in accuracy (65% down to 60%) to save 30% in execution time (1.1 ms down to 0.7 ms) which can improve the system's battery lifetime. Furthermore, it is recommended that each trained MLA is tested directly on the target device to evaluate the model's average CPU, memory and current requirements in addition to the model's accuracy. While the new method proposed in this thesis does increase the training time it would provide researchers a better indicator of their classifiers overall performance.



Highest Accuracy	Running Time
65.33%	1.1 ms
64.94%	1.0 ms
64.56%	1.1 ms



Shortest Running Time	Accuracy
0.77 ms	59.37%
0.82 ms	58.34%
1.1 ms	60.76%

Figure 5.2: The proposed cross-validation procedure examines both accuracy (left) and normalized execution time (right) to identify the best overall classifier

Finally, this thesis highlights the current challenges associated with using open source clinical datasets for machine learning research. Many researchers are beginning to publish their training sets on PhysioNet and other file sharing services to allow other researchers to use their previously collected datasets for new research. The MIMIC II database is one example of research groups and hospitals collaborating to collect and publish anonymized clinical and physiological data for research use. In addition, the increasing accessibility of patient's health records and the growing acceptance of wireless sensors provides larger and more diverse datasets for machine learning applications. However, as previously discussed a limitation of using these public databases is that researchers are restricted to data that is available. The MIMIC II database contained ECG signals that are intended for visual inspection only because the original designers did not anticipate the need for more detailed recordings. The limitations of the MIMIC II database prevented the authors from using more sensitive feature processing techniques in this thesis. In addition, key health metrics such as smoking information and family medical history were not readily available for all patients. The types of missing data and the insufficient storage of data limited the number of patients and type of features that could be studied in this thesis. Another common challenge is that datasets collected by researchers did not have potential machine learning or signal processing applications in mind which results in additional overhead before the datasets can be used. Therefore this thesis demonstrates the importance of developing better standards regarding the collection, content and general format when publishing databases for research use. The standards could also include procedures for machine learning and signal processing applications. While PhysioNet has some guidelines for publishing new datasets they focus only on ensuring that new databases can interface with PhysioNet viewers.

5.6 Limitations

The current system exhibits some limitations. First, the training dataset was composed entirely of ICU patients. These patients were under high levels of stress since they were ill and outside of their homes. As a result, the patients may have abnormally high readings of physiological metrics also known as *white coat syndrome* (Owens *et al.*, 1999). The ICU patients were under constant observation and on different medications that further impacts their vital signs. Medication usage and dosage were also not studied in this thesis due to the difficulty of extracting the information from the MIMIC database. Overall these patients are not the best representation of the population who would be using M4CVD in their day to day life. All of these factors help explain the classifier difficulty to create a dividing hypersurface between the two classes.

Second, the machine learning algorithms tested for M4CVD were not large in terms of their architecture. The SVM deployed in M4CVD had only a few hundred support vectors while some SVMs in the literature have thousands. However, the maximum number of support vectors that can be studied is always limited to the number of examples available during training. Similarly, MLP can have thousands of neurons and multiple layers while the MLP used in the studied system had less than 100 neurons in a single hidden layer. The consumption requirements of larger networks should be investigated. It is important to note that the MLP may not always have better hardware requirements and depends on multiple factors. It is critical for researchers to evaluate the hardware requirements of their own systems in addition to evaluating classifier accuracy.

Finally, the open sourced algorithms and processing libraries used in this thesis were not designed with mobile applications in mind. Mobile optimized algorithms may have different performance results. M4CVD also did not integrate with a health record system in real time so the hardware requirements for accessing and retrieving data was not evaluated.

Furthermore, no security and privacy protocols were implemented in M4CVD. As a result, the hardware requirements of the security and privacy components were not studied on a low resource device which may provide additional overhead to M4CVD.

Chapter 6

Conclusion and Future Work

In this chapter the thesis contributions are summarized and directions for future research are discussed.

6.1 Summary of Contributions

The proposed research was to develop a health monitoring system for cardiovascular disease based on a mobile device. The proposed system analysis was based on MLAs which analyzed features extracted from a clinical database containing both real-time physiological recordings and medical data from a population of 518 patients. It was successfully validated that the proposed system can use wearable sensors to record the physiological signals found in the MIMIC II clinical database to a mobile device . The suitability of the SVM and MLP for health monitoring applications was studied. Several features associated with CVD that can be monitored using wearable devices and/or health records were examined. This thesis also investigated the effect of two automatic class labels and three input feature formats on classifier accuracy. The process of designing the system also led to the investigation of the resource requirements for a health monitoring system on a low resource device. The work

in this thesis contributes towards the goal of personalized predictive monitoring.

6.1.1 Model to Monitor Cardiovascular Disease

This research has successfully demonstrated that a mobile device can be used as a health monitoring platform that can analyze a hybrid of both sensor and health record data. A machine learning algorithm was designed that successfully demonstrated the concept of classifying a patients CVD severity automatically. The proposed system can assist health care professionals by presenting relevant and summarized information on a patient's condition from all of the collected data. M4CVD allows the management of patient care to move away from episodic appointments towards care that is non-invasive, integrated and continuous. M4CVD is capable of monitoring clinically relevant events that would otherwise be missed between traditional medical appointments. Patients will also benefit from continuous feedback between their medical appointments. The proposed system would allow patients to become active participants in the management of their health. Unlike existing systems, M4CVD can perform the analysis locally without the need for continuous transmission to a remote server saving computational and financial costs. Finally, M4CVD can monitor a patient continuously even outside of a controlled clinical setting.

