Self-Coach: An Intelligent WBAN System for Heart Disease Prediction Using Non-Dominated Sorting Genetic Algorithm

by

© *Babak Emami-Abarghouei*

A thesis submitted to the

School of Graduate Studies

in partial fulfilment of the

requirements for the degree of

Master of *Science*

Department of *Computer Science*

Memorial University of Newfoundland

September 2016

St. John's Newfoundland

To my Beloved Mother

Abstract

Abstract Wireless Body Area Network (WBAN) is a new technology based on an advanced healthcare system and Wireless Sensor Network (WSN). This domain has been designed for monitoring patients using their physical signals by providing low-cost, wearable, unobtrusive solutions for the continuous monitoring of cardiovascular health and physical activity status. Recent studies have addressed the use of WBAN, by means of real-time comprehensive health monitoring systems such as the Personal Health Monitoring system (PHM). The aim of utilizing these systems is to provide a fast and early diagnosis; however, WBAN has not fulfilled its potential while using the existing methods of machine learning and data mining techniques. The proposed method is a new framework for early heart failure prediction systems based on an intelligent WBAN named Self-Coach. Self-Coach is an intelligent monitoring system due to the detection/prediction method it uses, which provides real-time health status monitoring using the Non-Dominated Sorting Genetic Algorithm.

In this study, Self-Coach is proposed as a new medical diagnosis method based on the hybrid approach. This approach applies the Support Vector Machine (SVM) to classification, where its parameter values are optimized and visualized by the Non-Dominated Sorting Genetic Algorithm-II (NSGA-II). To predict a potential health tolerance threshold for a particular patient, the optimal boundary curve between both the healthy and unhealthy class and the patient's possible health positions based on new conditions (blood pressure and heart rate) have been explored with NSGA-II.

The optimal boundary is a set of non-dominated offspring from the unhealthy class that has been generated with NSGA-II and verified with the SVM. These members are plotted as a Pareto Front curve known as the optimal boundary curve. Based on the SVM classification results, this is the most fitted curve which can separate the healthy class from the unhealthy.

To explore the potential health position of a particular patient, new possible positions will be generated with NSGA-II. To generate these points, the patient's dynamic data (blood pressure and heart rate) will be increased and simulated utilizing NSGA-II.

The potential tolerance threshold for each patient is the new health position for that particular patient based on the new health conditions which will cross the optimal boundary or be dominated by at least one of these members.

To evaluate the Self-Coach's performance, its experimental results (optimal boundary members and explored tolerance thresholds for each individual) have been verified using SVM and compared with the raw dataset classification results. Based on the simulated results, this method can provide 24-hour health monitoring care for elderly people and those who might have coronary heart disease. It is a new technology for this domain. Real-time data analysis is a significant part of Self-Coach, which makes it a good candidate for supporting a broad array of high-impact applications in the domain of the healthcare system, for training, rehabilitation, surgical recovery and as a home-based monitoring healthcare system.

— MUN School of Graduate Studies

Acknowledgements

First and foremost, I would like to express my deepest appreciation to my dear supervisor, Dr. Saeed Samet , for his support, encouragement, and invaluable comments. The completion of this master thesis would have not been possible without his supervision.

Special thanks to all e-health staff: Marian, Ann, Donald and Dr.Gerard Farrell for their support, encouragement, guidance. My kind regards to my dear friend Javad Rahimipour for his support, suggestions and patience. Much thanks to my dearest friends, Ali Farrokhtala, Amirali Daroudi, Ali Meghdadi, Khalil Eslamloo and Majid Afshar for their kindness, and for always cheering me up.

I take this opportunity to thank my dear sister Darya and my brother Pouyan for their support in all stages of my life. Last but not the least, my heartfelt appreciation goes to my beloved mother, Shahin, for her endless love and support through my life. I dedicate this thesis to her with all my love.

— MUN School of Graduate Studies

Contents

List of Tables

List of Figures

List of Abbreviation

- **American Civil War:** year 1861-5, the casualty lists, request for medical supplies, and X-ray images were transmitted over the telegraph.

- **Remote Patient Monitoring system (RPM):** A modern telemedicine system which is designed to reduce the cost, increase the efficient utilization of physician skills, and provide remote access to patients for continuous monitoring and real-time analysis of the patients' health information.

- **Wearable mobile health system:** A tiny medical sensors attached to the Wireless Sensor Network **(WSN)**. This service takes advantage of WSN to increase the efficiency of communication and is called the Wireless Body Area Network **(WBAN)**.

- **Wireless Body Area Network (WBAN):** It is an enabling technology, which has been proposed to support early detection of abnormal conditions and prevention of severe consequences.

- **Body Sensors Network (BSN):** Medical sensors around the body to capture the physical body signals.

- **Mobile Computing Center (MCC):** WBAN Computing center, transfer signals to data, record the incoming data, and data processing.

- **Medical Center (MC):** Medical data server, Physicians, and emergency contacts.

- **Support Vector Machine (SVM):** Is a well-known supervised machine learning models with associated learning algorithms. It has proposed to data analysis using classification and regression analysis technique.

- **Evolutionary Algorithms (EA):** Is a subset of evolutionary computation, a generic population-based meta-heuristic optimization algorithm.

- **Multi-objective Optimization Problem (MOP):** This approach has been widely accepted as a real world problem-solving application. Identification of the Pareto-optimal solutions (Pareto Front) is a significant use of these algorithms for problem-solving. This is a function to recognize a vector of satisfiable condition based on the mathematical description to show the accuracy of the function's performance.

- **Multi-objectives Optimization using Evolutionary Algorithm (MOEA):** Optimization algorithms exploit EA to speed up this identification process.

- **Pareto Front (PF):** By definition, Pareto Front describes a set of alternatives where there is no other solution/alternative that dominates it. In our case, the Pareto Front will be the optimal boundary between unhealthy and healthy class which contains a set of points that all healthy patients dominates it and it dominates all unhealthy data.

- **Crowding Distance (CD):** It is the Euclidian distance which will be calculated for each individual separately based their objectives on each Pareto graph.

- **Genotype:** The way to define the classs features in GA. Here is the patients medical features related to their heart health status

- **Phenotype:** The way to define the class label in GA. Here is denoted to the healthy or unhealthy class. The effect of the genotype (sample's features) will be shown by its phenotype (sample's class label).

- **Parent:** Those members who selected for generates new solution.

- **Children, Off-spring, new Solution:** New possible member, will be created from two parents using crossover operation or single parent using mutation function. These new members have not existed in the dataset, but technically they have their parents genome, and they can be existing. The phenotype of these members has been evaluated with the SVM.

- **Cost Function (CF):** the optimization functions to map multi features to two features. Here, the dataset's features have shown with to features or fitness function. F1 or X-axis is on feature alone here is heart rate per minute. F2 or Y-axis, a trade-off between all features including heartbeat feature.

- **Cost Position, Cost Health Position, Health Position:** is a graph to visualize the effect of solutions cost value.

- **Coronary Heart Disease (CHD):** Also known as ischemic heart disease (IHD), It is a group of diseases that includes: stable angina, unstable angina, myocardial infarction. The person at risk of this disease might have an experience of chest pain (a discomfort which may travel into the shoulder, arm, back, neck, or jaw) or heartburn.

Chapter 1

Introduction

1.1 Introduction

A rapid increase in the world population of elderly people has drawn researchers' attention to improve the healthcare system [54]. Within the next decade, in the United States alone, 76 million baby boomers (a generation born between the years 1946 to 1964) will reach retirement age. Current healthcare systems are not organized to have adequate services for such an aging population [62].To solve this problem, telemedicine systems were proposed in the early 19^{th} century [55]. The way which United States used this technology for the American Civil War was fascinating. During the American Civil War 1861-5 the casualty lists, request for medical supplies, and X-ray images were transmitted over the telegraph. Later in 1920, the United States' Seaman's Church provided medical care using radio signal [14].

Despite the fact that it was the first time that telemedicine took advantage of telecommunication technology, telemedicine has previously been used in cardiology clinical applications during the 1920-1940s in medical centers in Norway, Italy and France. The use of telemedicine in cardiology was later followed by radio consultation with medical centers in the 1920s, 1930s and 1940s for the patients aboard ships at sea as well as remote islands with radiology healthcare applications [57]. In 1983, remote diagnoses for dermatology, cardiology, pathology, radiology and endoscopy were utilized by University Hospital of North Norway Tromso in order to provide a long distance clinical healthcare using modern telecommunication systems [42].

Indeed, the modern telemedicine systems are a collection of wireless sensor nodes and mobile applications placed around or in a human body that are used to compute and exchange the body health information to the medical center immediately[54]. This provides a potential opportunity for the healthcare monitoring systems to involve a portable administrative monitoring system [37]. The use of a modern telemedicine system was proposed to reduce the cost, increase the efficient utilization of physician skills and provide remote access to patients for continuous monitoring and real-time analysis of the patients' health information based on Remote Patient Monitoring system (RPM) [54]. RPM is a section of telemedicine that includes diagnosis, monitoring, treatment and education, with the lowest healthcare costs, which has been designed to track a patient health status at long distance [37].

The telehealth system and wireless communication in RPM allow physicians to have a remote interface to collect and transmit patient's data to the medical center [37]. In the last decade, a variety of disease statuses, including cardiac care, diabetes, pulmonary disease, pharmaceutical compliance, co-morbidities and mental health, have been monitored by such technology [37]. The advancement in wireless communication has had a huge impact on the Wireless Sensor Network (WSN) for RPM.

Tiny and wearable wireless sensors such as AMON [5] or Medical Belt [45] have become widely used in the modern medical healthcare system and allow physicians to have access to a comprehensive telemonitoring system [53, 37]. By integrating a mobile device and a medical sensor such as the Live-Net [62] or MyHeart [30, 27] applications, it is possible to access patients' health information in any environment [53]. Since the last decade, the use of this technology has been widely implemented in the healthcare industry, developed based on the integration of wearable medical sensors, wireless sensors and mobile devices, and is known as a wearable mobile health system [3] which is integrated into the telemedicine system and e-healthcare as novel information technology [63]. This service utilizes WSN to increase the efficiency of communication, and is called the Wireless Body Area Network (WBAN)[37]. It is an enabling technology which has been proposed to support the early detection of abnormal conditions and prevention of severe consequences [68]. Self-Coach, which is proposed in this thesis, is a novel WBAN application designed for a telemonitoring healthcare system which provides active, reliable and 24-hour health monitoring for elderly and remote patients who have a heart problem. This application can predict the patient's actual health status based on their medical records (dataset history) and activities. Indeed, Self-Coach proposes a novel data processing method using a Genetic Algorithm which can suggest a health tolerance threshold for a patient in a healthy class who might have a chest pain experience. This methodology is widely discussed in Chapter 4 and 5.

1.1.1 Background of the Problem

WBAN is a system composed of small off-the-shelf, tiny sensor platforms with application-specific signal conditioning modules, which can be implemented on either inside or around the patient's body [18]. These sensors are used for collecting and monitoring the vital signs of patient's activities [16]. WBAN is defined as a set of autonomous integrated multiple sensors with the ability to provide processing capabilities that may reduce the amount of transmitted data to minimize the cost [68]. This application has been described as an integrated system, constructed based on four units: the Body Sensors Network (BSN), Mobile Computing Center (MCC), Medical Center, and Communication aspect [51]. BSN is the part of WBAN that is constructed of wearable or implantable bio-sensors to capture the patient's body signals. The captured data is transferred to the base station (MCC) using Bluetooth, wi-fi, or a cable line. MCC is the WBAN base station that aggregates and processes the signals. Real-time analysis and data processing are implemented in this unit. The processed data is uploaded on a medical server site for evaluation and analysis by physicians [51, 22]. Figure 3.1 shows a quick view of the WBAN's units.

WBAN utilizes data mining techniques to reduce the data processing time cost. Data mining and machine learning algorithms are the most significant applications that have been adopted for data processing on electronic datasets [53]. These techniques have been applied to explore the hidden relations between data features. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in a large dataset [18]. The study of machine learning and data-mining techniques allows researchers to discover useful relationships between unknown features [53].

In existing health monitoring systems, the system needs to have the basic information on the subject of study to learn various situations and make an accurate decision in similar conditions [60]. However, a lack of complete information in many engineering problems leads to uncertainty in decision making[35].

While real-time data processing has been proposed to increase the WBAN classification's accuracy, using conventional data classification methods such as k-nearest neighbor, support vector machine (SVM), and decision tree is still very time-consuming. The accuracy of decision making and health-class-label's prediction for a particular patient would not be very significant regarding the lack of datasets' information [60].

To overcome the above-mentioned issues, this thesis proposes the use of the Non-Dominated Sorting Genetic Algorithm (NSGA-II) for real-time monitoring systems. The proposed system is able to predict the patient's health status based on their medical records, current activity, blood pressure and the momentary heart rate.

1.2 Problem Statement

Classification and data processing are satisfied in a polynomial time with an order of 3 $O(m(n)^3)$, where (m) represents the number of objectives and (n) is the number of records [3]. While (m) has been reduced from 13 features (Cleveland) [4] to 6 using genetic algorithms [16] to speed up the processing time, using standard methods such as the decision tree or the Naive Bayes for data classification still requires a polynomial time [68]. Due to this high time complexity, the WBAN system has not been used as a prediction system [62].

Applying feature selection can speed up the classification process, but at the same time, the prediction accuracy decreases due to removing some critical features such as age, gender, and current health status. Consequently, since WBAN has been proposed as an autonomous healthcare system for monitoring aging and remote patients, removing some features such as age and gender will have an indirect effect on the final health phenotype and the accuracy of prediction will decrease [51].

Furthermore, most clinical datasets contain insufficient information due to their data gaps, which also can affect decision making. The proposed Method can be used to solve this issue by filling these gaps and preparing data for decision making, while at the same time, preserving all necessary dataset features. This Method has been applied only on the (Cleveland) [4]. However, with some modifications, it will be compatible for various datasets.

1.3 Objectives of the Study

- Review the existing healthcare and telemonitoring system frameworks.
- Explore the body features which have an effect on critical situations based on the relationship between the patient's features and significant aspects.
- Explore the relationships between human body signals, the patients' medical history, current activity and their heart's health status.

- Investigate the existing classifier algorithm.
- Increase the accuracy of decision making by improving the dataset gap using NSGA-II.
- Propose a new application based on a real-time data classifier algorithm to predict actual cardiovascular problems.

1.4 Contribution of the Study

Self-Coach is a new application designed for WBAN to provide 24-hour health monitoring care for elderly people and those who might have Coronary Heart Disease. Real-time data analysis is a significant part of this application so risky health conditions can be predicted ahead of time. As this method is a new add-on application for wearable devices, a basic understanding of wearable devices would be helpful. Thus, wearable devices are discussed in Chapter 2 to provide this background information.

In existing health monitoring systems, the system needs to have basic information on the subject of study to learn various situations and make an accurate decision in similar conditions [60]. However, a lack of complete information in many engineering problems leads to uncertainty in decision making[35].

The available datasets, such the Cleveland [4], include a number of patients with certain measured features. Each patient is labeled with a normal or abnormal health status. However. in these datasets there is no real case to measure the critical points (tolerance threshold) for individuals and raw data sets do not provide a good training set for data mining techniques to predict the tolerance threshold for a new patient.

A number of health graphs can be used to describe the dataset health graphs based on the aforementioned features. As an example, in Figure 1.1 the health graph for our dataset is presented. In this graph, the X-axis represents heart rate and the Y-axis is a trade off of all dataset features, (its defining function is discussed in Chapter 4). Healthy patients are marked with a green '*' sign, while unhealthy ones are marked with a red '+' sign.

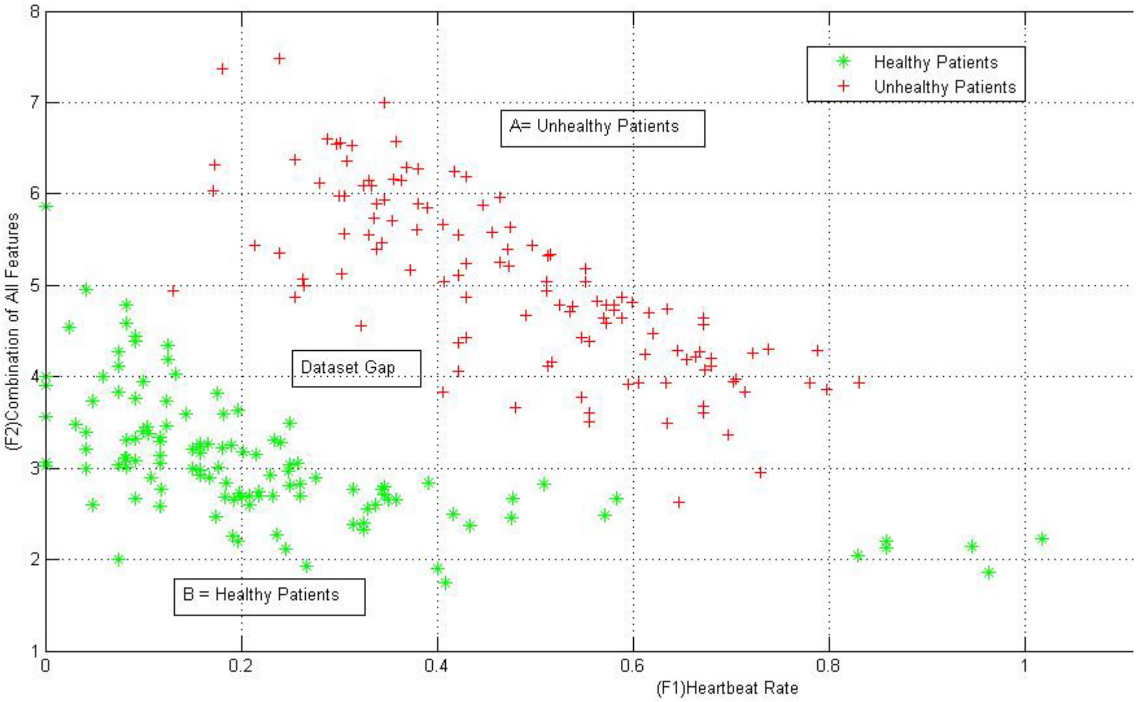


Figure 1.1: Health position graph

As can be seen in Figure 1.1 unhealthy patients are shown in the top right section of the graph, while healthy patients are listed in the bottom corner section. However, the data set is inconsistent, so there will not be a visible threshold to decide whether or not a new patient should be categorized as healthy or unhealthy. Besides for an existing patient in the dataset, some features are subject to change, thus the patient's health status might change in certain conditions but it is unknown when they will be moved to a different class. In our thesis, we aim to answer the two above questions.