This thesis evaluated the SVM and MLP to assess their suitability for deployment in a monitoring system on a mobile device. While the SVM had better accuracy (62%), the MLP was more efficient on the mobile device in terms of execution time and current consumption. The results demonstrate the importance, throughout the development process, for evaluating machine learning systems not just in terms of traditional classification accuracy but also computational resource consumption.

6.1.2 Automatic Class Labels and Input Formats

The work in this thesis has also demonstrated how automatic class labeling techniques can be used on previously unlabeled datasets without the need for manual labeling. The DRG and SAPS metrics can be used on previously unlabeled datasets enabling them to be analyzed by MLAs. Overall, the authors recommend the SAPS metric for severity estimation as it produced a better MLP (82%) classifier which suggests better separation between the low and high severity patient classes.

Different input formats and their effect on the classifier's accuracy were also investigated. A discretization stage was used before the classification module that incorporated medical knowledge regarding individual feature risk levels into M4CVD. The results in this thesis did not show a consistent improvement in system accuracy. Despite these results, incorporating medical knowledge in a pre-analysis stage has great potential in machine learning applications.

6.1.3 Evaluation of Model Resource Consumption

The last contribution in this thesis was to evaluate the resource requirements for each module found in a monitoring system: acquisition, signal processing, feature extraction and classification. The results of the literature review shows that most systems in the literature were not evaluated in terms of hardware consumption. When system resource consumption was studied the researchers only examined their system as a whole without testing the individual modules as was done in this thesis. The results in this thesis show that the MLA's complexity was not a barrier for adoption on a low resource device. In fact, the preprocessing modules that transformed the raw recorded data into MLA inputs was the most computationally expensive modules. The results in this thesis shows that the functions used for signal processing and feature extraction have the greatest areas for improvement

in terms of mobile efficiency. The hardware benchmarks presented in this thesis can serve as a reference to future researchers developing or improving their own mobile applications.

6.2 Future Directions

In this section recommendations for future work in three areas are discussed. First, M4CVD can be further assessed and evaluated with patients inside a clinical setting. Second, focus on improving system accuracy and extending the model presented in this thesis to analyze new input sources. Finally, researchers can focus on improving the machine learning algorithms accuracy and efficiency when deployed in M4CVD.

The M4CVD model presented in Chapter 3 was successfully implemented on a Raspberry Pi 2. While the Raspberry Pi 2 is mobile, the next step may be to implement the model on a smartphone for easier integration and connectivity with wearable sensors. The system can be integrated with other wearable devices and a hospital health record database. Clinical experiments can then be conducted to determine M4CVD's effect on disease management. Research can also focus on assessing the benefits of the system to health professionals and patients.

Future work will also focus on extending the proposed model capabilities to analyze additional physiological and clinical parameters. For example, a patient's diet and level of physical activity are additional key indicators when assessing cardiovascular health (Heart and Stroke, 2013). M4CVD could integrate with sensors located on a patient's smartphone and with existing mobile applications (e.g., My Fitness Pal) to monitor diet and physical activity. Activity recognition (e.g, accelerometers) could also improve the ECG analysis in M4CVD by associating the input features with information on the patient's current physical activities. The input features discussed in this thesis can also be expanded. Features such as long term (24 hour) ECG HRV features, beat classifications and frequency derived features

(power ratio) can also be implemented.

Additional research can be done to improve M4CVD's accuracy and efficiency. Feature selection algorithms (Guyon and Elisseeff, 2003) can be implemented to identify the important features that best contribute to M4CVD's accuracy while lowering the system's overall complexity. The model developed in this thesis can also be further extended for multi-class severity classification (e.g., low, medium and high) to provide more resolution on a patients overall risk level. Additional MLAs such as random forest trees and bayesian networks could also be investigated to evaluate their accuracy and mobile efficiency. Finally, future research must be done to improve the efficiency of the current generation of MLAs and signal processing libraries specifically for running in a mobile environment.

Bibliography

Alshurafa, N., Eastwood, J.-A., Pourhomayoun, M., Liu, J. J., and Sarrafzadeh, M. (2014). Remote health monitoring: Predicting outcome success based on contextual features for cardiovascular disease. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 1777–1781.

Alshurafa, N., Eastwood, J.-A., Nyamathi, S., Liu, J. J., Xu, W., Ghasemzadeh, H., Pourhomayoun, M., and Sarrafzadeh, M. (2015). Improving compliance in remote healthcare systems through smartphone battery optimization. *IEEE Journal of Biomedical and Health Informatics*, **19**(1), 57–63.

Andreu-Perez, J., Leff, D. R., Ip, H., and Yang, G.-Z. (2015). From wearable sensors to smart implants—toward pervasive and personalized healthcare. *IEEE Transactions on Biomedical Engineering*, **62**(12), 2750–2762.

Anguita, D., Ghio, A., Oneto, L., Parra, X., and Reyes-Ortiz, J. L. (2012). Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In *International Workshop on Ambient Assisted Living*, pages 216–223.

Anliker, U., Ward, J. A., Lukowicz, P., Troster, G., Dolveck, F., Baer, M., Keita, F.,

Schenker, E. B., Catarsi, F., Coluccini, L., *et al.* (2004). AMON: A wearable multiparameter medical monitoring and alert system. *IEEE Transactions on Information Technology in Biomedicine*, **8**(4), 415–427.

Averill, R. F., Goldfield, N., Steinbeck, B., Grant, T., Muldoon, J., Brough, A., *et al.* (2003). All patient refined diagnosis related groups (apr-drgs). *Version*, **15**, 98–054.

Banaee, H., Ahmed, M. U., and Loutfi, A. (2013). Data mining for wearable sensors in health monitoring systems: A review of recent trends and challenges. *Sensors*, **13**(12), 17472–17500.

Baron, M. J. S., Velasquez, J. J., Cifuentes, C. A., and Rodriguez, L. E. (2011). An approach to telemedicine intelligent, through web mining and instrumentation wearable. In *Computing Congress (CCC), 2011 6th Colombian*, pages 1–5.

Batista, G. E., Monard, M. C., *et al.* (2002). A study of k-nearest neighbour as an imputation method. *HIS*, **87**(251-260), 48.

Bellos, C., Papadopoulos, A., Fotiadis, D. I., and Rosso, R. (2010a). An intelligent system for classification of patients suffering from chronic diseases. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 2890–2893.

Bellos, C., Papadopoulos, A., Rosso, R., and Fotiadis, D. I. (2011a). Chronious: A wearable platform for monitoring and management of patients with chronic disease. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 864–867.

Bellos, C., Papadopoulos, A., Rosso, R., and Fotiadis, D. I. (2011b). Heterogeneous data fusion and intelligent techniques embedded in a mobile application for real-time chronic

disease management. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 8303–8306.

Bellos, C., Papadopoulos, A., Rosso, R., and Fotiadis, D. I. (2011c). A support vector machine approach for categorization of patients suffering from chronic diseases. In *Wireless Mobile Communication and Healthcare*, pages 264–267.

Bellos, C., Papadopoulos, A., Rosso, R., and Fotiadis, D. I. (2012). Categorization of patients' health status in copd disease using a wearable platform and random forests methodology. In *Biomedical and Health Informatics (BHI), 2012 IEEE-EMBS International Conference on*, pages 404–407.

Bellos, C., Papadopoulos, A., Rosso, R., and Fotiadis, D. I. (2013). Clinical validation of the chronious wearable system in patients with chronic disease. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pages 7084–7087.

Bellos, C. C., Papadopoulos, A., Rosso, R., and Fotiadis, D. I. (2010b). Extraction and analysis of features acquired by wearable sensors network. In *Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference on*, pages 1–4.

Bellos, C. C., Papadopoulos, A., Rosso, R., and Fotiadis, D. I. (2014). Identification of copd patients health status using an intelligent system in the chronious wearable platform. *Biomedical and Health Informatics, IEEE Journal of*, **18**(3), 731–738.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, **13**(Feb), 281–305.

Bhargava, R., Kargupta, H., and Powers, M. (2003). Energy consumption in data analysis for on-board and distributed applications. In *Proceedings of the ICML*, volume 3, page 47.

Boursalie, O., Samavi, R., and Doyle, T. (2015). M4CVD: Mobile machine learning model for monitoring cardiovascular disease. In *The 5th International Conference on Current & Future Trends of Information & Communication Technologies in Healthcare*.

Bousseljot, R., Kreiseler, D., and Schnabel, A. (1995). Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet. *Biomedical Engineering*, **40**(s1), 317–318.

Buist, M. D., Jarmolowski, E., Burton, P. R., Bernard, S. A., Waxman, B. P., and Anderson, J. (1999). Recognising clinical instability in hospital patients before cardiac arrest or unplanned admission to intensive care. a pilot study in a tertiary-care hospital. *The Medical Journal of Australia*, **171**(1), 22–25.

Camm, A. J., Malik, M., Bigger, J., Breithardt, G., Cerutti, S., Cohen, R., Coumel, P., Fallen, E., Kennedy, H., Kleiger, R., *et al.* (1996). Heart rate variability. standards of measurement, physiological interpretation, and clinical use. *European Heart Journal*, **17**(3), 354–381.

Chan, M., Estève, D., Fourniols, J.-Y., Escriba, C., and Campo, E. (2012). Smart wearable systems: Current status and future challenges. *Artificial Intelligence in Medicine*, **56**(3), 137–156.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, **2**, 27:1–27:27.

Chen, W., Wei, D., Zhu, X., Uchida, M., Ding, S., and Cohen, M. (2005). A mobile phone-based wearable vital signs monitoring system. In *Computer and Information Technology, 2005. CIT 2005. The Fifth International Conference on*, pages 950–955.

Clark, R. A., Inglis, S. C., McAlister, F. A., Cleland, J. G., and Stewart, S. (2007). Telemonitoring or structured telephone support programmes for patients with chronic heart failure: systematic review and meta-analysis. *BMJ*, **334**(7600), 942.

Clifford, G. D., Azuaje, F., and McSharry, P. (2006). *Advanced methods and tools for ECG data analysis*.

Clifton, L., Clifton, D. A., Watkinson, P. J., and Tarassenko, L. (2011). Identification of patient deterioration in vital-sign data using one-class support vector machines. In *Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on*, pages 125–131.

Clifton, L., Clifton, D. A., Pimentel, M. A., Watkinson, P. J., and Tarassenko, L. (2014). Predictive monitoring of mobile patients by combining clinical observations with data from wearable sensors. *Biomedical and Health Informatics, IEEE Journal of*, **18**(3), 722–730.

Comito, C. and Talia, D. (2015). Evaluating and predicting energy consumption of data mining algorithms on mobile devices. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pages 1–8.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20**(3), 273–297.

Dai, S., Bancej, C., Bienek, A., Walsh, P., Stewart, P., and Wielgosz, A. (2009). Tracking heart disease and stroke in Canada. *Chronic Diseases in Canada*, **29**(4), 192–193.

Depari, A., Flammini, A., Sisinni, E., and Vezzoli, A. (2014). A wearable smartphone-based system for electrocardiogram acquisition. In *Medical Measurements and Applications (MeMeA), 2014 IEEE International Symposium on*, pages 1–6.

Eekhout, I., de Boer, R. M., Twisk, J. W., de Vet, H. C., and Heymans, M. W. (2012). Missing data: a systematic review of how they are reported and handled. *Epidemiology*, **23**(5), 729–732.