In [35], to fill the dataset gap and increase the certainty of decision-making, a hybrid fuzzy-evolutionary algorithm, named fuzzy-evidential hybrid inference engine, is proposed. To pre-analyze the tolerance thresholds for each patient, we faced a similar problem, and in order to improve the dataset gap, Genetic Algorithm (GA) is used to generate new possible patients in both classes to find the optimal boundary curve (Pareto Front) to set the border between the two classes. In the following sections, our approach is described in greater detail.

1.4.1 Improve the dataset's gap using NSGA-II

We can find the desired boundary by finding the Pareto Front (Pareto Boundary) of the dataset. By definition, Pareto Front describes a set of alternatives where there is no other solution/alternative that dominates it. In our case, the Pareto Front will be the optimal boundary between unhealthy and healthy classes which contains a set of points where all healthy patients dominate it and it dominates all unhealthy data. However, prior to finding this optimal boundary, the dataset needs to be improved; thus, new possible patients are generated in both classes to fill the gaps in the dataset. This thesis used NSGA-II to generate new possible patients and explore all possible health positions. The details of NSGA-II are discussed in depth in Chapter 4.

The unhealthy class is selected to generate the new simulated health positions. Our goal is to find the optimal boundary between the two classes, which can be achieved by finding the best possible unhealthy positions that dominate all unhealthy data and be dominated by all healthy data. To get more accurate data, the algorithm NSGA-II is applied in several iterations. During each iteration, a new set of points, called the Pareto Front, is generated. These new points, along with the raw dataset are presented in Figure 1.2.

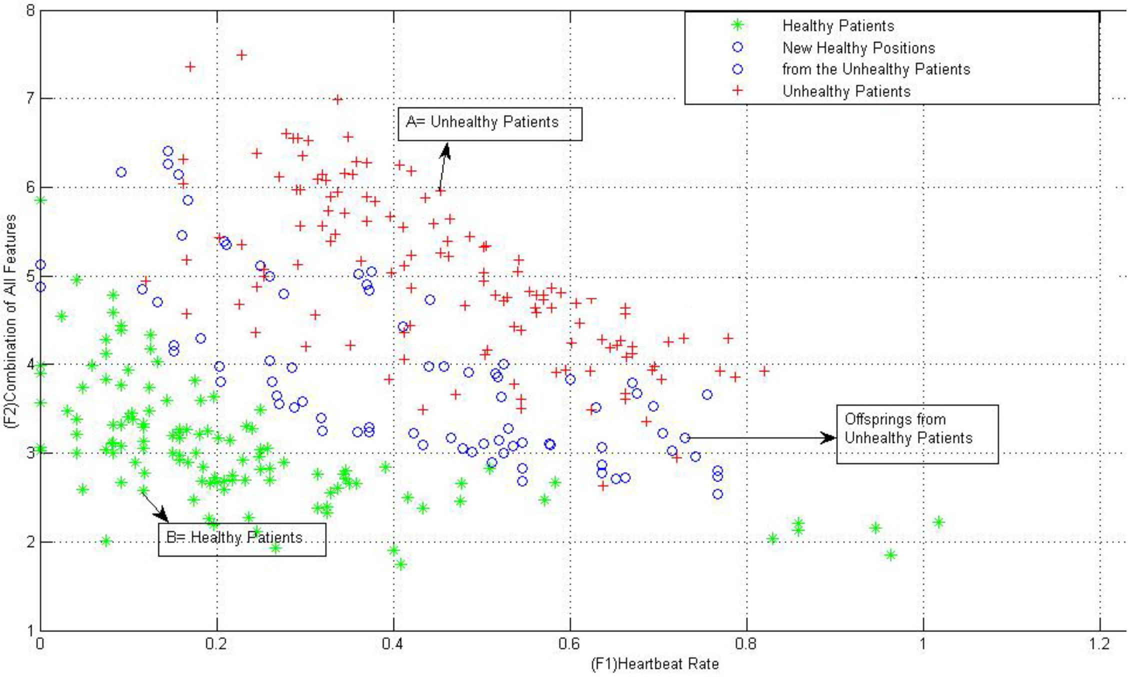


Figure 1.2: Generates new possible non-dominated unhealthy patients

1.4.2 Optimal Boundary Extraction

Now what needs to be done is to label the new health positions as healthy or unhealthy. To address this question, the SVM algorithm is used. Using SVM, it is required to introduce a training data set to learn and a test dataset to evaluate. In our case, the training set is the raw data set which is the ground truth, and the simulated data is evaluated as the test set. The optimal boundary then will be the set of points which have been labeled as unhealthy and dominate all other unhealthy points. In figure 1.3 the calculated optimal boundary is presented.

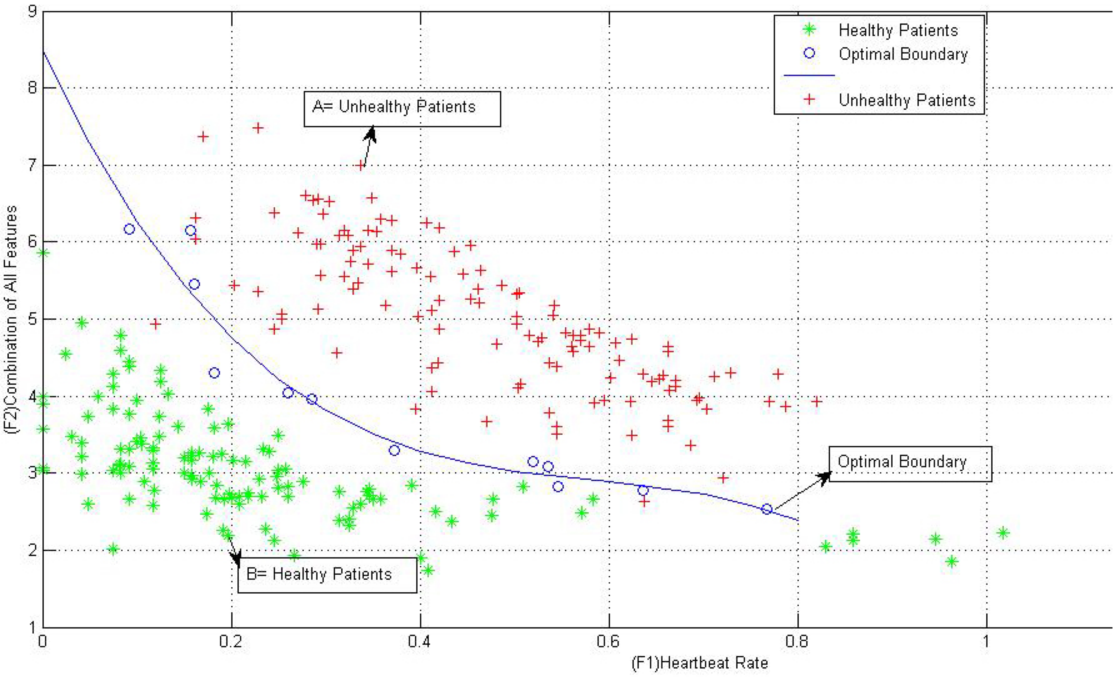


Figure 1.3: Exploring the most fitted Pareto Front curve

1.4.3 Tolerance Threshold Exploration for a Particular Patient

Patients' statuses are subject to change based on the new conditions. So under specific conditions any healthy patient can experience critical situations and be re-categorized in the unhealthy class. A tolerance threshold, as discussed in this thesis, is a new health position for the patient representing the new health status which would be dominated by at least one point in the optimal boundary. A change in any feature results in a new health position.

Technically, only some features can be modified based on daily conditions such as activity, using medicine, stress or emotional level. The features that do not change suddenly such as age, sex, and weight have been categorized as static features. On the other side, we have dynamic features, which include features with flexible curves due to the patient's daily activity. Heart rate, blood pressure, and blood sugar are categorized in this class.

Tolerance threshold points for each patient are points which show the patient's health status and could be categorized in the unhealthy class based on the patient's new condition. Indeed, the patient might experience chest pain or heart failure if its health position is dominated by one of these non-dominated points in the optimal boundary curve. In other words, by increasing the patient's dynamic features, their health position will tend to the right side of the graph. The first point which is dominated with the optimal boundary points will be recorded as the patient's tolerance threshold.

Following the dataset limitations, we did not have enough features for the study, and some features such as skin temperature, stress, respiration rate, sweating rate the patient's activity, and their medicine usage have not been mentioned in our dataset. Thus, the only available dynamic data are on heart rate and blood pressure. These features have been used in NSGA-II to generate new possible health positions for each particular patient. These two factors are not only subject to sudden changes but also have been also found to be critical features for heart disease problems [23]. The effect of these features on heart disease has been analyzed in [66]. The author also mentioned the importance of reducing blood pressure in risky situations, regardless of the use of anti-hypertensive medication [23].

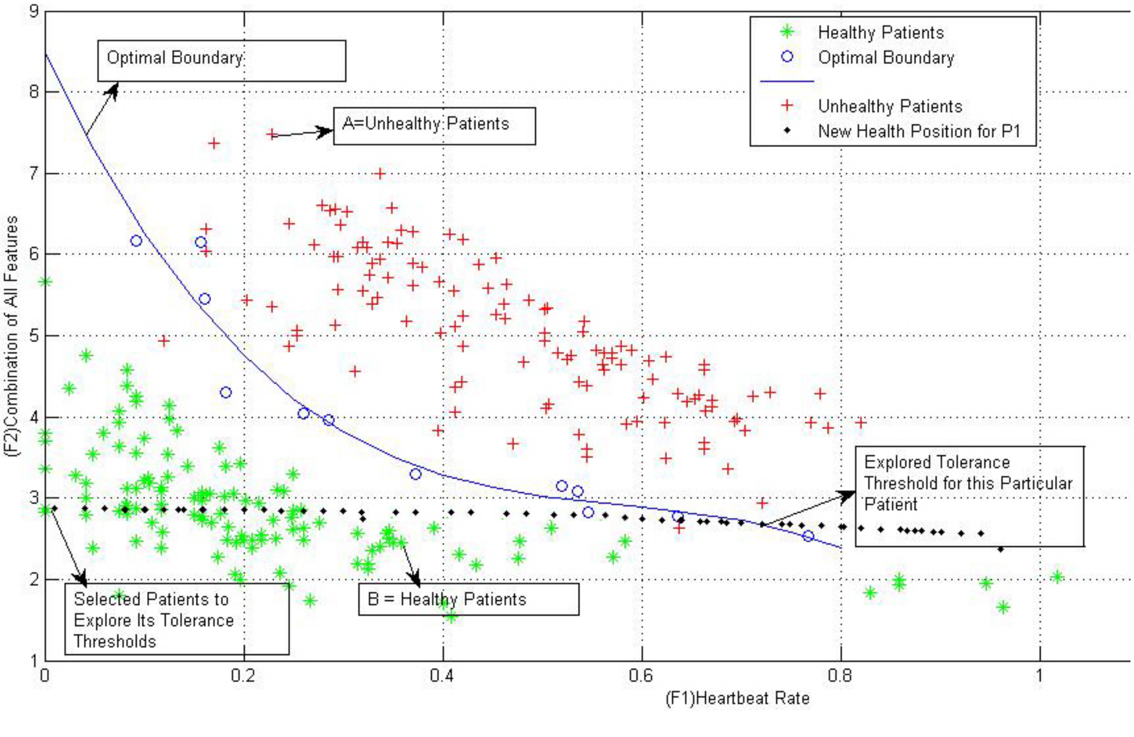


Figure 1.4: Tolerance threshold exploration for particular patient

An example of tolerance threshold exploration for a particular patient is shown in Figure 1.4. In this graph, the blue line describes a change in the health position of a particular patient by increasing their heart rate. For this graph, the patient's blood pressure is not subject to change. The tolerance threshold is where the blue line crosses the optimal boundary (black curve) which means the point at which the patient's health position is dominated by the optimal boundary. In our study, the number of the tolerance threshold is calculated for each patient based on various blood pressure measurements . This subject is explored in-depth in Chapter 5.

In each set of graphs, one of the dynamic data (Heart Rate) in the X-axis is increased from the starting point to 1.0 to simulate all possible positions due to an increase in the dynamic data, then the second one (Maximum Blood Pressure) is increased one point in the Y-axis and all possible positions are simulated again. This step has been simulated 10 times for each patient to explore all possible tolerance threshold points based on their blood pressure from 0 to 1; (8 to 20 bpm). Table 3.1 shows a quick view of the data set features and the normalization process.

1.5 Scope of the Study

The scope of this study can be defined as follows:

- Analyze and process the medical dataset related to heart disease (Cleveland) [4] with the NSGA-II algorithm and SVM.
- Use WEKA as the software for data mining and MatLab as a programming language for data analysis and to implement the proposed method.

- In this study, all data are simulated with NSGA-II to explore the patient's health status positions, and all simulations are done using MatLab.
- The Cleveland dataset, which has been used in this thesis, is a standard dataset that was published in 1998 for heart disease classification and future studies in this domain. Table 1.1 shows the Cleveland dataset's features that are used in this thesis. Appendix-A Shows the details of these features.

Table 1.1: Dataset features [4]

Feature	Value
Age	20 to 100
Sex	F, M
Chest Pain (CP)	angina, abnang , notang, asympt
Resting Blood Pressure (BPS)	90 to 200
Cholesterol (chol)	120 to 560
Fasting Blood sugar (fbs)	$< 120 =$ (true or false)
Resting ECG (restecg)	Normal , Abnormal , High
Max Heart Rate (HR)	50 to 190
Exercise induced angina (exang)	True , False
OldPeak	0.4 to 4.4
Slope	1,2,3
Number of vessels colored (ca)	0,1,2,3
Thal	normal, fixed, reverse
Class label	Healthy , Unhealthy

Chapter 2

Literature review

This chapter briefly reviews the use of the Wireless Body Area Network (WBAN) as an interactive application of healthcare system. WBAN has several objectives such as design, using bio-medical sensors, sensors' energy-efficiency, mobile application, wireless communication, and information security, but with regards to the thesis scope, the mobile application for WBAN has been explored in this chapter.

This study is followed by a review of applicable machine learning techniques such as SVM and decision tree. It concludes with an exploration of the optimization approach.

2.1 Wireless Body Area Network

A rapid increase in the world population of elderly people has motivated researchers to improve the healthcare system [54]. Since traditional healthcare systems are not organized to provide adequate service for an aging population [62], the use of telemedicine systems were proposed in the 19^{th} century [55]. The aim of this service is to reduce the cost and increase the efficient utilization of physician skills, providing a remote access to patients for continuous monitoring and real-time analysis of the patient's feedback based on Remote Patient Monitoring (RPM) [54]. To increase the efficiency of these applications, the RPM has been introduced as a section of telemedicine to treat distant patients [37].

Wearable Health Monitoring System (WHMS) is a branch of the RPM applications which takes advantages of wearable biosensors to reduce the cost of healthcare and improve the efficiency of monitoring systems [68]. For the last two decades, using WHMS has been widely accepted in the global healthcare system. WHMS has been used as an autonomous healthcare system which provides a comprehensive home health status monitoring system [3].

To address this demand a variety of system prototypes and commercial products have been proposed to provide real-time feedback for patients' health status. Also, the use of WHMS has been addressed to manage and monitor the issues of chronic diseases, elderly people's healthcare, rehabilitation system, and patients with disabilities [51, 8]. According to World Health Organization (WHO) report in [28], the global number of available physicians per 1000 citizens is roughly around 1.4. However, this ratio has been raised only a little in developed countries to approximately 2 to 4 healthcare providers per 1000 inhabitants. With the anticipated rise in population the healthcare system will be faced with big challenges such as provision of recovery/waiting room or improvised first care centers. To address this problem WBAN has been proposed as a new interactive WHMS application to provide a remote healthcare and comprehensive monitoring health-status system [68].

2.1.1 WBAN Objectives

In terms of healthcare settings, WBAN has been proposed to monitor, collect and process patient's phisical body signals and provide a digital medical records. WBAN is a collection of wireless medical sensor nodes and mobile applications placed around or in a human body to capture, compute and deliver the body signals information to the medical center [54]. Using mobile devices and wearable sensors provides a comprehensive telemonitoring system [37], since patients' body-signals are assessable in any environment [53]. The significance of using WBAN can be explored in three different areas [22]:

1. Ease of data collection with a comfortable interface
2. Scalability to support a majority of inhabitants
3. Real-time monitoring, processing and estimation to improve physician assessments

By utilizing WBAN applications, data collection is not limited to medical centers, since patients' data can be captured at home [57]. Using WBAN, patient's signals are recorded all day,in the attempt to enhance the clinical decision making [68].

Table 2.1: Biosensors implemented in WBAN [32]

ID	Wearable	Implantable	Description
EEG	headphone		Electroencephalogram
ECG	Wearable		Electrocardiography
EMG	Wearable		Electromyography
HF	Wearable		Heat Flux
HS	Wearable		Heart Sounds
HR	Wearable	Cochlear Implants	Heart Rate
BP	Wearable		Blood Pressure
SpO2	Wearable		Plus oximetry
A	Wearable		Activity
T	Wearable		temperature
GSR	Wearable		Galvanic Skin Response
GS	Wearable	Implantable	Glucose Sensor

2.1.2 WBAN Application

To provide a comprehensive monitoring system, WBAN takes advantage from some wearable medical sensors as well as implantable sensors such as: Electroencephalogram (EEG), Electrocardiography (ECG) and Electromyography (EMG) [32]. The feature of these sensors are shown in Table 2.1.