Ellis, R. J., Zhu, B., Koenig, J., Thayer, J. F., and Wang, Y. (2015). A careful look at ecg sampling frequency and r-peak interpolation on short-term measures of heart rate variability. *Physiological Measurement*, **36**(9), 1827.

Enders, C. K. (2010). *Applied Missing Data Analysis*.

Engelse, W. and Zeelenberg, C. (1979). A single scan algorithm for qrs-detection and feature extraction. *Computers in Cardiology*, **6**(1979), 37–42.

Esfandiari, N., Babavalian, M. R., Moghadam, A.-M. E., and Tabar, V. K. (2014). Knowledge discovery in medicine: Current issue and future trend. *Expert Systems with Applications*, **41**(9), 4434–4463.

Gao, H., Duan, X., Guo, X., Huang, A., and Jiao, B. (2013). Design and tests of a smartphones-based multi-lead ecg monitoring system. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pages 2267–2270.

Gattinoni, L., Radrizzani, D., Simini, B., Bertolini, G., Ferla, L., Mistraletti, G., Porta, F., Miranda, D. R., *et al.* (2004). Volume of activity and occupancy rate in intensive care units. association with mortality. *Intensive Care Medicine*, **30**(2), 290–297.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*.

Goldberg, R. J., Ciampa, J., Lessard, D., Meyer, T. E., and Spencer, F. A. (2007). Long-term survival after heart failure: a contemporary population-based perspective. *Archives of Internal Medicine*, **167**(5), 490–496.

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, **101**(23), e215–e220.

Graf, A. B., Smola, A. J., and Borer, S. (2003). Classification in a normalized feature space using support vector machines. *IEEE Transactions on Neural Networks*, **14**(3), 597–605.

Guidi, G., Pettenati, M. C., Melillo, P., and Iadanza, E. (2014a). A machine learning system to improve heart failure patient assistance. *Biomedical and Health Informatics, IEEE Journal of*, **18**(6), 1750–1756.

Guidi, G., Pettenati, M. C., Melillo, P., and Iadanza, E. (2014b). A machine learning system to improve heart failure patient assistance. *IEEE Journal of Biomedical and Health Informatics*, **18**(6), 1750–1756.

Gultepe, E., Green, J. P., Nguyen, H., Adams, J., Albertson, T., and Tagkopoulos, I. (2014). From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system. *Journal of the American Medical Informatics Association*, **21**(2), 315–325.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, **3**(Mar), 1157–1182.

Hagan, M. T. and Menhaj, M. B. (1994). Training feedforward networks with the marquardt algorithm. *IEEE transactions on Neural Networks*, **5**(6), 989–993.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, **11**(1), 10–18.

Hampton, J. R. (2013). *The ECG made easy*.

Haykin, S. (2009). *Neural Networks and Learning Machines*, volume 3.

He, T., Clifford, G., and Tarassenko, L. (2006). Application of independent component analysis in removing artefacts from the electrocardiogram. *Neural Computing & Applications*, **15**(2), 105–116.

Heart and Stroke (2013). The Canadian Heart and Stroke Foundation. Heart disease recovery road.

Hillman, K., Chen, J., Cretikos, M., Bellomo, R., Brown, D., Doig, G., Finfer, S., Flabouris, A., Investigators, M. S., *et al.* (2005). Introduction of the medical emergency team (MET) system: a cluster-randomised controlled trial. *Lancet*, **365**(9477), 2091–2097.

Horn, E. and Lee, S. (1965). Electronic evaluations of the fetal heart rate patterns preceding fetal death: further observation. *Am J Obster Gynecol*, **87**, 824–826.

Iapichino, G., Mistraletti, G., Corbella, D., Bassi, G., Borotto, E., Miranda, D. R., and Morabito, A. (2006). Scoring system for the selection of high-risk patients in the intensive care unit. *Critical care medicine*, **34**(4), 1039–1043.

Juen, J., Cheng, Q., and Schatz, B. (2015). A natural walking monitor for pulmonary patients using mobile phones. *Biomedical and Health Informatics, IEEE Journal of*, **19**(4), 1399–1405.

Jung, E.-Y., Kim, J., Chung, K.-Y., and Park, D. K. (2014). Mobile healthcare application with emr interoperability for diabetes patients. *Cluster Computing*, **17**(3), 871–880.

Kailanto, H., Hyvarinen, E., and Hyttinen, J. (2008a). Mobile ECG measurement and analysis system using mobile phone as the base station. In *Pervasive Computing Technologies for Healthcare, 2008. PervasiveHealth 2008. Second International Conference on*, pages 12–14.

Kailanto, H., Hyvarinen, E., and Hyttinen, J. (2008b). Mobile ECG measurement and analysis system using mobile phone as the base station. In *2nd International Conference on Pervasive Computing Technologies for Healthcare*, pages 12–14.

Katsaras, T., Milsis, A., Rizikari, M., Saoulis, N., Varoutaki, E., and Vontetsianos, A. (2011). The use of the healthwear wearable system in chronic patients' early hospital discharge: Control randomized clinical trial. In *Medical Information & Communication Technology (ISMICT), 2011 5th International Symposium on*, pages 143–146.

Keele, S. (2007). Guidelines for performing systematic literature reviews in software engineering. In *Technical report, Ver. 2.3 EBSE Technical Report. EBSE.*

Kleiger, R. E., Miller, J. P., Bigger, J. T., and Moss, A. J. (1987). Decreased heart rate variability and its association with increased mortality after acute myocardial infarction. *The American journal of cardiology*, **59**(4), 256–262.

Köhler, B.-U., Hennig, C., and Orglmeister, R. (2002). The principles of software QRS detection. *Engineering in Medicine and Biology Magazine, IEEE*, **21**(1), 42–57.

Krause, A., Ihmig, M., Rankin, E., Leong, D., Gupta, S., Siewiorek, D., Smailagic, A., Deisher, M., and Sengupta, U. (2005). Trading off prediction accuracy and power consumption for context-aware wearable computing. In *Wearable Computers, 2005. Proceedings. Ninth IEEE International Symposium on*, pages 20–26.

Krause, A., Smailagic, A., and Siewiorek, D. P. (2006). Context-aware mobile computing:

Learning context-dependent personal preferences from a wearable sensor array. *Mobile Computing, IEEE Transactions on*, **5**(2), 113–127.

Kunnath, A. T., Nadarajan, D., Mohan, M., and Ramesh, M. V. (2013). Wicard: A context aware wearable wireless sensor for cardiac monitoring. In *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*, pages 1097–1102.

Laguna, P., Jané, R., and Caminal, P. (1994). Automatic detection of wave boundaries in multilead ECG signals: validation with the CSE database. *Computers and Biomedical Research*, **27**(1), 45–60.

Le Gall, J.-R., Loirat, P., Alperovitch, A., Glaser, P., Granthil, C., Mathieu, D., Mercier, P., Thomas, R., and Villers, D. (1984). A simplified acute physiology score for ICU patients. *Critical Care Medicine*, **12**(11), 975–977.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, **521**(7553), 436–444.

Lee, K. H., Kung, S.-Y., and Verma, N. (2011). Improving kernel-energy trade-offs for machine learning in implantable and wearable biomedical applications. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1597–1600.

Leite, C., Sizilio, G., Neto, A., Valentim, R., and Guerreiro, A. (2011). A fuzzy model for processing and monitoring vital signs in icu patients. *BioMedical Engineering Online*, **10**, 68.

Li, Q., Mark, R. G., and Clifford, G. D. (2008). Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a kalman filter. *Physiological Measurement*, **29**(1), 15.

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*.

Liu, N., Lin, Z., Koh, Z., Huang, G.-B., Ser, W., and Ong, M. E. H. (2011). Patient outcome prediction with heart rate variability and vital signs. *Journal of Signal Processing Systems*, **64**(2), 265–278.

Luo, G. (2015). MLBCD: a machine learning tool for big clinical data. *Health Information Science and Systems*, **3**(1), 1–19.

Malliani, A., Pagani, M., Lombardi, F., and Cerutti, S. (1991). Cardiovascular neural regulation explored in the frequency domain. *Circulation*, **84**(2), 482–492.

Mancini, G. J., Gosselin, G., Chow, B., Kostuk, W., Stone, J., Yvorchuk, K. J., Abramson, B. L., Cartier, R., Huckell, V., Tardif, J.-C., *et al.* (2014). Canadian cardiovascular society guidelines for the diagnosis and management of stable ischemic heart disease. *Canadian Journal of Cardiology*, **30**(8), 837–849.

Martis, R. J., Acharya, U. R., and Adeli, H. (2014). Current methods in electrocardiogram characterization. *Computers in Biology and Medicine*, **48**, 133–149.

Mattila, J., Ding, H., Mattila, E., and Särelä, A. (2009). Mobile tools for home-based cardiac rehabilitation based on heart rate and movement activity analysis. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pages 6448–6452.

Mayton, B., Dublon, G., Palacios, S., and Paradiso, J. A. (2012). Truss: tracking risk with ubiquitous smart sensing. In *Sensors, 2012 IEEE*, pages 1–4.

Melillo, P., Izzo, R., Orrico, A., Scala, P., Attanasio, M., Mirra, M., De Luca, N., and Pecchia, L. (2015a). Automatic prediction of cardiovascular and cerebrovascular events using heart rate variability analysis. *PloS one*, **10**(3), e0118504.

Melillo, P., Izzo, R., Orrico, A., Scala, P., Attanasio, M., Mirra, M., De Luca, N., and Pecchia, L. (2015b). Automatic prediction of cardiovascular and cerebrovascular events using heart rate variability analysis. *PloS one*, **10**(3), e0118504.

Moody, G. B. and Mark, R. G. (2001). The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, **20**(3), 45–50.

Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., and Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *Neural Networks, IEEE Transactions on*, **12**(2), 181–201.

Murthy, V. K., Grove, T. M., Harvey, G. A., and Haywood, L. J. (1978). Clinical usefulness of ECG frequency spectrum analysis. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 610.

O'Brien, I., O'Hare, P., and Corrall, R. (1986). Heart rate variability in healthy subjects: effect of age and the derivation of normal ranges for tests of autonomic function. *British Heart Journal*, **55**(4), 348–354.

Okoli, C. (2015). A guide to conducting a standalone systematic literature review. *Communications of the Association for Information Systems*, **37**(1), 43.

Oresko, J. J., Jin, Z., Cheng, J., Huang, S., Sun, Y., Duschl, H., and Cheng, A. C. (2010). A wearable smartphone-based platform for real-time cardiovascular disease detection via electrocardiogram processing. *Information Technology in Biomedicine, IEEE Transactions on*, **14**(3), 734–740.

Owens, P., Atkins, N., and OBrien, E. (1999). Diagnosis of white coat hypertension by ambulatory blood pressure monitoring. *Hypertension*, **34**(2), 267–272.

Özkaraca, O., Işık, A. H., and Güler, İ. (2011). Detection, real time processing and monitoring of ecg signal with a wearable system. In *Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on*, pages 424–427.

Pan, J. and Tompkins, W. J. (1985). A real-time qrs detection algorithm. *Biomedical Engineering, IEEE Transactions on*, (3), 230–236.

Pandian, P., Mohanavelu, K., Safeer, K., Kotresh, T., Shakunthala, D., Gopal, P., and Padaki, V. (2008). Smart Vest: Wearable multi-parameter remote physiological monitoring system. *Medical Engineering & Physics*, **30**(4), 466–477.