2.1.3 WBAN Requirements

Since 2000, interest in using WBAN services and remote monitoring healthcare has risen steadily. Using this application as a health controller device needs to satisfy some objectives including being small and light, comfortable and easy to use and providing reliable records [22]. During the last decade 16 features have been proposed to evaluate the WBAN's performance [50]. The definitions of these features are found in Table 2.2.

Table 2.2: Basic WBAN requirements [50]

ID	Feature	description
F1	Wearability	The system must have low weight and size.
F2	Appropriate placement on the body (Comfortability)	The system has to be unobtrusive and comfortable , in order not to interfere with the user's movements and daily activity.
F3	Aesthetic issues	The system should not severely affect the user's appearance.
F4	Data encryption and security	Encrypted transmission of measured signals and authentication requirement for private data access
F5	Operational lifetime	Ultra low power consumption for long-term, maintenance-free health monitoring.
F6	Real Application	The developed system is applicable (and useful) to real-life scenarios/health conditions.

F7	Real-time Application	The wearable system produces results, e.g. display of measurements, alerts, diagnosis etc, in (or near) real-time
F8	(CCR)Complexity and Computational Requirements	The number of operations and computational power required by the system to achieve desirable results.
F9	Ease of use	The system incorporates a friendly, easy-to-use user interface.
F10	Performance and test in real cases	Sufficient results and performance statistics are provided to verify the system's functionality in real cases.
F11	Reliability	The system produces reliable results.
F12	Cost	The amount of money required to produce and purchase the proposed wearable system.
F13	Interference Robustness	Availability and reliability of wirelessly transmitted physiological measurements.
F14	Fault Tolerance	The system produces reliable results under any circumstances, such as various kinds of patient's movements.
F15	Scalability	Potentiality of upgrading, enhancing and easily incorporating additional components to the developed system.

F16	Decision Support	The implemented system includes some type of diagnosis / decision mechanism or an algorithm / pattern recognition system for context aware sensing of parameters.

Based on these requirements, the next section presents an overview of the current WBAN system and remote monitoring system.

2.2 An Overview of Existing WBAN Services

Since the last decade, a number of WBAN frameworks have been proposed to improve the efficiency of the remote health monitoring system. However, the earlier devices such as American Medical Alert Corp (AMAC) [33] did not have flexibility for communication between the patient and physician. In the latest application, Live-Net [62], the device will transfer the data to the correspondent medical center automatically. Some of the most significant WBAN applications have been explored in Table 2.3.

Table 2.3: Selected telemedicine Application

Project Title or Description	Hardware/Communication	Measured Biosignas	Data analyzing
American Medical Alert Corp (AMAC)[33]	E-Questioner /cell phone	No sensor , work based on Digital Questioner	E-record-analyses at medical center
Tele-Station Unit [1]	Central coordinator kit act as a hub/ telephone line	ECG, HR, SpO2, BP and weight scale	E-record-analyses at Philips server
Live-NET [62]	PDA, microcontroller board / wires, 2.4GHz radio, GPRS	A, ECG, EMG, GSR, T, R, Sa02, BP	MIT, and Cambridge university
AMON (EU IST FP5 program) [5]	Wrist-worn device / GSM link	BP, T, Sa02, ECG, A	EU FP5 IST
Medical Belt [45]	belt, chest sensors[45] / wires , Bluetooth	ECG , 3-Axis	Philips biomedical research group
My-Heart (EU IST & PDA, Textile FP6 program) [30, 27]	electronic sensors on clothes / conductive yarns, GSM, Bluetooth	ECG, R, other vital signs, A	European Commission cooperation between 33 companies

2.2.1 AMAC

American Medical Alert Corp (AMAC), developed the Asthma Monitoring System (AMS) for remote monitoring of asthma, which is the most common chronic disease of children [33]. During the last decade, it was the number one cause of Emergency Department(ED) visits and hospitalizations. Annually in New York City, almost 300,000 children, nearly 17 percent population of children, have suffered Asthma. The prevalence of this disease is almost 3 times as likely to lead to hospitalization for children of low-income neighbourhoods compared with those who live in higher-income neighbourhoods.

The AMAC is an electronic monitoring system based on self managment to assist patients with different chronic conditions such as congestive heart failure and diabetes mellitus. Considering the AMAC project, the AMS kit is given to the patients free of charge. It is an electronic device as big as a hand with four keys and a cell phone. The patients (children at least 8 years old) are asked a short list of questions (varying from encounter to encounter) with this machine every day. Those answers are sent to a central site for analysis. Considering their conditions, the healthcare provider may call them to estimate the child's health status, and adjust the appropriate medication if necessary, or have the child come in to see the physician.

The proposed system is working based on a very simple interface and there is no need for special training for children. The AMS kit has been designed with four different color keys and the trainer should to explain the meaning of each colors. It is supported with both English and Spanish languages. According to data published by the Metro-Plus Health-Plan, from August 2004 through April 2006, a total of 59 subjects were examined with this method. Overall, 50 subjects had visited ED at least once in the prior year, and 24 subjects had at least 1 hospitalization. During this study, it expected that these children would have mean of 2.4 ED visits per month and at least 1 hospitalization every 7 weeks; I.e. they were expected to have 2 to 3 ED visits per month and 12 to 16 hospitalizations during the 24-month study period, based on the data from Coney Island Hospital. These ratios dropped significantly as a result of using the AMS kits .

AMAC provides a home care system called Health Buddy, which allows its patients (especially teenagers) to send their health status to a physician in a medical center. However, it is a very simple electronic kit designed for asthma, Frost and Sullivan Leading Company was developed it to provide remote healthcare system for a variety of medical services such as cardiac care, diabetes, pulmonary disease, pharmaceutical compliance, co-morbidities and mental health.

2.2.2 Tele-Station Unit Project

This project (Tele-Station Unit project) has been designed by the Phillips company to monitor patients' health status [1]. Due to the WBAN requirements, the Tele-Station Unit provides a more proactive healthcare technology by putting more controls in the hands of physicians. The patients' body signals captured by Tele-station units (ECG, HR, SpO2, BP and weight scale) are aggregated by a coordinator device (Glucometer Connector Cable) for transferring to the Phillips database server over the telephone line. It also can send the prescribe medicines if needed. Figure 2.1 shows a quick view of these medical units.

The Tele-station is a simple monitoring application that works with Phillips wireless medical devices. Patients use wireless measurement devices to take their own vital signs and automatically transmit objective measurement results to Phillips' secure server. Unlike the AMAC, which is designed based on a digital questioner; the accuracy of care in this method has been improved when the patients' vital signals are captured, analyzed and transferred automatically, including when patients are sleeping [1].

This monitoring system, has been constructed based on separated units, designed to capture body signals, and analyze and forward them to the medical center. This monitoring system has been known as a primitive gateway for a WBAN, because of the use of these wireless measurement devices. A level of WBAN goals has been achieved by this united system, but this application is not a real WBAN because:

- tele-station units are bulky and patients have to carry out these estimators manually, instead of using a comfortable sensors attached to the body.
- These units have their own embedded power, but mobility has not been achieved by this system, because it is only capable of transmitting the data over a phone line.

Regarding the WBAN requirements, mentioned in table 2.2, the shape, weight and size of wearable medical devices should be kept small and comfortable for an efficient and user friendly monitoring application [51].

Figure 2.1: A quick view of the tele-Station units [1]

2.2.3 LIVE-NET

Live-Net is the basis of the WBAN system, which has been produced by the Media Laboratory of MIT, and Cambridge university [62]. It is a flexible wearable platform, which has been produced for long-term health monitoring with real-time data analysis. Based on the MIT Wearable Computing Group's distributed mobile system architecture, Live-Net is able to continuously monitor a wide range of physiological signals. Using this application, physicians can obtain their patients' daily data, including their activity and emotional data in their accounts.

While Tele-Station was designed based on multiple units for signal processing, Live-NET is constructed with a central processing unit. Figure 2.2 shows a quick view of Live-Nets' structure. As illustrated in this figure, It contains three major components:

- Biomedical sensors (EMG, ECG and EEG)
- Cables: the connections between sensors and base station
- Personal Data Assistant (PDA): a base station or processing unit
  - Application Program Interface (API): communication software and detection application
  - Machine learning and data analyzing inference: for real-time data processing

Live-Net's on-time data analysis makes it an intelligent system, which is able to detect repeating patterns in complex human behaviour. In addition, the accuracy of heart failure detection has also been improved due to a semi real-time data classification and analysis of data feedback.

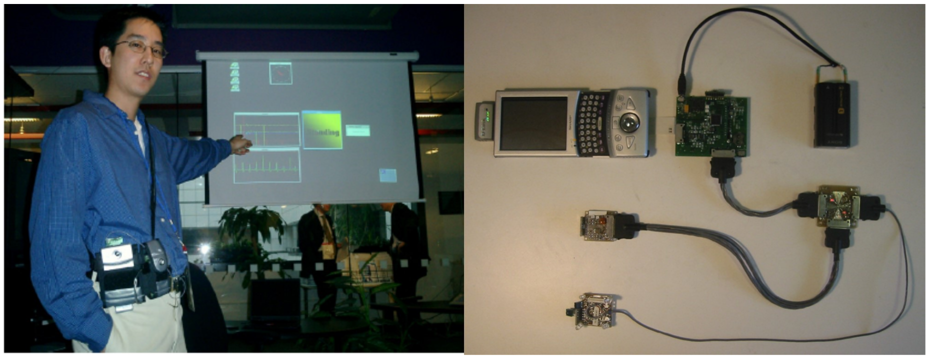


Figure 2.2: The Live-Net wearable streaming real-time ECG/motion setup [62]

Live-Net methodology: using a Linux-based PDA device as a central processing unit makes it flexible for a wide range of objectives such as privacy, accuracy, connectivity, and scalability. To facilitate the PDA data collection, the Live-Net system uses Swiss-Army-Knife 2 (SAK2) as a modular sensor hub for data collection (a coordinator hub, data aggregation and connection between PDA and many sensors). SAK2 is a flexible data acquisition board which is able to transfer megabits of data connected to a Wireless network in 2.4 GHz and various interface ports (Daughter board connector, RS-232 serial, I2C)

To achieve long-term health monitoring, some biosensors are integrated into the SAK2 mainboard. Using this physiological sensing board which called BioSense, enabled the Live-Net application to record patients' daily activities. The board incorporates a three-dimensional (3D) accelerometer, ECG, EMG, galvanic skin conductance, a serial-to-I2C converter (which allows the simultaneous attachment of multiple third party serial-based sensing devices to the sensor network), and independent amplifiers for temperature/respiration/other sensors that can be daisy-chained to provide a flexible range of amplification for arbitrary analog input signals.

Due to this modification, the sensor hub is allowed to have a wide connectivity with commercial medical sensors such as pulse oximetry, respiration, blood pressure, EEG, blood sugar, humidity, core temperature, heat flux and CO2 sensors. The Live-Net system can also be outfitted with BlueTooth, Secure Data (SD), or Compact Flash (CF) based sensors, and communication devices including GSM/ GPRS modems, GPS units, image and video cameras, memory storage, and even full-VGA headmounted displays. Use of the external GPS device, allows more specific classification for different activity classes;(running, sleeping, walking).

In terms of data classification, a MIThril Real-Time Context Engine was installed on PDA to enable lightweight, modular and on-time context classifiers application. As a result of using the context engine, a light-weight machine learning algorithm was implemented on the PDA, which allows systems to classify and identify the variety of user-state contexts 3 times per day using the Bayesian Network algorithm. Live-Net's accuracy and productivity were improved using this embedded machine learning technique. Nowadays, the use of this application has been widely accepted in different areas such as monitoring soldiers' health on the battlefield and a self-organized monitoring system for Parkinson's disease.

- Live-Net application have been widely accepted in different areas.
- Hypothermia study at the United States Natick Army Laboratories.
- Study on the effects of medication on the dyskinesia state of Parkinson's patients by neurologists at Harvard Medical School
- Pilot epilepsy classifier study with the University of Rochester Center for Future Health
- Study of the course of depression treatment with psychiatrists at Harvard Medical School

Overall, the WBAN's targets have been partially achieved using Live-net. The wearability, mobility and real-time data analyzing are the main advantages, which have been successfully achieved by this method. Nevertheless, the system is not comfortable because the device is bulky and include a wired connections for biosensors.

2.2.4 AMON

Around 2005, the advanced care and alert portable telemedical monitor project (AMON), was financed by the EU FP5 IST program, especially for monitoring high-risk cardiac/respiratory patients [5]. AMON was designed to enable three measures; continuous data collection, emergency case detection based on multiple signals estimation and a GSM modem for transferring the captured data to the Tele Medicine Center (TMC).

While the tele-station was designed with a united core and the Live-Net used sticky biosensors around the body connected to a central core in cable mode, the AMON was designed as a single wrist-worn device using multiple embedded biomedical sensors. Figure 2.3 shows the AMON device with five embedded sensors: SpO2, BP, HR, ECG and Galvanic Skin Response Sensor (GSR Sweating) and the central processing unit attached to the GSM modem. Unlike Live-Net, real-time processing and monitoring are not supported by this system. This system worked in offline mode and at certain times, the compressed data was sent to TMC for analyzing and estimating. TMC was an external medical center designed to capture, analyze and detect the patients' health status. The signals were recorded every two minutes and the aggregated data were sent three times a day.

This type of monitoring system is very close to the Tele-station and the AMAC project, which were designed for home care monitoring. Using integrated sensors makes it more comfortable compared with the Live-net, but using offline data processing in TMC did not makes AMON useful as a monitoring system.

Although, the WBAN goals were not successfully achieved with this method, it has made two significant contributions: compression of the multiple sensors in a wearable device, and improvement of the detection accuracy by removal of the ECG noises. This method was an attempt to increase the accuracy of tele estimation by comparing the ECG signal with SPO2 and HR signals at the same time. Moreover, the acceleration kit was used to mitigate the potential generated noise. Measurement of blood saturation using the reflectance sensor does not provide reliable results.

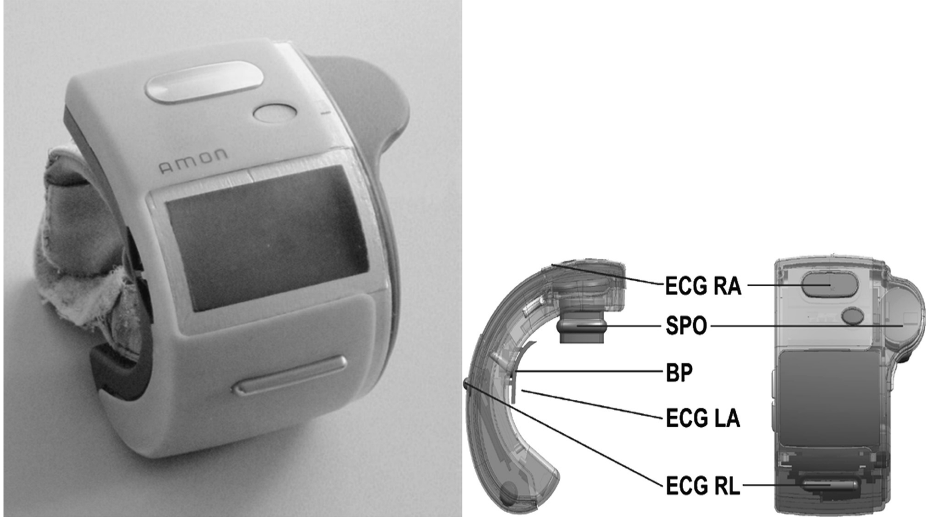


Figure 2.3: AMON prototype with integrated sensors [5]

2.2.5 Medical Belt

In order to facilitate wearability, comfort and mobility another medical device was developed at the Philips research center in Aachen, Germany (biomedical research group), to monitor patients' ECG signal for Cardiovascular disease (CVD); called the wearable medical belt [45]. This belt designed to be worn on the patient's chest, was developed with three integrated dry electrodes to capture the ECG signals. In addition, SPO2 and accelerator detection (2-axis) sensors have been attached to this medical belt to improve detection's accuracy. These low-power sensors and the storage device are arranged to record the patient's activity for 48 hours.

As Figure 2.4 shows, this product is easy to use, comfortable, portable and not as bulky as the tele-station and Live-Net or the AMON devices. It has improved the patient's life quality by attaching the medical sensors to clothing, which is a daily part of our life and in close contact to the signal's source. The idea of attaching the monitoring device to clothing has a huge impact on WBAN and telemedicine. The medical belt has been designed based on simple offline methodology, which is integrated with 3 dry sensors and 65 megabits internal storage. The captured signals on the internal storage derive are sent to the home-pc with a Bluetooth connection and the aggregated data are transferred over the ADSL-Internet to the corespondent medical center for analyzing.

Although it was not well designed for WBAN when it could not support online and out door care, integration of the medical sensor to clothing was a unique idea, which has swamped the researchers interest to design medical clothing for the telemedicine system. Howbeit, the medical belt was a new idea for designing wearable sensors, but as shown in Figure 2.4, it is still a wearable medical component and is not a real daily cloth.

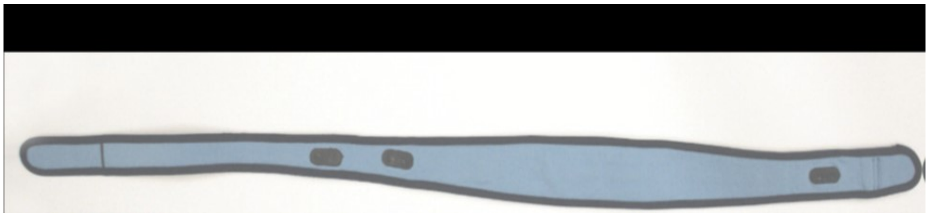


Figure 2.4: Wearable belt approach with three integrated dry electrodes for monitoring patient's ECG and activity [45]

2.2.6 My-Heart

Like the Phillips medical belt, the My-Heart is another application to monitor patients' heart status. This is an international project supported by the European Commission, a cooperation between 33 companies such as Nokia, Philips, Vodafone and Medtronic from 10 European countries to detect CVD [27]. Almost 20% of European citizens have or are very likely to have this disease. The European Union healthcare system spent annually a hundred billion Euros on CVD cure, according to the decadal CVD care report, in the 2003 Guidelines a 10-year risk of CVD death of 45% or more was arbitrarily considered high risk [26]. With the aging population, it is a challenge for the European Union healthcare system to provide affordable healthcare to its citizens.

To mitigate this potential ratio, the My-Heart application was proposed to provide a comprehensive healthcare monitoring system from the early detection of CVD. It was designed to analyze patients' body signals using wearable sensors providing patients with access to care. The project was planned to take 45 months, from 2004 to 2007, to design wearable medical-clothing to monitor patients' ECG signals. The stress management and patients' depression analysis are interesting parts of this project. While other methods focused on patients' heart signals, My-Heart proposed a comprehensive monitoring system which considered the patients' psychological elements as well as their health status. This project was constructed based on four main objectives: Activity Coach, Take Care, Neuro Rehab and Heart Failure Management [30]. Figure 2.5 shows a quick view of these applications.

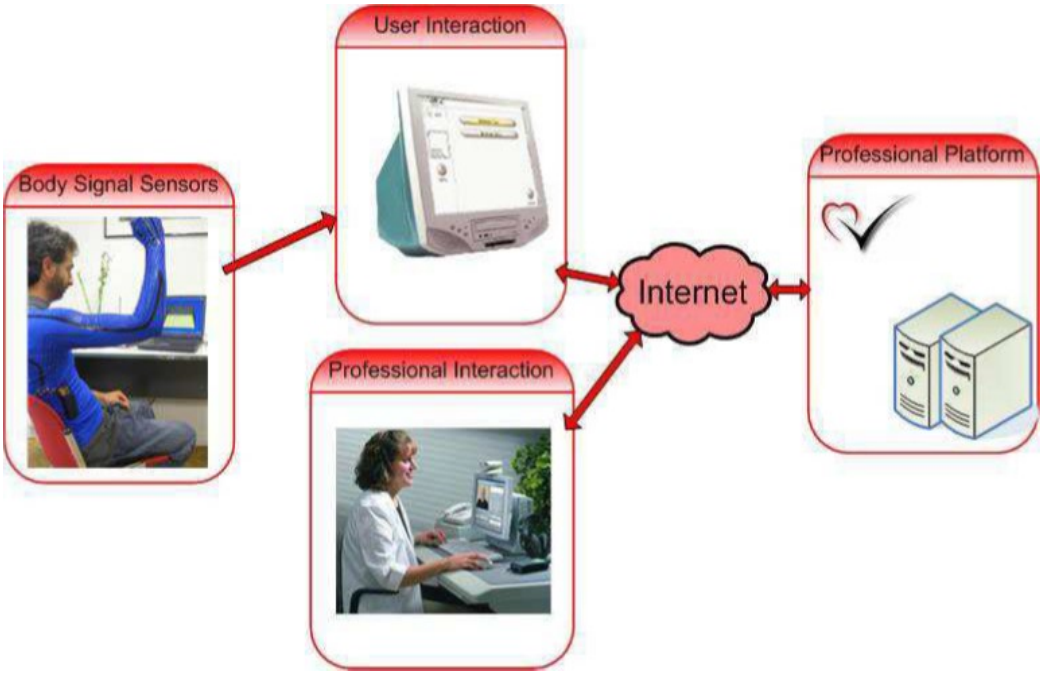


Figure 2.5: My-Heart projects' architecture

2.2.6.1 Activity Coach

The maximum benefit to users occurs when the coach and physician can monitor wearers' bio-signals during outdoor activity. The user wears a medical T-shirt integrated with a ECG sensor, called the Body Signal Sensor (BSS). The captured signals are transferred to the mobile device over the Bluetooth channel. The Personal Mobile Coach (PMC) is a mobile device which is configured for the outdoor scenarios to generate appropriate feedback and send it to the medical center for analyzing by healthcare providers including physicians and coaches.

The Fitness Coach Service Centre (FCSC) is the professional platform that provides online services to the user. The user's data are stored in this center, and analyzed with special training algorithm, and the results are uploaded on the web-based application. Regarding this architecture, a specific training schedule will be given to users based on their daily activity and capacity.

2.2.6.2 Take Care

The aim of this project is to improve patient's life style by analyzing daily-life parameters: blood pressure, weight and cholesterol. Like the activity coach project, this application also uses the same smart T-shirt integrated with ECG sensors to monitor daily activity and sleep time. The BSS signals are transferred to the PMC, and the compressed data are uploaded to the server site for analysis and results are given to patients using web-based application.

2.2.6.3 Neuro Rehab

The objective of neuro rehab is to improve and shorten of the rehabilitation process with motor and cognitive exercises in the rehabilitation ward and in the patient's home. This project provides an intensive remote rehabilitation care system for patients using wearable technology, speech therapy tools, learning tools and communication tools. It is structured in three stages:

- The patient station: the monitoring application, BSS and PMC transfer the patient's data to the therapist site
- The therapist site: monitors the incoming data from the physician for analysis and decision-making; it works manually and there has not been an accurate algorithm proposed for this stage.
- The server site: is the central monitoring server including hosts, databases, configuration rules, exercises, session recordings, demographic data, and the rehabilitation protocol.

2.2.6.4 Heart Failure Management

To improve the quality of life and life expectancy for peoples with heart failure, this application uses early prediction to improve patient self-care management. To upgrade the patients' heart status and quality of life, this application uses heart signals that are relevant to heart failure, on a daily basis. This application also has the same units, including the BSS, PMC, web-based application, and the medical station.

The data is automatically analyzed in order to detect changes in the patient's health status early enough to allow early therapeutic intervention, thus avoiding severe deterioration and hospitalization.

2.3 WBAN Objectives

As can be seen, over the last decade, using WBAN has been widely accepted as a patient monitoring system. However, this application has been proposed as an off-line monitoring system in an AMAC application in the last few years, there are some remarkable applications such as the My-Heart and Medical Belt projects proposed based on semi-on-time data processing which increased the efficiency and usability WBAN.

To facilitate this approach, WBAN takes advantage of data miming technique. The data mining and machine learning algorithms are the most significant applications that have been adopted for data processing on electronic datasets [37]. Today, physicians are taking advantage of these techniques to improve the accuracy of patients' health status prediction such as My-Heart [30].

2.4 Data Mining Techniques

Presently, a huge amount of data is collected and saved in electronic storage units called databases. These databases are expanding year by year [65]. Over the past two decades, the use of digital records has been widely adopted in various domains, such as Electronic Health Records (EHR), Client/Hosts Intrusion Detection Systems

(IDS), and Intelligence Business [10]. Due to new technologies and the continual upgrading of datasets, it is easy to find datasets with terabytes of data for analysis [65].

In the upcoming decade the use of big data and data mining at the Internet Data Center (IDC) will increase by 50 times [70]. The use of big data and analysis of electronic records requires adaptable technology. Data mining algorithms are the most significant applications that have been adopted for big data and big electronic datasets [37]. These techniques have been applied to explore the hidden relations between data features. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large datasets [53]. Machine learning is a significant technique which provides data patterns for data mining. This application can be defined as a computer program that learns from event ”E” regarding a number of defined tasks ”T” that leads to a performance measure ”P” [44].

Machine learning algorithms are designed to probe datasets. These datasets may have a continuous, categorical, or binary sets of features [37]. The results of these algorithms (data feature extraction) are called nuggets of knowledge and are retrieved from large sets of data. To solve this problem of features' data investigation, three main domains have been proposed: classification, association rule, and clustering [65].

Generally, when working with machine learning algorithms, datasets will be separated into two sets (training or learning and test sets). Machine learning algorithms will learn and explore the data features from the learning sets and evaluate the learned pattern (extracted data features relationship) by analyzing the test sets [60].

Machine learning algorithms can analyze data from labeled and unlabeled datasets, which is called supervised and unsupervised learning, respectively. In the supervised mode, all datasets' rows have been labeled with a specific class, which is called a class label. In contrast, in the unsupervised mode, datasets' rows have not yet been classified with a particular class label [58].

In the supervised mode, the outputs regarding the use of the labeled data and clustered data are known. Today, researchers are taking advantage of this technique to discover those unknown features' useful relations [37]. Currently, this technique is being applied to discover exciting and hidden relations of different usages such as patient monitoring systems [30, 62], disease diagnosis [69], heart disease diagnosis [38] and decision making for management [2].

Today, using data mining includes the following tasks [39]:

- Description
- Estimation
- Prediction
- Classification
- Clustering
- Association

Description: Using data mining to determine and describe the relationships within data and their patterns. Some data mining methods such as decision tree are more suitable for a transparent explanation.

Estimation: Is roughly like the classification method, but has been designed for the numeric target variables.

Prediction: When using the estimation and description methods to detect the actual class of data, the data class can be methodically predicted. Indeed, both description and estimation methods are subsets of this class.

Classification: The predefined classes are analyzed with a learning algorithm to determine these classes. This task known as a supervised learning, concern the use of a learning set of objects to make a classification [67].

Clustering: Breaks down records to different clusters. A cluster is a collection members with similar features which has different elements from other clusters. Clustering is different from classification; in clustering, there is no target variable. There is no estimation or prediction of the value of target variables. Clustering seeks to cluster data into appropriate clusters with maximum similarity between elements of each cluster and minimum similarity between elements of the other clusters. It is called unsupervised learning because the classes are unknown and are discovered from data. Clustering can be used for exploratory data analysis and to visualize data, to discover similar instances [58].

Association: This task of data mining is used for finding the relationship between two or more attributes. Association tries to find the rules between some attributes that can play a significant role to predict potential events [39].

Considering the research objectives, which include the prediction of future cases based on real-time data analysis, in this section, supervised data classification is described. In the following section, classification using the decision tree, SVM and the multi-optimization problem based on genetic algorithms is discussed.

2.4.1 Data Classification Algorithms

In this section, well-known classification methods including Decision Tree and SVM, which are known as the most significant classifier algorithms, are reviewed.

2.4.1.1 Decision Trees

Murthy in [47] proposed using the decision tree algorithm for newcomers' machine learning. In this method, data is sorted into different categories based on features' values. In this method, features and their values are represented by nodes and their branches. Figure 2.6 shows a simple classification using this method for Table 2.4 [37].

Classification using this algorithm is depicted in Figure 2.7 as an example, to show how data are classified based on features. As shown in this graph, any problem is solvable using a decision tree algorithm, but this method is inefficient due to its time complexity. This algorithm can find the optimal solution based on heuristic deep

Table 2.4: Training Set dataset [37]

at1	at2	at3	at4	Class
a1	a2	a3	a4	yes
a1	a2	a3	b4	yes
a1	b2	a3	a4	yes
a1	b2	b3	b4	No
a1	c2	a3	a4	yes
a1	c2	a3	b4	No
b1	b2	b3	b4	No
c1	b2	b3	b4	No

search, which is solved in an NP-complete problem. Due to the time complexity of this algorithm, it is not recommended for real-time monitoring applications [47].

Indeed, based on greedy searching this algorithm will continue its learning until all possible solutions are found. Depth-First Search (DFS) is simply an exhaustive search which is used for the decision tree learning step. A DFS algorithm will go down to the tree to find a solution, and after evaluating the feasibility of solutions, it will try to find other solutions using backtrack [56]. Equation 2.1 shows the time complexity of the upper bound of this algorithm. In [52] it is proved that the complexity of this algorithm is in NP-Complete. All the best possible solutions will be explored using the decision tree algorithm, but due to the time complexity of this algorithm, it is not feasible to use in a real-time monitoring system.

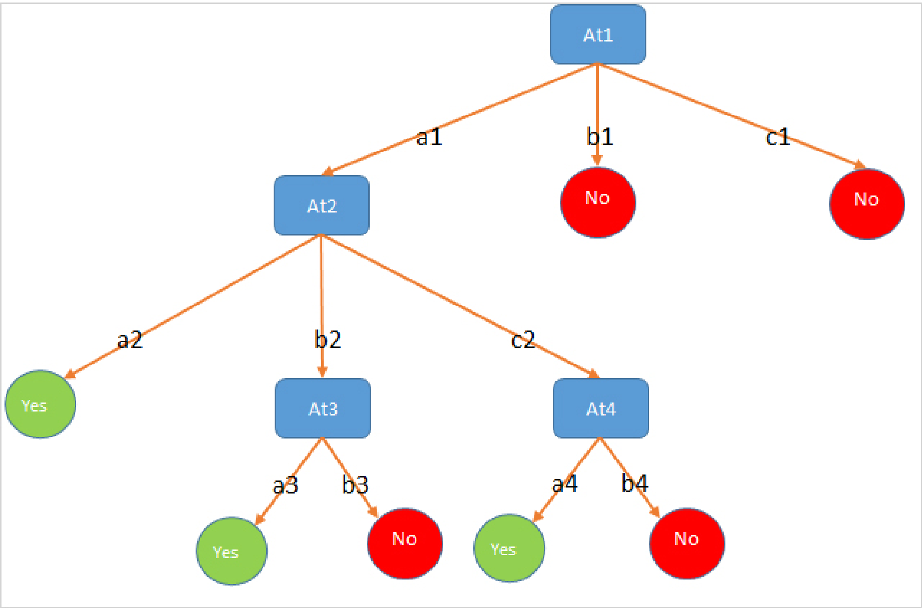


Figure 2.6: Classification using the decision tree algorithm

$$IfFeatureControl = \left\{ \begin{array}{l} \\ \mathbf{b} \ is \ the \ branching \ factor \\ \mathbf{d} \ is \ the \ depth \ where \ the \ solution \ is \\ \mathbf{h} \ is \ the \ height \ of \ the \ tree \ (so, \ \mathbf{d} \ \leq \mathbf{h}) \\ Then \\ DFS \ takes \ O(b^h)O \ time \ and \ O(h) \ space \\ BFS \ takes \ O(b^d) \ time \ and \ O(b^d) \ space \\ IDDFS \ takes \ O(b^d) \ time \ and \ O(d) \ space \end{array} \right. \tag{2.1}$$

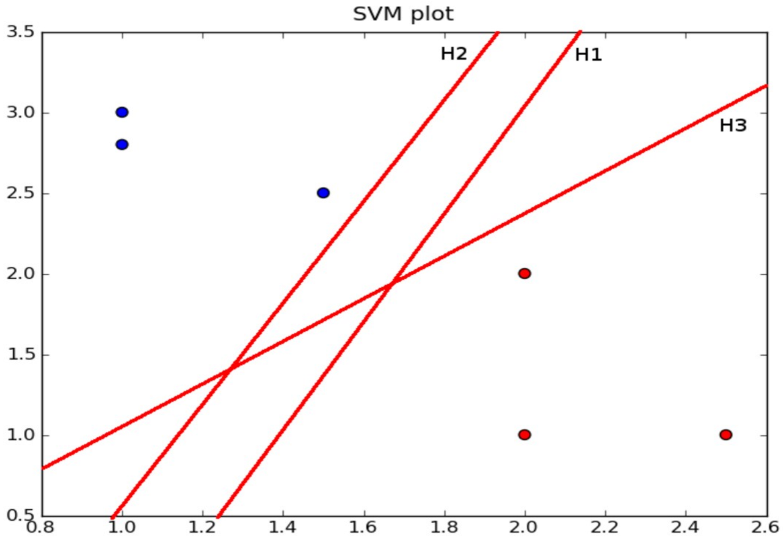


Figure 2.7: SVM decision boundaries

2.4.1.2 Support Vector Machine (SVM)

The Support Vector Machine (SVM) is the significant classifier algorithm, known as state-of-the-art introduced in [36, 9]. It is a very popular classifier algorithm because of high accuracy classification and ability to analyze the big data.

The advantages of SVM are: unlimited feature are supported with SVM and it also provides a non-linear decision boundaries for linear classifiers [36]. SVM algorithms are useful for solving a learning problem with two classes. The datasets will be divided in two classes using SVM optimal boundary regions. Figure 2.7 shows a schema of classification for two sort data (blue and red) using SVM optimal bindery.

Clearly, each problem has multiple possible decision boundaries, however just one of them represents the optimal margin. The one of these lines known as the optimal boundary which have an equal distant from both support vectors' classes. The nearest vectors to the optimal boundary are known as Support Vector or SV. They called SV because the decision boundary will be changed if one of them changes, when others learning vectors have no effect on the decision boundary. Figure 2.8 shows the optimal SVM boundary and its support vectors

Indeed, they are the margin of optimal frontier. SV and optimal frontier have a constant distance which is known as margin. The goal of using SVM is to find the maximum margin between two classes. The optimal boundary in SVM is very close to the Pareto frontier in NSGA, which is used in this thesis. In this thesis SVM frontier has been improved using NSGA Pareto Front points. This section has been deeply discussed in Chapter 5.

2.5 Data Classification using Optimization Algorithm

Over the years, various procedures have been proposed to improve the sorting of data using a calibration process between the quantity and quality of measured data from several single storm events. The first method of multi-target tracking (multi-single events) was proposed by Maalel and Huber in [41, 6], based on analyzing the results of observation and expectation results. Today, sorting and clustering known as multiobjective optimization is still a challenging problem for multi-dimensional data [40].