Patel, S., Hughes, R., Hester, T., Stein, J., Akay, M., Dy, J. G., and Bonato, P. (2010). A novel approach to monitor rehabilitation outcomes in stroke survivors using wearable technology. *Proceedings of the IEEE*, **98**(3), 450–461.

Patel, S., Park, H., Bonato, P., Chan, L., and Rodgers, M. (2012). A review of wearable sensors and systems with application in rehabilitation. *Journal of Neuroengineering and Rehabilitation*, **9**(1), 1.

Prabhakara, M. and Kulkarni, V. (2014). Real time analysis of EEG signals on android application. In *Advances in Electronics, Computers and Communications (ICAECC), 2014 International Conference on*, pages 1–4.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raghavendra, B., Bera, D., Bopardikar, A. S., and Narayanan, R. (2011). Cardiac arrhythmia detection using dynamic time warping of ECG beats in e-healthcare systems. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2011 IEEE International Symposium on a*, pages 1–6.

Raj, S., Maurya, K., and Ray, K. C. (2015). A knowledge-based real time embedded platform for arrhythmia beat classification. *Biomedical Engineering Letters*, **5**(4), 271–280.

Regan, T. (2005). Current Sense Circuit Collection.

Rosenblatt, F. (1961). Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, DTIC Document.

Rosu, M.-C. (2014). Preliminary evaluation for an ecg monitoring system. In *Electronics, Computers and Artificial Intelligence (ECAI), 2014 6th International Conference on*, pages 73–80.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, **47**(3), 537–560.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**(3), 581–592.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, DTIC Document.

Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L.-W., Moody, G., Heldt, T., Kyaw, T. H., Moody, B., and Mark, R. G. (2011). Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical Care Medicine*, **39**(5), 952.

Saykrs, B. M. (1973). Analysis of heart rate variability. *Ergonomics*, **16**(1), 17–32.

Shih, D.-H., Chiang, H.-S., Lin, B., and Lin, S.-B. (2010). An embedded mobile ecg reasoning system for elderly patients. *Information Technology in Biomedicine, IEEE Transactions on*, **14**(3), 854–865.

Solar, H., Fernández, E., Tartarisco, G., Pioggia, G., Cvetković, B., Kozina, S., Luštrek, M., and Lampe, J. (2013). A non invasive, wearable sensor platform for multi-parametric remote monitoring in CHF patients. *Health and Technology*, **3**(2), 99–109.

Stone, A. A., Shiffman, S., Schwartz, J. E., Broderick, J. E., and Hufford, M. R. (2003). Patient compliance with paper and electronic diaries. *Controlled Clinical Trials*, **24**(2), 182–199.

Sufi, F., Fang, Q., Khalil, I., and Mahmoud, S. S. (2009). Novel methods of faster cardiovascular diagnosis in wireless telecardiology. *Selected Areas in Communications, IEEE Journal on*, **27**(4), 537–552.

Sun, F.-T., Kuo, C., Cheng, H.-T., Buthpitiya, S., Collins, P., and Griss, M. (2012). Activity-aware mental stress detection using physiological sensors. In *Mobile Computing, Applications, and Services*, pages 211–230.

Takata, K., Ma, J., Apduhan, B. O., Huang, R., and Jin, Q. (2008). Modeling and analyzing individual's daily activities using lifelog. In *Embedded Software and Systems, 2008. ICESS'08. International Conference on*, pages 503–510.

Tao, K. M. (1993). A closer look at the radial basis function networks. In *The Asilomar Conf. on Signals, Systems & Computers*, pages 401–405.

Torres-Huitzil, C. and Nuno-Maganda, M. (2015). Robust smartphone-based human activity recognition using a tri-axial accelerometer. In *Circuits & Systems (LASCAS), 2015 IEEE 6th Latin American Symposium on*, pages 1–4.

Tsien, C. L. and Fackler, J. C. (1997). Poor prognosis for existing monitors in the intensive care unit. *Critical Care Medicine*, **25**(4), 614–619.

Vapnik, V. N. and Chervonenkis, A. J. (1974). Theory of pattern recognition.

Vapnik, V. N. and Vapnik, V. (1998). *Statistical learning theory*, volume 1.

Villalba, E., Salvi, D., Ottaviano, M., Peinado, I., Arredondo, M. T., and Akay, A. (2009). Wearable and mobile system to manage remotely heart failure. *Information Technology in Biomedicine, IEEE Transactions on*, **13**(6), 990–996.

Wagstaff, D. A., Kranz, S., and Harel, O. (2009). A preliminary study of active compared with passive imputation of missing body mass index values among non-hispanic white youths. *The American Journal of Clinical Nutrition*, **89**(4), 1025–1030.

Wakabayashi, I. (2004). Relationships of body mass index with blood pressure and serum cholesterol concentrations at different ages. *Aging clinical and experimental research*, **16**(6), 461–466.

Wells, B. J., Nowacki, A. S., Chagin, K., and Kattan, M. W. (2013). Strategies for handling missing data in electronic health record derived data. *Generating Evidence & Methods to improve patient outcomes*, **1**(3), 7.

WHO *et al.* (1978). International classification of diseases:[9th] ninth revision, basic tabulation list with alphabetic index.

World Health Organization (2010). Burden: mortality, morbidity and risk factors. *Global Status Report on Noncommunicable Diseases*, **2011**.

Zanaty, E. (2012). Support vector machines (SVMs) versus multilayer perception (MLP) in data classification. *Egyptian Informatics Journal*, **13**(3), 177–183.

Zheng, Y.-L., Ding, X.-R., Poon, C. C. Y., Lo, B. P. L., Zhang, H., Zhou, X.-L., Yang, G.-Z., Zhao, N., and Zhang, Y.-T. (2014). Unobtrusive sensing and wearable devices for health informatics. *IEEE Transactions on Biomedical Engineering*, **61**(5), 1538–1554.