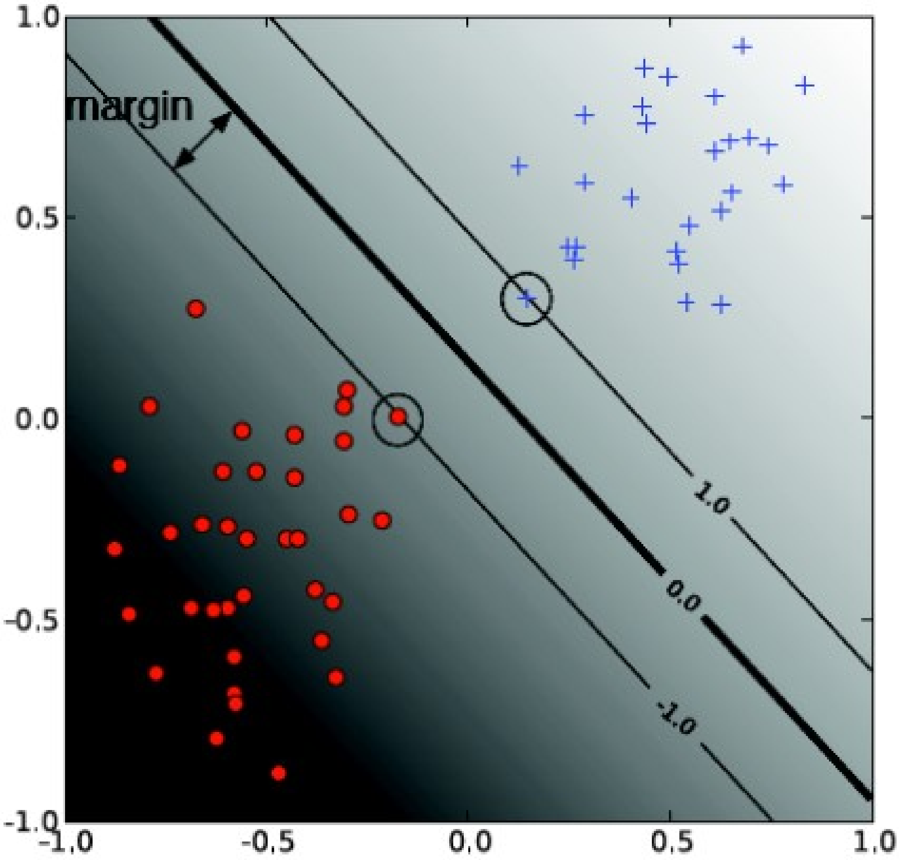


Figure 2.8: SVM concepts

The significance of study in this area is not only because of the multiobjective nature of most real-world problems, but also because it may solve open questions in this area. Although optimization has no universal definition, because the best answers are related to problems and human decision making, applying the Evolutionary Algorithm (EA) to solve optimization has provided an objective for problem-solving called the Multiobjective Evolutionary Algorithm (MOEA) [11]. By using the EA for optimization problems, the best possible solutions will be discovered using the Pareto Frontier theory [59].

A Multiobjective Optimization Problem (MOP) can be defined as using a function to recognize a vector of a satisfiable condition based on a mathematical description to show the accuracy of the function's performance. Indeed, in terms of optimization, all the acceptable solutions will be discovered [12]. Finding these optimal solutions is an NP-Hard problem regarding some conflict between commensurable (measured in the same units) and non-commensurable features. With regard to this conflict, in case of optimization there are some objectives which should be minimized and some which need to be maximized. These objectives will be represented with $F(k)_i(X)$ where k is the total number of objectives and (i) is a positive integer number. The objectives vector can be represented in a vertical or horizontal matrix.

While traditional optimization methods can find a single objective in each simulation, using EAs allows the tracking of an entire set of solutions in a single simulation. EA is a population-based algorithm which allows the generation of several elements of the Pareto optimal set in a single run [16]. It is the main motivation for using EAs to solve problems.

2.5.1 Optimization

The nature of MOP has been inspired by single target optimization. To formulate MOP, there is a set of possible solutions instead of a single target. These objectives are predictable due to the Pareto Front Optimization Theory [6]. To solve the MOP using Pareto Front theory, all objectives of (xi) should be compared with all other members' objectives. Indeed, the selection is essentially a compromise of one complete solution over another in multiobjective space.

2.5.2 Pareto Frontier Terminology

The aim of MOP is to find good trade-offs between multiple objects rather than a single solution, as in global optimization. The notion of optimum most commonly adopted and generalized by Vilfredo Pareto [12]. This theory was modified by [59], and named the Edgeworth-Pareto optimum.

2.5.3 Pareto Frontier Theory

$$\begin{aligned}&\{(x_i) \xrightarrow{domiante} (x_{i+1}) IFF \Longrightarrow \forall (X)k_j | x \in [X], k \in [K\ and (i,j) \in [N];\\&(x_i)k_j \leq (x_{i+1})k_j and \exists (x_i)k_j {<} (x_{i+1})k_j\}\end{aligned} \tag{2.2}$$

Equation 2.2 shows the Pareto frontier equation. In this case, there is some phenotype or data with k size of an objective or genome. Regarding the Pareto frontier theory, all members are compared with one member (Xi). In this trade-off, all objectives of (Xi) will be compared with all objectives of $(Xi + 1)$. (Xi) is non-dominated with $(Xi + 1)$ if and only if (Xi) is smaller or equal to $(Xi + 1)$ for all objectives, and if at least in one comparison it has a lower cost than $(Xi+1)$. After all comparisons, nodes' ranks will appear in the ranking table. This table shows the rank of each node, and the number or comparisons where node (Xi) has been dominated by another node. Those nodes that have not been dominated in each comparison will sort as our Pareto front curve. This subject has been discussed in depth in Chapter 4.

2.5.4 Genetic Algorithm (GA)

Genetic Algorithm (GA) is the most significant branch of EA and was first proposed by John Holland in the 1960s [43]. This is an optimization technique for exploring possible optimal solutions. When optimization techniques are working based on random processes, GA's features make it far more powerful than a random search operation. Chromosomes' population, elitism selection based on a fitness function, producing new offspring with the use of parent crossover and random mutation are the significant uses of GA. These features allow GA to explore possible optimal solutions even if the dataset has discontinuity and derivatives.

The GA's chromosomes are represented in binary strings, with two possible alleles for each chromosome's locus: 0 and 1. GA's fitness function will evaluate and select the elitism chromosomes for a new population. Due to the correspondent problem, chromosomes will be scored by GA's fitness function, which shows how well that chromosome can solve the problem.

In this thesis, GA has been proposed to find the optimal Pareto front offspring because of a lack of continuity of patients, and diversity of features. Using this technique allows the exploration of possible optimal solutions for both healthy and unhealthy classes. The significance of using GA has discussed in depth in Chapter 4 and 5.

2.5.5 Using MOEA for Feature Selection

The history of using MOEA to deal with data mining and feature selection dates back to the last two decades. MOEA has drawn researchers' attention to solving the multi-objective problem when the most suitable solutions are explorable as a Pareto front curve [24]. The Pareto front curve is a sort of non-dominated solution, as the member of this curve will represent the best possible solutions to solve the correspondent problem. The Reduced Pareto Set Genetic Algorithm with elitism (RPSGAe) [25] is proposed to reduce the number of solutions on the efficient frontier. Using this application, simulation results are sped up when elitism members will be selected for the next generation.

To deal with feature selection and increase the accuracy of the classification hybrid classification's framework is proposed in [7]. In this method, MOEA has been utilized in the classification with SVM algorithm to minimize the risk of the classification [7].

In [48] a set of classifiers based on the hierarchical MOEA operation are proposed to increase the accuracy of features selection. Furthermore, in [29] NSGA-II was applied in the neural network as the classifier function to optimize the feature selection and reduce the global error of classification.

Besides using GA to improve the accuracy of feature selection and classification, in [38] GA has been applied in the fuzzy system to improve the transparency of classification to deal with Coronary Heart Disease (CHD). The use of GA and optimal Pareto front to increase the accuracy of classification was enhanced in [69] as well. In this study, the author proposed using NSGA-II to extract the optimal kernel point in the SVR (Support Vector Regression) classification model to deal with the early diagnosis of four diseases (liver disorder, breast cancer, diabetes, and hepatitis).

A hybrid decision-making system which is proposed in [69] consists of the NSGA-II and SVR classification approach. To improve the accuracy of decision-making, all data have been evaluated with SVM in the training set. Then, data have been simulated with NSGA-II and newly generated solutions have been verified with SVM as a test set. The evaluated data for each class have been added to the dataset and retrained the SVMs patterns for future classification. When new suitable solutions have participated in the training set, the accuracy of decision making has improved from 95.57 using supervised fuzzy clustering to 98.76 using the proposed hybrid classification method.

Following the problem that is mentioned in [35, 69] this thesis also takes advantage of NSGA-II to improve the accuracy of decision-making and explores the possible optimal boundary curve between the unhealthy and healthy region. As mentioned in section 1.6, we used NSGA-II to improve the dataset gap, defined the optimal boundary and explored possible tolerance thresholds for each particular patient. The estimated points for each patient would be submitted as its tolerance thresholds to the Self-Coach application. Self-Coach is a new WBANs application used to monitor a patient's health status based on their heart activity. In real-time, this application will be used to monitor dynamic health data and compare them with the patient's restricted points.

Chapter 3

Research Method

The rapid growth of the global population of elderly people, and the resulting increases in healthcare and transportation expenses, and rising average lifespan, are the catalysts for innovating aspects of the Wireless Body Area Network (WBAN), a new healthcare application for a remote patient monitoring system [15]. WBANs consist of smart wearable devices that can process information and communicate with other devices over wireless channels [18]. The captured patient's physical signals are transmitted to the mobile device for data processing, aggregation and storage. The processed data then will be transferred to physicians for final evaluation and decision making. The WBAN application is a combination of four individual units: a Body Sensors Network (BSN), Mobile Computing Center (MCC), Medical Center (MC) and Communication Paths (CP) [51]. Figure 3.1 shows a quick view of the WBAN's units.

To speed up the data processing and decision making, MCC exploits machine learning and knowledge discovery approaches. The aim of this thesis is to explore the

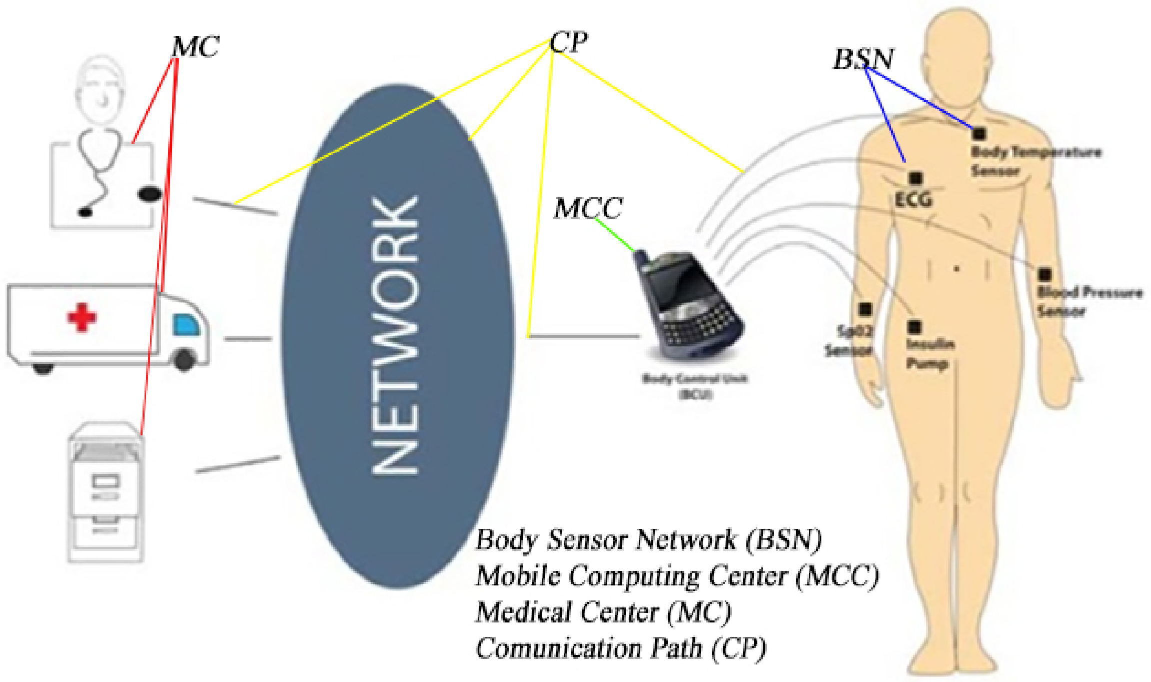


Figure 3.1: Quick view of WBAN application

features of the WBAN application, propose a new method to enhance current MCC applications with the capability of predicting potential health tolerance thresholds, and develop a Self-Coach application which can provide an autonomous active health monitoring system. The proposed application is named Self-Coach because it can detect the body features' abnormality related to heart health status, and has the capability to alarm the user if their health status approaches one of its explored tolerance thresholds. The Self-Coach application has been simulated in Matlab, and the simulation's results have been evaluated with Support Vector Machine (SVM) in the WEKA. This chapter includes a brief review of these steps.

3.1 Thesis Methodology

The methodology of this thesis has been formulated in 5 phases. A brief view of these 5 steps is summarized in Figure 3.2.

Phase 1: Introduction In this step, the background of the telemonitoring healthcare system is explored. This objective has been discussed in Sections 1.1 to 1.3. The current healthcare system has some limitations such as an inadequate number of physicians, which causes an additional cost for patients who require transportation. To reduce the transportation burden, WBAN is utilized as a new application for remotely monitoring patients.

Phase 2: WBAN Terminology The proposed application has been designed to provide a comprehensive health monitoring system for remote patients. Over the years, some features such as portability, user-friendliness, comfortability and low cost have been added to this domain to improve WBAN's functionality. Although WBAN brings many advantages for developing the tele-monitoring health system, it seems it faces some limitations to reaching its goal. Current WBAN applications cannot be called active (real-time) monitoring systems because they use embedded machine learning techniques in their MCC unit. As an example, Live-Net is called semi-active, since the recorded body information will be analyzed three times per day. Another example is the Medical Belt, which records data and processes them off-line. Although WBAN has improved physicians' decision making through the use of the machine learning technique, the time complexity of computing and a lack of dataset information leads to inefficiency in these methods in determining the risk percentage of each patient, and WBAN can not yet provide a real-time health status prediction.

To date, even in the newest WBAN applications, which have been growing in number in the last two years , there is no heart failure prediction feature. Existing applications such as Live-Net or Medical Belt, do not have this function to predict the potential health tolerance threshold based on the patient's activity and current health status. These new technologies have been designed to monitor the user's daily activity, estimate the calorie usage per activity, and monitor heart activity. This step has been explored in Sections 1.1-2 and Sections 2.1 to 2.5.

Phase 3: Purpose of the Study Since using common data mining methods such as Decision Tree, SVM, and Basin Network are time-consuming processes, WBAN has not been widely used as an efficient prediction system. To improve WBAN productivity, and deliver real time heart failure predictions, this study proposes the use of pre-process analysis using the NSGA-II algorithm. The Self-Coach application which is proposed in this thesis is studied in Section 1.4 and Sections 4.1 to 4.4.

Phase 4: Implementation and Evaluation To implement and evaluate the results, three steps are considered:

Table 3.1: Feature Normalization's chart

Feature	Value	Normalized	Description
Age	20 to 100	0 to 1	$1forage \geq 100$

Sex	0,1	0,1	F= 0, M=1
Chest Pain (CP)	1,2,3,4	0,0.33	angina, abnang
		0.66,1.0	notang, asympt
Resting Blood Pressure (BPS)	90 to 200	0 to 1	$1forBPS \geq$ 200
Cholesteral (chol)	120 to 560	0 to 1	$1forchol \geq$ 500
Fasting Blood sugar (fbs)	true,false	1,0	FBS $<$ 120 = 0
Resting ECG (restecg)	1,2,3	0, 0.5, 1	Normal , Abnormal , High
Max Heartbeat Rate (HR)	50 to 185	0 to 1	$HR \geq$ 185 = 1
	50-70	0	
	70-85	0.1	
	85-100	0.2	
	100-115	0.3	
	115-130	0.4	

	130-145 145-155 155-165 165-175 175-185 1 $\leq$ 185	0.5 0.6 0.7 0.8 0.9 1	
Exercise induced angina (exang)	1	0	True False
OldPeak	0.4 to 4.4	0 to 1	$OldPeak \geq$ 4.4 =1
slope	1,2,3	0, 0.5, 1.0	
number of colored vessels (ca)	0, 1 2,3	0 , 0.33 0.66 , 1.0	
thal	0,1,2	, 0.5 , 1.0	normal, fixed, reverse
class			Healthy , Unhealthy

- **Implementation:** Implement NSGA-II with Matlab, and analyze the dataset accuracy using WEKA.

- **Normalization:** since using NSGA-II can support data at the scale of 0 to 1, this dataset is normalized between 0 to 1; 0 represents the best case and 1 is high risk. Table 3.1 shows a quick view of the data set features and the normalization process. This table is discussed in greater detail in Section 4.1.1.

- **Simulation:** The simulation consists of three steps:

  - **Step 1: Preparation**. Figure 3.3 shows the simulation phases. In terms of optimization, the best possible solutions (explored in the Pareto set in each iteration) that are very close to the real data will be separated in a new dataset as a test set. This new set is analyzed separately and the analysis will be repeated to have the highest detection accuracy and minimum noise. The reliable data (Pareto Front members with the highest accuracy of classification) are submitted for simulation with NSGA-II. The data analysis and evaluation steps have been done for each simulation phase (using 6, 10, and 13 features) with the SVM algorithm. Table 3.2 shows the dataset and explored Pareto Front set's accuracy in each simulation phase.

  - **Step 2: Pareto fronts simulation**. In our case, the Pareto Front is the desired optimal boundary between the healthy and unhealthy classes which contains a set of points at which all healthy patients dominate it and it dominates all unhealthy data. However, prior to finding this optimal boundary, the dataset needs to be improved. Thus, new possible patients are generated in both classes in order to fill the gaps in the dataset. This thesis used Non-Dominate Sorting Genetic Algorithm (NSGA-II) to generate new possible patients and explore all possible health positions. The details of NSGA-II have been discussed in depth in Chapter 4.

Table 3.2: Accuracy of classification using the SVM for each simulation

Title	True Positive TP Rate	False Positive FP Rate	Precision
Whole features	0.953	0.049	0.954
All members using 6 features	0.927	0.073	0.927
PF for 6 features	0.909	0.09	0.909
10 selected features	0.93	0.07	0.93
PF for 10 features	0.931	0.071	0.931
PF for 13 features	0.951	0.051	0.951

The unhealthy class has been selected to generate the new simulated health positions since our goal is to find the optimal boundary between the two classes which can be achieved by finding the best possible unhealthy positions. The best possible unhealthy position is the data point where a slight modification in its feature will move it to the healthy class. To get more accurate data, the algorithm NSGA-II is applied in several iterations. In each iteration a new set of points, called the Pareto Front, will be generated.

For our dataset, this thesis uses two dynamic features, heart rate and maximum blood pressure; our dataset does not have a minimum blood pressure, temperature, or respiration rate All possible positions for the patients are simulated with these two sets of graphs. This step has been thoroughly discussed in chapters 4 and 5.

- **Step 3: Tolerance threshold discovery**.

  A patient's health status is subject to change based on new conditions. So under a particular condition, any of the healthy patients can experience critical situations and be categorized in the unhealthy class. A tolerance threshold, which is discussed in this thesis, is a new health position for the patient representing the new health status, which will be dominated by at least one point in the optimal boundary.

  A change in any feature results in a new health position. Technically, only some features will be modified based on daily conditions such as activity, the use of medicine, stress or emotional level. The features that do not change suddenly such as age, sex, and weight have been categorized as static features. On the other side, we have dynamic features, which include features with flexible curves due to the patient's daily activity. Heart rate, blood pressure, and blood sugar are categorized in this class. With respect to the fact blood sugar has no specific value in our dataset (it only has true or false, which is presented as 1 and 0), heart rate and blood pressure have been chosen as this thesis's dynamic factors.

  These two factors are not only subject to sudden changes but also are categorized as critical features for heart disease problems [23]. The tolerance threshold points for each patient are points which show how the patient's health status would be categorised in the unhealthy class based on the patient's new conditions. Indeed, the patient might experience chest pain or heart failure if their health position was dominated by one of the points in the optimal boundary curve. In other words, by increasing the patient's dynamic features, their health position will tend to the right side of the graph. The first point which is dominated by the optimal boundary points will be recorded as the patient's tolerance threshold.

Phase 5: Documentation Figure 1.4 shows a quick view of our methodology. In this figure, one can find the optimal boundary, unhealthy and healthy area, the new health position for the particular patient, and their tolerance threshold. Explored points for each patient will be evaluated with the SVM algorithm as a test dataset. The verified points will be recorded as the patient's health tolerance thresholds. The confirmed thresholds will be submitted to Self-Coach. Thus, Self-Coach only needs to compare a set of values instead of a complete data processing cycle, which certainly can be done on a real-time basis.

The proposed method is named Self-Coach because it provides a non-stop health/activity monitoring system for each particular patient directly. Self-Coach not only provides a remote monitoring system, but also will coach the patient when their health status is in danger, and can alarm the patient to stop the current activity.

Note: In this thesis, the methodology of this application has been proposed, and Self-Coach will be implemented in the future.

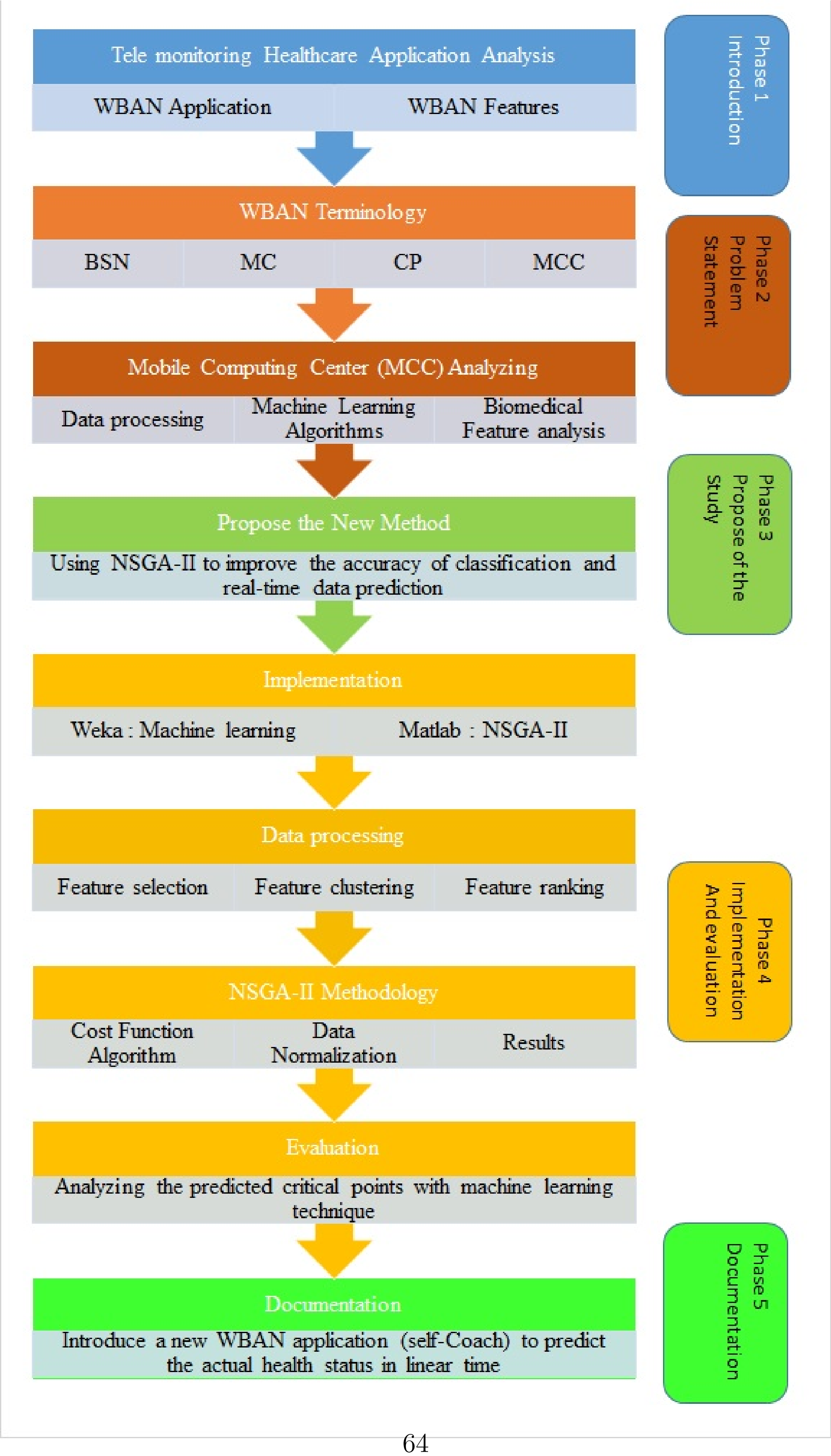

Tele monitoring Healthcare Application Analysis
WBAN Application
WBAN Features
WBAN Terminology
BSN
MC
CP
MCC
Mobile Computing Center (MCC) Analyzing
Data processing
Machine Learning Algorithms
Biomedical Feature analysis
Propose the New Method
Using NSGA-II to improve the accuracy of classification and real-time data prediction
Implementation
Weka : Machine learning
Matlab : NSGA-II
Data processing
Feature selection
Feature clustering
Feature ranking
NSGA-II Methodology
Cost Function Algorithm
Data Normalization
Results
Evaluation
Analyzing the predicted critical points with machine learning technique
Documentation
Introduce a new WBAN application (self-Coach) to predict the actual health status in linear time
Phase 1 Introduction
Phase 2 Problem Statement
Phase 3 Propose of the Study
Phase 4 Implementation And evaluation
Phase 5 Documentation

Figure 3.2: Overview of methodology

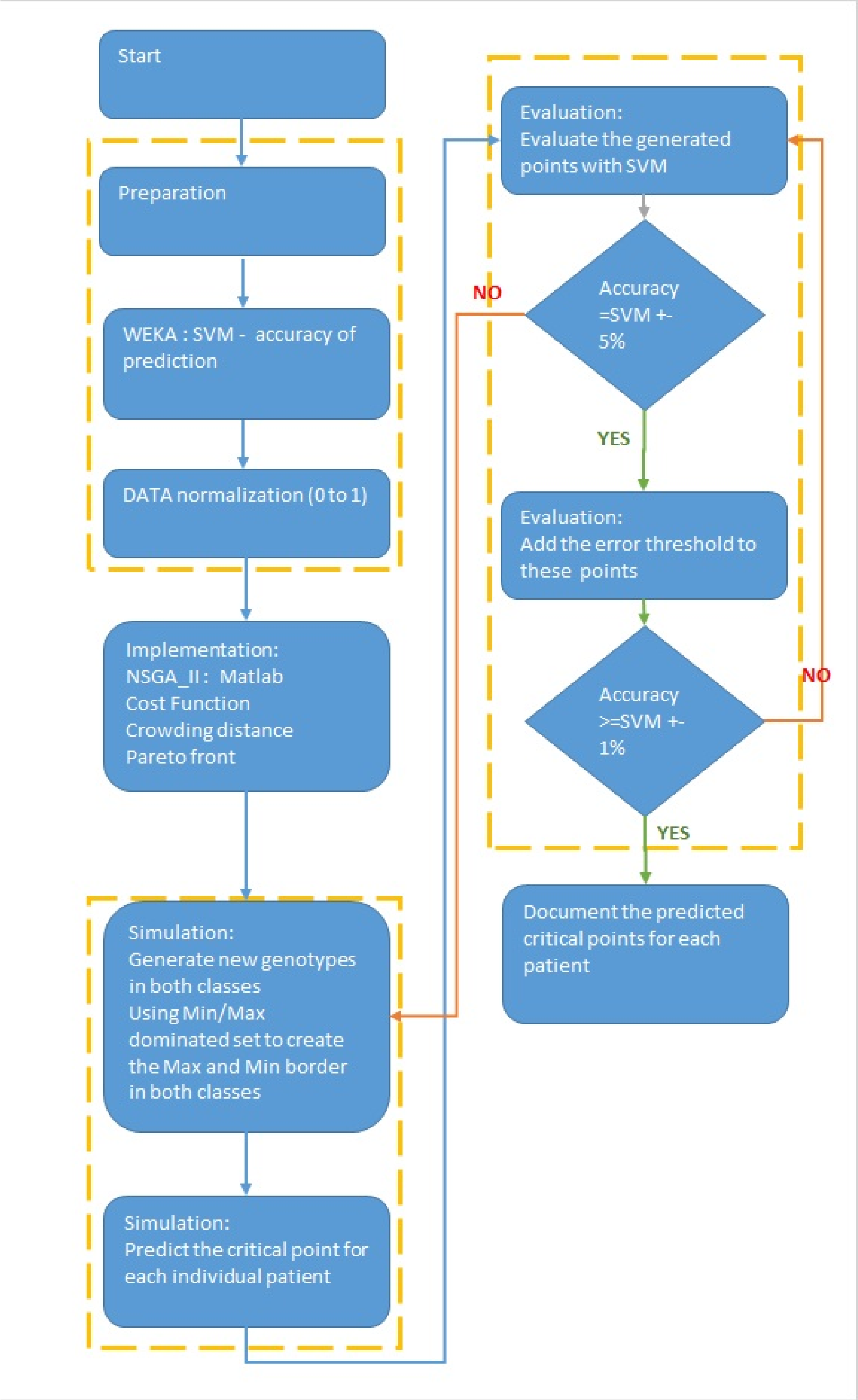

Start
Preparation
WEKA : SVM - accuracy of prediction
DATA normalization (0 to 1)
Implementation:
NSGA_II : Matlab
Cost Function
Crowding distance
Pareto front
Simulation:
Generate new genotypes in both classes
Using Min/Max dominated set to create the Max and Min border in both classes
Simulation:
Predict the critical point for each individual patient
Evaluation:
Evaluate the generated points with SVM
Accuracy =SVM +- 5%
NO
YES
Evaluation:
Add the error threshold to these points
Accuracy >=SVM +- 1%
NO
YES
Document the predicted critical points for each patient

Figure 3.3: Overview of implementation

Chapter 4

Implementation

For the last three decades, using a Multiobjective Optimization approach has been widely accepted as a real world problem-solving application. Identification of the Pareto-optimal solutions is a significant use of these algorithms for problem-solving. Optimization algorithms exploit Evolutionary Algorithms (EA) to speed up this identification process. As a result of this combination, MOEA can generate new solutions (offspring) from current solutions (parents) to speed up the non-dominated solutions exploration. Having a minimum distance inthe selected dominated sets with maximum diversity is a big problem for using MOEA. To formulate this diversification problem, MOEA takes advantage of sorting approaches[12].

Non-Dominates Sorting Genetic Algorithm (NSGA-II) [16], is the most significant MOEA application which has addressed the MOEA's problem. Since all possible solutions (new health positions) will be explored using the NSGA-II technique, in this thesis, this method is used to explore all possible healthy and unhealthy phenotypes. These solutions provide adequate health risk estimation for the real phenotype (patient). This process calculates the patient's health position distance for both Pareto sets (healthy and unhealthy curves). Furthermore, by increasing the dynamic features for each individual a tolerance threshold is explored, which is the position where a healthy phenotype will be labeled as an unhealthy phenotype or going to have a chest pain experience. Figure 4.1 shows a quick view of this theory.

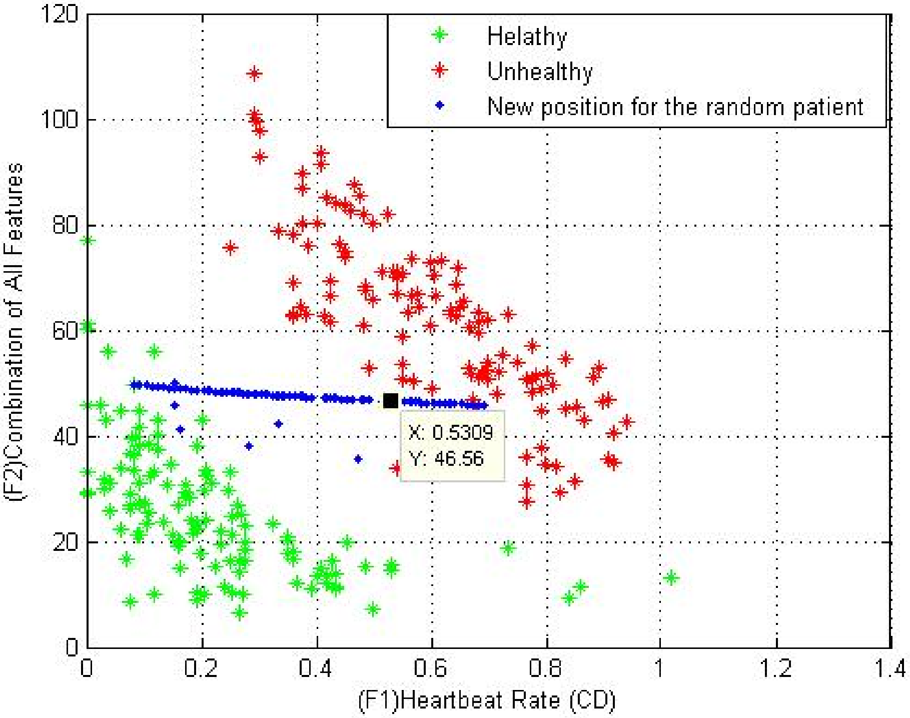


Figure 4.1: Quick view for tolerance threshold exploration methodology

The methodology of this implementation has been discussed in this chapter. Subsequently, there is a brief discussion about these datasets' features, dataset normalization technique, optimization using NSGA-II, implementation, and evaluation. The results of these simulations are shown in Chapter 5.

4.1 Dataset Features

In this thesis the Cleveland dataset (heart disease) [4] is used. This database contains 75 factors, 13 features which have shown the most correlation with the heart disease problem are used in this thesis. These features are proposed with the dataset as well. Table **??** shows this study's selected features.

In this thesis, the patients are separated into two classes: healthy and unhealthy. To increase the accuracy of simulation results these features are separated into two categories; dynamic and static attributes. The features that do not change suddenly such as age, sex, weight, and other medical records are our static features, and some features such as heart rate, blood pressure, and blood sugar that have very flexible curves due to the patient health status, are categorized as dynamic features. The results of these simulations are presented with a two-dimension (2-D) graph, where the X-axis represents the hear rate (Hr) and the Y-axis denotes the trade-off between these dynamics and statics features. Figure 4.1 shows a quick view of this idea.

These features' taxonomy are used to simulate the new phenotypes, (position) for patients regarding their new heart rate. Since NSGA-II can support data on the scale of 0 to 1, the dataset features' value have been normalized between 0 to 1; when 0 represents the best case (Healthy), and 1 represents the unhealthy (high-risk) step. This subject has been discussed in depth in Section 4.1.1.

4.1.1 Normalization Technique

Based on the NSGA-II Cost-function's requirement, the dataset is normalized in the range of 0 to 1. To address the normalization step, a standard definition for all criteria is defined based on the severity of the features' effects. Table 3.1 shows the normalization metrics. These configurations have been brought from our NECEC Conference's paper in 2015 [19].Table 3.1 shows a quick view of the data set features and the normalization process

4.1.2 Feature Selection

To evaluate the results and make a good case for comparison between these results and data mining results, this simulation began with six selected features proposed in [3]. The main problem with using data mining technique is the method of feature selection. As can be seen in Appendix 1, some important features such as age, and sex were not used. With a normal feature selection technique, those features with the highest correlation value would be selected. Although, by using this method the maximum prediction accuracy with minimum features can be achieved, some features that have a significant role to predict the future class for a dynamic dataset might be omitted. In this case, the acceptable hear rate in the healthy class (a tolerance threshold for each patient beyond which patient might have a heart failure experience) and age have a reverse relationship. Besides, simulation results shows that hear rate threshold will decrease by increasing some other features such as blood pressure.