Zhu, M., Zhang, Z., Hirdes, J. P., and Stolee, P. (2007). Using machine learning algorithms to guide rehabilitation planning for home care clients. *BMC medical informatics and decision making*, **7**(1), 1.

Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry*, **39**(4), 561–577.

Appendix A

All Advanced Search Queries

Table A.1: Five popular academic databases were used with advanced search queries in the literature review

Query	Database	#Results	Selected
(data mining) AND (wearable OR wearable system) AND (mobile OR smartphone) AND (CVD OR cardiovascular disease) AND (patient monitoring OR monitoring) AND (ECG OR electrocardiogram) AND (Healthcare) AND (Telemedicine) AND (SVM OR support vector machine) AND (Biomedical) AND (Machine learning)	Google Scholar	136	18
"data mining" AND "wearable" OR "wearable system" AND "mobile" OR "smartphone" AND "cvd" OR "cardiovascular disease" AND "monitoring" AND "ecg" AND "Telemedicine" AND "svm" OR "support vector machine" AND "Machine learning" AND"Biomedical"	IEEE Xplore	488	33
(data mining) AND (wearable OR wearable system) AND (mobile OR smartphone) AND (cvd OR cardiovascular disease) AND (patient monitoring OR monitoring) AND (ecg OR electrocardiogram) AND (svm OR support vector machine) AND (Machine Learning) OR (Healthcare) OR (Telemedicine) OR (Biomedical)	Science Direct	14	3
"data mining" AND "wearable" OR "wearable system" AND "mobile" OR "smartphone" AND "cvd" OR "cardiovascular disease" AND "monitoring" AND "ecg" AND "Telemedicine" AND "svm" OR "support vector machine" AND"Machine learning" AND"Biomedical"	Springer Link	1566	12
"data mining" AND "wearable" OR "wearable system" AND "mobile" OR "smartphone" AND "cvd" OR "cardiovascular disease" AND "monitoring" AND "ecg" AND "Telemedicine" AND "svm" OR "support vector machine" AND"Machine learning" AND"Biomedical"	PubMed	154	1

Table A.2: Selection of Papers for Data Extraction

Category	Paper	Q_1	Q_2	Q_3	Q_4	Total
CAT3	(Bellos *et al.*, 2011b)	2	2	2	2	8
CAT3	(Bellos *et al.*, 2013, 2014)	2	2	2	2	8
CAT3	(Bellos *et al.*, 2010a, 2011b, 2012)	2	2	2	0	6
CAT3	(Solar *et al.*, 2013)	2	2	2	0	6
CAT1	(Gultepe *et al.*, 2014)	2	0	2	2	6
CAT1	(Guidi *et al.*, 2014a)	2	0	2	2	6
CAT2	(Clifton *et al.*, 2014)	2	0	2	1	5
CAT2	(Alshurafa *et al.*, 2014)	2	1	2	0	5
CAT3	(Krause *et al.*, 2006)	2	1	2	0	5
CAT3	(Oresko *et al.*, 2010)	2	2	0	0	4
CAT1	(Liu *et al.*, 2011)	2	0	2	0	4
CAT1	(Chen *et al.*, 2005)	0	2	2	0	4
CAT3	(Torres-Huitzil and Nuno-Maganda, 2015)	2	2	0	0	4
CAT2	(Luo, 2015)	2	0	0	2	4
CAT3	(Juen *et al.*, 2015)	2	2	0	0	4
CAT3	(Kunnath *et al.*, 2013)	0	2	2	0	4
CAT3	(Raghavendra *et al.*, 2011)	2	2	0	0	4
CAT1	(Leite *et al.*, 2011)	2	0	2	0	4
CAT1	(Katsaras *et al.*, 2011)	0	2	2	0	4
CAT2	(Clifton *et al.*, 2011)	2	0	2	0	4
CAT1	(Bellos *et al.*, 2010b)	2	0	0	2	4
CAT1	(Villalba *et al.*, 2009)	0	2	2	0	4
CAT1	(Takata *et al.*, 2008)	0	2	2	0	4
CAT3	(Krause *et al.*, 2005)	2	2	0	0	4
CAT3	(Melillo *et al.*, 2015a)	2	0	2	0	4
CAT2	(Shih *et al.*, 2010)	2	0	0	0	2
CAT1	(Pandian *et al.*, 2008)	0	0	2	0	2
CAT1	(Kailanto *et al.*, 2008a)	0	2	0	0	2
CAT1	(Anliker *et al.*, 2004)	0	0	2	0	2
CAT1	(Prabhakara and Kulkarni, 2014)	0	2	0	0	2
CAT1	(Depari *et al.*, 2014)	0	2	0	0	2
CAT1	(Jung *et al.*, 2014)	0	2	0	0	2
CAT1	(Gao *et al.*, 2013)	0	2	0	0	2
CAT1	(Mayton *et al.*, 2012)	0	0	2	0	2
CAT1	(Baron *et al.*, 2011)	0	0	2	0	2
CAT1	(Patel *et al.*, 2012)	2	0	0	0	2
CAT1	(Zhu *et al.*, 2007)	0	2	0	0	2
CAT1	(Özkaraca *et al.*, 2011)	0	0	0	0	0
CAT2	(Raj *et al.*, 2015)	0	0	0	0	0
CAT1	(Rosu, 2014)	0	0	0	0	0