Thus, to increase the efficiency of simulation, all 13 features are used instead of 6 selected features for the e-health conference [21]. The results of this step have been presented in Section 5.2.

4.2 Simulation with NSGA-II

Exploring the optimal solution of conflicting goals or objectives is called an optimization problem. While in the classical optimization there was a single optimal solution explored, a multi optimal solution explores using the modern optimization method, called: Multi-Objective Optimization Problems (MOPs) [11]. The set of these optimal solutions that is examined using MOP is called the Pareto optimal set. These points represent the best possible results, the non-dominated points, which can exist in order to address the problem optimization [64].

The solution is non-dominated with other solution if and only if it is smaller or equal to the other one for all objectives and if at least in one comparison it has a lower cost than the other solution [11].

Classical MOP methods work based on breaking down the objectives to a multi-single objective. Using these methods to explore all possible non-dominated solutions needs many simulations [16]. Over the past decade, MOP techniques have been taking advantage of Evolutionary Algorithms (EA) named MOEAs. The primary reason for using EA to solve the MOP is the ability to find the multi-Pareto optimal set in one single run [18]. A number of MOEA methods have been proposed since 2000 such as Non-dominated Sorting Genetic Algorithm (NSGA) [17], Non-dominated Sorting Genetic Algorithm-II (NSGA-II)[18], Pareto Simulated Annealing (PSA) [61] and Multi-Objective Genetic Algorithms (MOGA) [46]. The Non-dominated Sorting

Genetic Algorithm (NSGA), which is proposed in [17] is one of the most efficient EA methods. Some weaknesses of MOEA such as high computational complexity $O(m(N^3))$, and lack of elitism parent selection, were improved by applying a Crowding Distance (CD) area as a second sorting function, named $NSGA-II$ [18]. In this chapter, the original $NSGA-II$ algorithm has been explored and implemented in the Matlab software.

4.2.1 NSGA-II Methodology

These functions have been proposed to find the optimal solution with an Evolutionary Algorithm (EA) using the Pareto Front theory. As mentioned previously, the best possible solutions will be explored using NSGA-II. This thesis takes advantage of this opportunity to find the optimal boundary between two classes, healthy and unhealthy, to increase the accuracy of classification and explore the potential tolerance thresholds for each patient.

In this thesis, patients belonging to the unhealthy class of the Cleveland dataset [4] have been simulated 50 times with 50 iterations to explore the best possible solutions, and those solutions are categorized in the unhealthy class with the optimal minimum possible features. These are the non-dominated solutions that still belong to the unhealthy class. These explored points and their parents are in the same class (same phenotype) but they have different feature values (different genotype) than their parents. The class label (phenotype) of each new child has been verified with SVM.

To evaluate new members, whole datasets have been analyzed as a training set and newly explored points (non-dominated set) in each iteration have been added to a dataset as a test dataset. Those members that have shown the highest classification accuracy using SVM have been selected as the optimal boundary curve.

These explored points have been added to the unhealthy class to train SVM for tolerance threshold exploration. To explore the potential threshold, (a new condition in which a healthy phenotype might experience chest pain) three random patients have been selected. Table 5.1 shows these three patients' features.

In this thesis, three different simulations are shown: simulations with 6, 10 and 13 features, to have a good comparison between the number of participant features and the accuracy of the explored points for each patient. Table 4.1 shows participant features for each simulation phase. As mentioned in the problem background section, using feature selection can speed up the classification process, but the accuracy of this prediction will not be significant as a result of removing some features. The accuracy of heart disease prediction using six features showed a remarkable result in [3], but regarding potential health tolerance threshold exploration, using a small number of features is not enough. Table 5.10 shows how the potential health tolerance thresholds will be changed based on the number of participant features.

The results of this comparison show that the explored turning points would be changed by applying feature selection. When a trade-off between all features would show the patient's health status, to have accurate prediction results, we are not allowed to remove any of these features.

Table 4.1: Selected features for each simulation

Feature ID	Feature	NECEC -6 features	Aldrich-10 features	E-health-13 features
1	Heart Rate (HR)	*	*	*
11	Slope		*	*
12	Number of vessels colored (CA)			*
3	Chest Pain (CP)	*	*	*
10	Old Peak		*	*
14	Thal		*	*
6	Cholesterol	*	*	*
2	Age	*	*	*
13	Sex		*	*
9	Exercise Induced Angina (Exang)	*	*	*
4	Resting Blood Pressure (Rest BPS)	*		*
8	Rest ECG			*
5	Col-normal			*
7	Fasting Blood Sugar (FSB)			*

The NSGA-II is an open MOEAs' method which can be used to simulate and estimate most of the MOPs such as the Traveling Salesman Problem [34], Knapsack, Ham-Path, Ham-Cycle and Subset Sum [31] in polynomial time ($O(m(N^2))$). It has been designed based on two functions for sorting and selecting the best possible solutions. The result of this algorithm is a Pareto graph, or the best possible solutions which exist for the problem. This method has the following properties [18]:

- Elitism parent selection for the next generation
- Crowding distance value to select the population
- Presents the non-dominated set as a result

The Pareto curve's points are calculated with NSGA-II's cost-function. Based on some parameters such as some objectives, the nature of the problem and expected curves, a number of standard Cost Functions (CF) have been proposed. Table 4.3 shows the selected CFs. Since there are about 13 selected features and a connected curve is required, ZDT4 and ZDT2 are used for this simulation. The results of CF are plotted on a 2-D graph called a Pareto graph. We use these points to calculate the CD for each solution.

In Table 4.3 a number of cost functions which are standard for NSGA-II are listed; however, for this thesis, not all of them were practical, as some functions such as FON, ZDT1, ZDT3, and ZDT6 will not provide a connected non-convexed Pareto Front curve [12]. Therefore, based on the features suggested in [3], we used ZDT4 for our first two simulations (where the number of selected features are 6 and 10) and ZDT2 for the simulation with 13 features.

ZD4 has not been used for simulation with 13 features since it is providing promising results for up to 10 features. We also experimented with ZD2 for the first 2 sets of 6 and 10 features. However, according to the simulation result mentioned in Table 4.2, ZDT2 is not suitable for a small number of features. Thus, the best combination was to use ZD4 for the dataset with fewer features and use ZD2 as a cost function for the dataset with 13 features.

Table 4.2: Comparison between accuracy ZDT2 and ZDT4 using the SVM for each simulation

Title	TP Rate	FP Rate	Precision
Whole features	0.953	0.049	0.954
All members			
using 6 features			
ZDT4	0.927	0.073	0.927
ZDT2	0.893	0.167	0.893
PF for 6 features			
ZDT4	0.909	0.090	0.909
ZDT2	0.885	0.183	0.885
All members 10 selected features			
ZDT4	0.930	0.070	0.930
ZDT2	0.910	0.093	0.910
PF for 10 features			
ZDT4	0.931	0.071	0.931
ZDT2	0.909	0.098	0.909

NSGA Methodology has been constructed based on four main tasks:

- Merging the population: represented as the parent population (Pt) plus the offspring population.

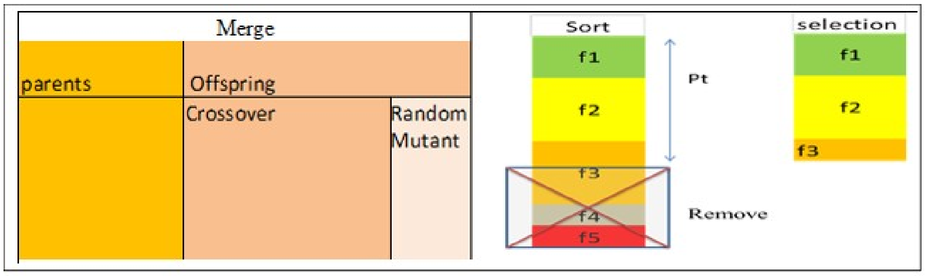


Figure 4.2: Non-Dominated sorting scheme.

- Sorting the population (cluster to the different PF): sorting the merged population into the different front (dominated level) based on their dominated ranks (non-dominated solutions ”f1” to the worst cases ”fn”)
- Tournament Selection: selecting the new PT of best cases for the next generation’s round.
- Parent Selection: selecting elitism parents for the next generation (picking up elitism parents and duplicating the best one with crossover and mutation).

Figure 4.2 shows a quick view of these steps; the merge population, sorted by different layers and elitist selection.

4.3 Implementation

In the following sections, implementation of the NSGA-II is described based on [18] using Matlab as the programming language. The initial solutions are the patients

Table 4.3: Multi objectives cost functions [12]

Problem	n	Variable Bounds	Objective Function	Optimal Solution
FON	3	[-4,4]	$f_1(x) = 1 - e^{(-\sum_{i=1}^{3}(x_i - \frac{1}{\sqrt{3}}^2))}$ $f_2(x) = 1 - e^{(-\sum_{i=1}^{n}(x_i + \frac{1}{\sqrt{3}}^2))}$	$x_1 = x_2 = x_3$ $[\frac{-1}{\sqrt{3}}, \frac{1}{\sqrt{3}}]$
ZDT1	30	[0,1]	$f_1(x) = x_1$ $f_2(x) = g(x)(1 - \sqrt{\frac{x_1}{g(x)}})$ $g(x) = 1 + 9(\sum_{i=2}^{n} x_i)/(n-1)$	$x_1 \in [01]$ $x_i = 0$ $i = 2, ..., n$
ZDT2	30	[0,1]	$f_1(x) = x_1$ $f_2(x) = g(x)[1 - (\frac{x_1}{g(x)})^2]$ $g(x) = 1 + 9(\sum_{i=2}^{n} x_i)/(n-1)$	$x_1 \in [01]$ $x_i = 0$ $i = 2, ..., n$
ZDT3	30	[0,1]	$f_1(x) = x_1$ $f_2(x) = g(x)[1 - \sqrt{\frac{x_1}{g(x)}} - \frac{x_1}{g(x)} \sin(10\pi x_1)]$ $g(x) = 1 + 9(\sum_{i=2}^{n} x_i)/(n-1)$	$x_1 \in [01]$ $x_i = 0$ $i = 2, ..., n$
ZDT4	10	$x_1 \in$ [0,1] x_i ∈[-5,5] $i = 2, ..., n$	$f_1(x) = x_1$ $f_2(x) = g(x)[1 - \sqrt{\frac{x_1}{g(x)}}]$ $g(x) = 1 + 10(n-1) +$ $\sum_{i=2}^{n}[x_i^2 - 10\cos(f\pi x_i)]$	$x_1 \in [01]$ $x_i = 0$ $i = 2, ..., n$
ZDT6	10	[0,1]	$f_1(x) = 1 - exp(-4x_1)\sin^6(6\pi x_1)$ $f_2(x) = g(x)[1 - (f_1(x)/g(x))^2]$ $g(x) = 1 + 9[(\sum_{i=2}^{n} x_i)/(n-1)]^{0.25}$	$x_1 \in [01]$ $x_i = 0$ $i = 2, ..., n$

from the Cleveland dataset [4]. The presented results are the best possible solutions, which can exist for both healthy and unhealthy classes.

4.3.1 NSGA-II Functions

To calculate the non-dominated set, NSGA-II requires two functions, the Cost Function (CF) and Crowding Distance (CD), which are used for computing the new phenotype position and elitism selection for the parent selection tournament, respectively.

4.3.1.1 Cost Function (CF)

The NSGA-II is defined as an open method which can solve and estimate most of the MOPs. The open part of this approach is named CF. In this research, ZDT2 and ZDT4 are simulated as CF [12, 17]. The selected cost function is presented in Table 4.3. The results of the CF present the positions of these solutions on the 2D graph. This position is used to estimate the likelihood of closeness of a patient to the border of the unhealthy class (Pareto graph for the unhealthy class).

ZDT4 and ZDT2 work with up to 10 and 30 features, respectively. The first simulations (Figures 5.2 to 5.4) are simulated with the first CF. Figures (5.5 to 5.8) are simulated with the second CF.

Figure 4.3 shows the simple simulation results. As can be seen, in this graph, there are two fitness functions F1 and F2. These functions show the value of the dynamic data and the trade-off between all features. As mentioned in Chapter 3, Phase 4, the features are separated into two classes, dynamic and static. Dynamic features are called dynamic because during daily activity they can poses various values based on the patient's activity and emotions. With regards to the patient's phenotype, a limit scale of these values is acceptable. These features are treated in a separate class.

These simulations have two dynamic features: heartbeat and maximum of blood pressure. In each set of simulations, one of these features is selected as dynamic data for F1 or X-Axis in the graph, and other dynamic data are treated as static data with F2 or the Y-Axis. Unlike the dynamic features, static features never change suddenly during daily activities; for example, patient's age, sex, the number of vessels colored (CA) and chest pain class never change.

There are some very important dynamic features such as blood sugar, sweating ratio, body temperature and breathing which are not available in our dataset. Thus, in this thesis these features are not considered. In the future study, more accurate results can be achieved by adding such features.

4.3.1.2 Crowding Distance (CD)

This is the Euclidian distance, which will be calculated for each individual separately based on their objectives on each Pareto graph [49]. In other words, when all the Pareto graph members are at the same level of domination, have dominated with the previous Pareto graph and dominate the next Pareto members, the CD value can show the effect of that particular solution on the graph. Equation 4.1 shows the way to calculate this Euclidian distance for each solution.

$$CD = \begin{cases} (CD)_i^1 = \frac{(|f_1^{(i+1)} - f_1^{(i-1)}|)}{(f_1^{Max} - f_1^{Min})} \\ (CD)_i^2 = \frac{(|f_2^{(i+1)} - f_2^{(i-1)}|)}{(f_2^{Max} - f_2^{Min})} \\ (CD)_i = (CD)_i^1 * (CD)_i^2. \end{cases} \tag{4.1}$$

All individuals will be sorted based on their CD values in each Pareto set. This value will be used for population selection. Basically, to select the elitism members, those members from the first PF sets will be selected first and from the last Pareto front, those members with the highest CD value will be elected. Section 4.3.2 shows how the elitism members will be selected using the CD value.

4.3.2 Sorting Methodology

NSGA-II sorts its solutions by using a non-dominated ranking matrix. To fill the ranking table, two steps need to be completed:

- N_p: represents the number of solutions which dominated the patient P
- S_p: a set of solutions which dominates with patient P

After finishing the comparison for all solutions and completing the ranking matrix, the set of solutions are called a first front or non-dominated set, which have not been dominated by any other solutions, or have the N_p equal to 0. Then, these sets of solutions are removed temporarily from the ranking table and their effect on the other solutions and their N_p index are replaced by INF (infinity). Again, there is a new set of solutions which have not been dominated. These are the second front or second non-dominated set. This step is repeated until all solutions are sorted at the appropriate front level

Table 4.4: Simple dataset

Solution	Feature 1	Feature 2
sol1	7	6.5
sol2	10	6
sol3	9	8
sol4	10.5	9.5
sol5	11	7

Table 4.4 shows a simple dataset including five solutions with two features. This dataset is simulated with the NSGA-II. The ranking matrix of this dataset has been presented in Table 4.5.

As shown in Table 4.5, each solution is compared with all solutions. The results of this comparison can be either 1 or 0, which shows whether the first solution has dominated the second solution or not. In this table, when Sol1 dominates Sol2 equal to 0, this means Sol2 has not been dominated with Sol1. In contrast, when Sol1 dominates Sol4 equal to 1, this means Sol4 has dominated by Sol1. The details column presents the S_p and the member of each Pareto graph with F_i when $i \in \{0, 1, 2, , N\}$. Furthermore, the last row in each section of this table shows the N_p for each solution or the number of times each one has been dominated. In each section, those solutions which have N_p equal to 0 have been selected as members of the Pareto graph for that particular section.

The next section demonstrates that the previous Pareto front members have been removed temporarily. Therefore, the other solutions are not affected by those removed members. Thus, those members with $N_p = 0$ are labeled as the next Pareto front members.

While Non-Dominated Sorting is very time-consuming $(O(m(N^2)))$ when m is the number of features and N is the population size, this cost has been reduced to $(O(m(N)))$ as a result of using the ranking table in $NSGA-II$ [18]. To have a quick view of this mechanism, the results of Table 4.5 are plotted in Figure 4.3.

As can be seen, on the left side of this figure there are unlabeled data, and these data have been labeled in a different Pareto front on the right side. Hence, all solutions have been sorted in the appropriate Pareto front. The next step is to select the elitism population (PT) for the next generation as mentioned in Figure 4.2. To deal with this step, NSGA-II has two policies:

- Select all best members (non-dominated set) and if more solutions are still needed to complete the required population, select from the next front and repeat this until the required population is selected.
- To select the best member from one Pareto front, those solutions with maximum CD are selected.

These steps are shown in Figure 4.2. Using this method, the members from the $F3$ will be chosen based on their CDs' value. In this case, the maximum and minimum points of the curve are selected first, and then those solutions with the greater CD value are chosen. This calculation has been shown in Equation 4.1.

After sorting all members of the last Pareto front, those members which have greater CD values are selected. Figure 4.4 shows this tournament.

4.3.2.1 Simulation With Real Data From The Cleveland Dataset

In this section, the sorting technique using NSGA-II for a random set of the Cleveland dataset has been shown to add more clarity on the methodology. Table 4.7 shows a random set of patients that belongs to the Cleveland dataset. As is shown in this table, each patient is presented with a unique ID and 13 features.

In the following section 4.3.1.1, the patient's genotype (features) must be encoded with an appropriate CF to prepare for optimization; therefore, these 12 patients have been submitted to the NSGA-II algorithm. In the first simulation round, the CF has been applied on each patient's features separately.