Appendix B

Circuit Diagrams

B.1 ECG Acquisition Circuit

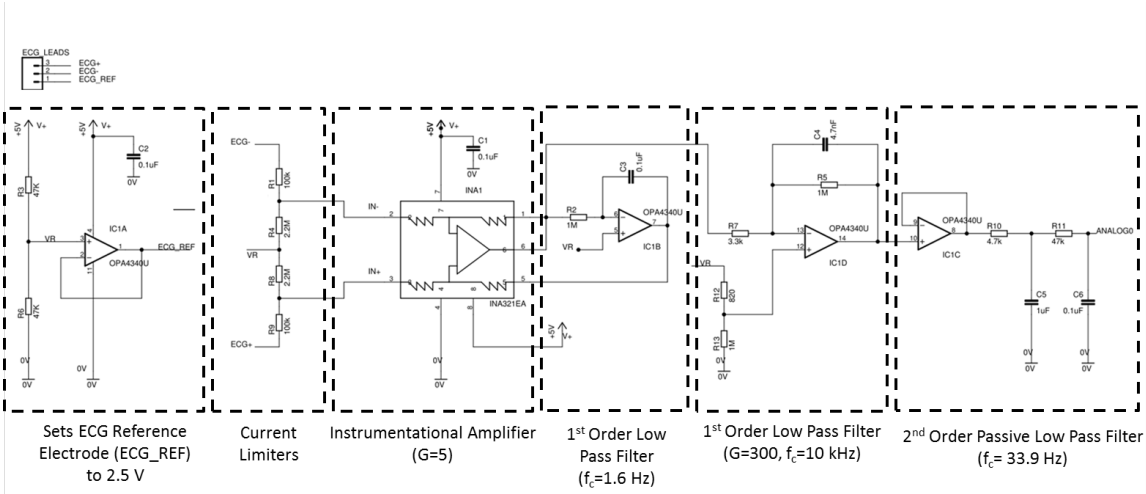


Figure B.1: The circuit used to acquire, amplify and filter the patient's ECG signal (Libelium, 2015).

B.2 Raspberry Pi Shield Configuration

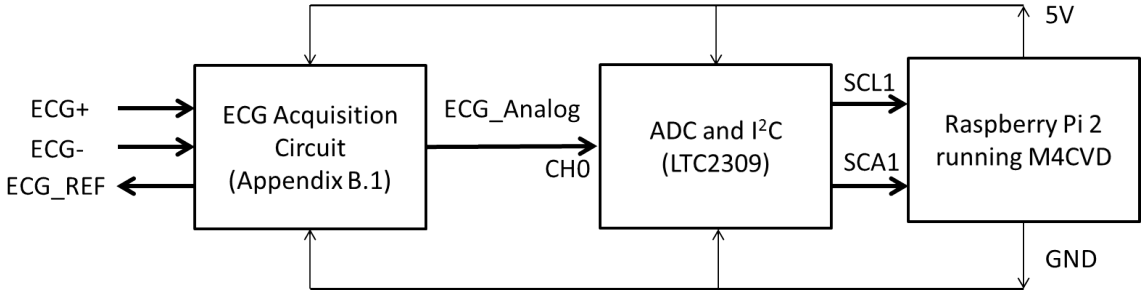


Figure B.2: The Raspberry Pi 2 runs M4CVD and provides power and ground to the ECG Acquisition Circuit and ADC.

B.3 BP Acquisition Circuit

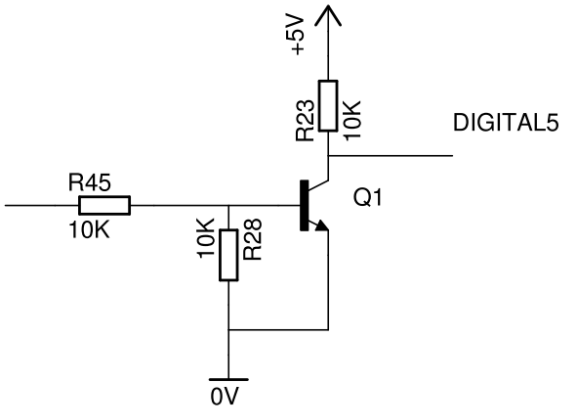


Figure B.3: The circuit used to transmit readings from the commerical BP monitor to the Raspberry Pi (Libelium, 2015).

Appendix C

Input Features Studied in Thesis

Table C.1: 24 features were initially considered as inputs for M4CVD. 11 features were successfully implemented and validated. 13 features were not successfully validated

Feature	Studied in this thesis?	Reason for not studying the feature (if applicable)
1. Age	Y	
2. Gender	Y	
3. Body Mass Index	Y	
4-5. Systolic\Diastolic Blood Pressure	Y	
6. Heart Rate	Y	
7. Mean R-R Interval	Y	
8. Heart Rate Varability	Y	
9. Standard Deviation of R-R (SDNN)	Y	
10. Square Root of Mean Difference of R-R (rMSSD)	Y	
11. Percentage of R-R interval greater than 50 ms (pNN50)	Y	
12-14. Cholesterol (Total, High / Low Density Lipoproteins)	N	Initially examined in (Boursalie *et al.*, 2015) but further research has shown that cholesterol has high correlation to BMI (Wakabayashi, 2004) resulting in cholesterol being excluded from M4CVD.
15-17. PR, QT and QRS points	N	
18. QRS wave interval	N	Validation of ECG peak detection was successful on MIT-BIH database but unsuccessful on the MIMIC II ECG signals.
19. Q wave amplitude	N	
20. Q-T interval	N	
21. Power in normalized low frequency band (0.04-0.15)	N	
22. Power in normalized high frequency band (0.15-0.4)	N	The MIMIC II database decimated its ECG signals to save storage space (Saeed *et al.*, 2011). The decimation process destroyed the frequency component of the ECG signal so frequency features could not be used as inputs for M4CVD.
23. Total Power in normalized band (0.04-0.4)	N	
24. Power ratio between low and high normalized power band	N	