Equation 4.2 shows how the patient's features will be mapped with two fitness functions ($F1 \ = \ X_1, \ and \ F2 \ = \ (f(x_1, x_2, x_3, ..., x_{13})))$. Table 4.8 shows 13 features of each patient, which have been encoded into two new functions known as the optimization objectives.

$$CF = ZDT2 = \begin{cases} F_1(x) \ = \ X - axis = \ x_1 \ \& \ x_1 \ \in \ [01] \\ F_2(x) \ Y - axis \ = g(x) \ [1 \ - \ (\frac{x_1}{g(x)})^2] \ \& \ x_i \ \in \ [01] \\ g(x) \ = \ 1 + 9 \ (\sum_{i=2}^{n} x_i)/(n-1) \ \& \ i \ = \ 2, ..., n \end{cases} \tag{4.2}$$

Now, these new values (F1 and F2) for each patient will be submitted in the next step of simulation, which is known as the sorting step. This step has been mentioned in section 4.3.2. The result of applying the sorting step on the patient's fitness values (F1 and F2) is presented in Table 4.8. By utilizing NSGA-II these 12 random patients have been sorted in three different Pareto Fronts based on their dominated rank. Figure 4.5 shows these three different PF curves.

After the sorting step, a new rank value (CD value) will be given to these patients by utilizing the CD function which is defined in section 4.3.1.2.

As mentioned in section 4.3.1.2, all members in each Pareto Front have been sorted based on their CD value in Table 4.6. The CD value will help NSGA-II to select elite parents for the next generation. For example, in this case, if 6 parents will be required for the next population, 6 first patients from the top of this table will be selected ($P1,\ P2,\ P6,\ P12,\ P3\ and\ P5$). The first and last member of each Pareto set are known as the most efficient members in the curve. Thus, their position in the curve is very important, and NSGA-II will replace their CD values with INF (infinity). This means they will be selected first in each Pareto set. The selected members are the elite parents which will participate in the next iteration.

Table 4.5: Simple ranking matrix

	sol1	sol2	sol3	sol4	sol5	details
sol1	0	0	1	1	1	$S_P =\{3,4,5\}$
sol2	0	0	0	0	1	$S_P =\{5\}$
sol3	0	0	0	1	0	$S_P =\{4\}$
sol4	0	0	0	0	0	$S_P =\{ \}$
sol5	0	0	0	0	0	$S_P =\{ \}$
N_P	0	0	1	2	2	$F_1 =\{1 , 2\}$

	sol1	sol2	sol3	sol4	sol5	details
sol1	inf	inf	0	0	0	$S_P =\{ \}$
sol2	inf	inf	0	0	0	$S_P =\{ \}$
sol3	inf	inf	0	1	0	$S_P =\{ 4 \}$
sol4	inf	inf	0	0	0	$S_P =\{ \}$
sol5	inf	inf	0	0	0	$S_P =\{ \}$
N_P	inf	inf	0	1	0	$F_2 =\{3,5\}$

	sol1	sol2	sol3	sol4	sol5	details
sol1	inf	inf	inf	0	inf	$S_P =\{ \}$
sol2	inf	inf	inf	0	inf	$S_P =\{ \}$
sol3	inf	inf	inf	0	inf	$S_P =\{ \}$
sol4	inf	inf	inf	0	inf	$S_P =\{ \}$
sol5	inf	inf	inf	0	inf	$S_P =\{ \}$
N_P	inf	inf	inf	0	inf	$F_3 =\{4\}$

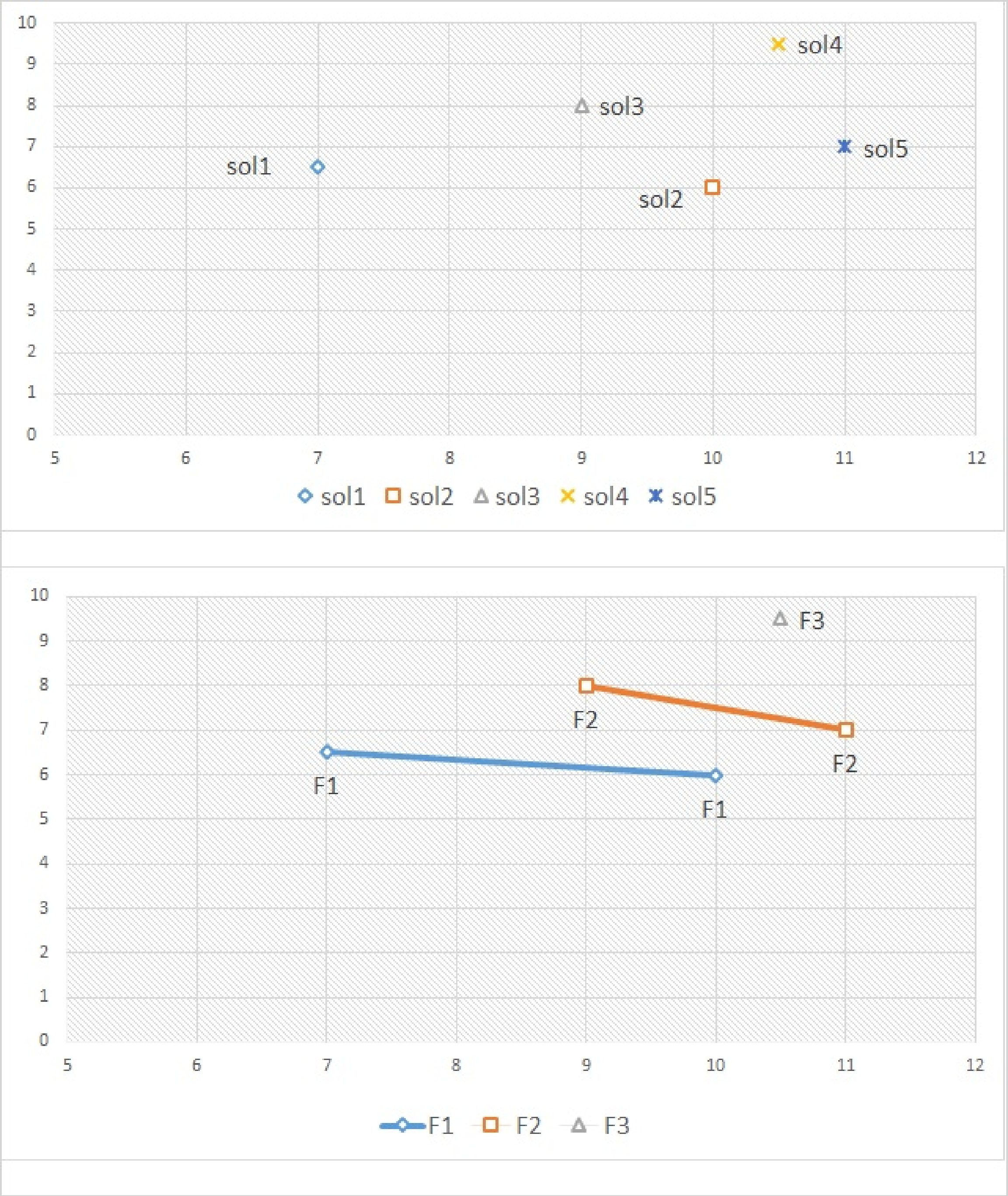


Figure 4.3: Sorting unlabeled data using NSGA-II.

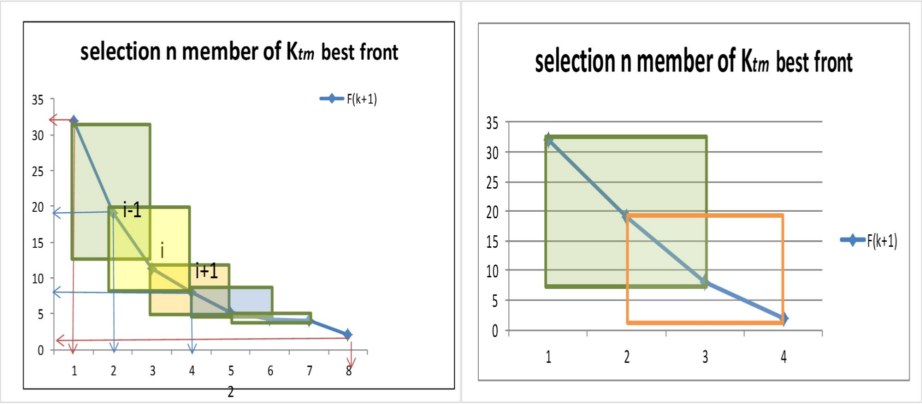


Figure 4.4: Selecting the best solution using CD value.

Table 4.6: NSGA-II CD Table Result

ID	Rank	F1	F2	CD
P1	1	0.233333333	3.843386081	INF
P2	1	0.35	3.500346535	INF
P6	1	0.408333333	2.969880625	1.208
P12	1	0.525	2.903817992	0.992
P3	2	0.35	3.84117628	INF
P5	2	0.408333333	3.810931956	INF
P7	2	0.425	3.624567817	1.3402
P10	2	0.491666667	3.571187739	0.87901
P8	2	0.433333333	3.634567817	0.0789
P4	3	0.391666667	4.014715698	INF
P11	3	0.5	3.705115924	INF
P9	3	0.441666667	3.805016513	0.8934

Table 4.7: Simple patients with multi features

ID	HR	Slope	CA	CP	Oldpeak	Thal	Chol	Rest BPS	Age	Exang	Sex	Restecg	FBS
P1	0.23	0.00	0.00	0.66	0.00	1.00	0.25	0.40	0.50	0.00	1.00	0.00	0.00
P2	0.35	0.00	0.00	0.33	0.00	0.00	0.25	0.30	0.50	0.00	1.00	0.00	1.00
P3	0.35	0.00	0.00	0.66	0.10	0.00	0.00	0.80	0.50	0.00	1.00	0.00	1.00
P4	0.39	0.00	1.00	1.00	0.00	0.00	0.25	0.10	0.50	0.00	1.00	0.00	1.00
P5	0.41	0.50	0.00	0.00	0.00	0.00	0.00	0.20	0.50	0.00	1.00	0.50	0.00
P6	0.41	0.50	0.00	0.66	0.00	1.00	0.00	0.40	0.50	0.00	0.00	0.50	0.00
P7	0.43	0.00	0.00	1.00	0.00	0.00	0.25	0.50	0.50	1.00	1.00	1.00	0.00
P8	0.43	0.50	0.00	0.66	0.30	0.00	0.00	0.20	0.50	0.00	1.00	0.50	1.00
P9	0.44	0.50	0.00	1.00	0.10	0.00	0.50	0.40	0.50	0.00	0.00	0.50	0.00
P10	0.49	0.50	0.00	0.00	0.30	0.00	0.50	0.60	0.50	0.00	1.00	0.00	1.00
P11	0.50	0.00	0.00	1.00	0.00	0.00	0.25	0.40	0.50	0.00	0.00	0.50	0.00
P12	0.53	0.00	1.00	0.66	0.00	0.00	0.50	0.40	0.50	0.00	1.00	1.00	1.00

Table 4.8: NSGA-II Rank Table Result

ID	Rank	F1	F2
P1	1	0.233333333	3.843386081
P2	1	0.35	3.500346535
P6	1	0.408333333	2.969880625
P12	1	0.525	2.903817992
P3	2	0.35	3.84117628
P5	2	0.408333333	3.810931956
P7	2	0.425	3.624567817
P8	2	0.433333333	3.634567817
P10	2	0.491666667	3.571187739
P4	3	0.391666667	4.014715698
P9	3	0.441666667	3.805016513
P11	3	0.5	3.705115924

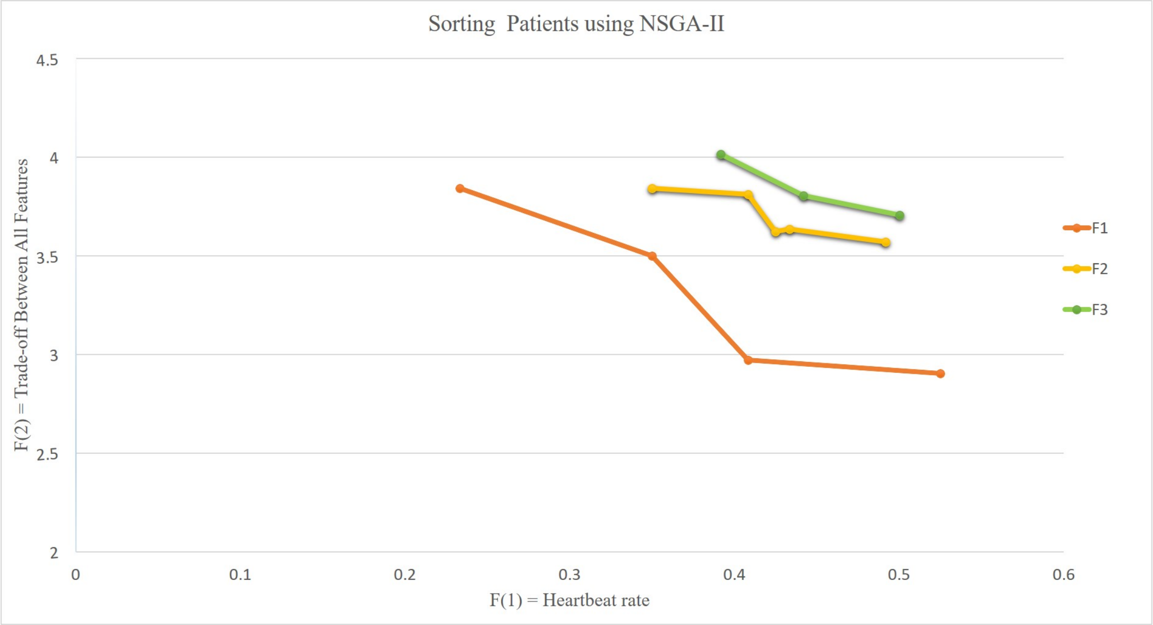


Figure 4.5: Sorting patients in different PFs using NSGA-II.

4.4 Sorting the New Population using NSGA-II

The new population consists of two generations, the parents population and children. New offspring (children or new members) will be generated using either Crossover or Mutation.

4.5 Offspring Generation

For creating a new generation, NSGA-II follows two main steps, parent selection and duplication, and divides this generation into two groups : crossover and mutant children. This is based on the main method which has been mentioned in [17]. This thesis followed the original steps for parent selection and generation. The original method contains a binary tournament for parent selection, and arithmetic crossover for crossover children and Gaussian mutation to generate mutant children. Using this configuration, NSGA-II will be able to find new solutions close to the Pareto-optimal front [18].

4.5.1 Parent Selection

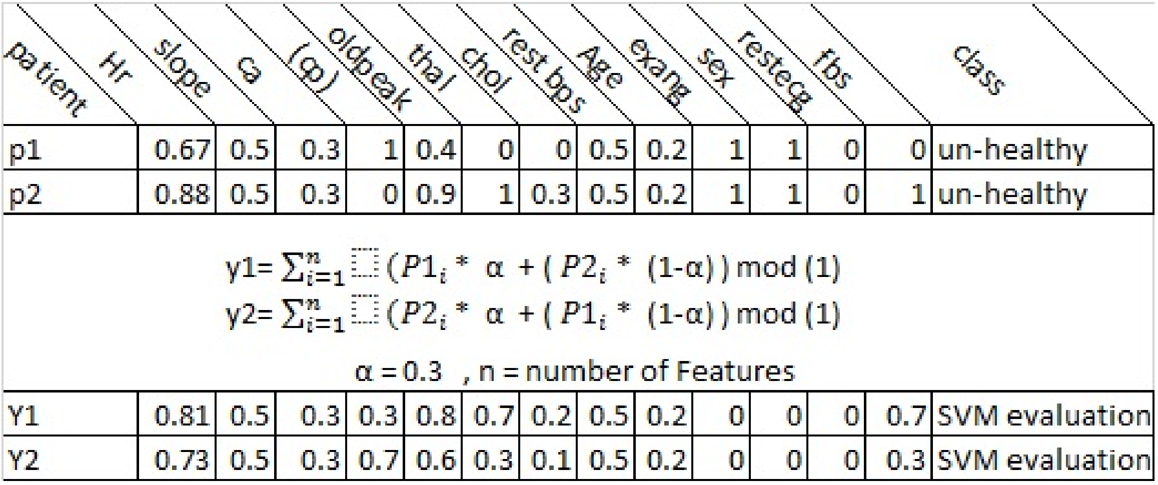
The original NSGA-II was proposed based on the Binary Tournament for parent selection: select two random members and send the most fitted one as a new parent for the offspring generation process. To generate new children, selected parents will be duplicated using the mutation or crossover method.

4.5.2 Arithmetic Crossover

Crossover is a genetic operator used to generate a new offspring or a new genotype, which is known as the reproduction of a new genome. Using this operator, GA can produce two new solutions (sample) from two parents. Based on the problem and the cost function, a number of crossover methods would work for GA, including single point, double points, and Uniform crossover. Following the proposed crossover operator in [18], Arithmetic Crossover has been chosen to simulate this thesis.

Arithmetic Crossover is a type of uniform crossover function which allows NSGA to generate new controllable offspring using equation 4.3. In this equation, $P_1\ and\ P_2$ have been selected based on a binary tournament in the parent selection step and subset of x_{11}, x_{12} , ... x_{1n} represents parents' features or parents' genome. α is a positive variable between 0 to 1 which is used to control new children's renovation. For example, children will have the average of their parents' features if α is equal to 0.5. When parent features could be any variable between 0 to 1, we bounded the children's features to 1. With this limitation, newly generated features will be evaluated, whether or not they are between 0 and 1. This operation is shown in Figure 4.6.

$$ArithmeticCrossover = \begin{cases} P_1 = x_{11}, x_{12}, x_{13}, ..., x_{1n},\ n =\ number\ of\ features \\ P_2 = x_{21}, x_{22}, x_{23}, ..., x_{2n},\ n =\ number\ of\ features \\ \alpha = N \Rightarrow 0 \leq N \leq 1 \\ Child_1\ \equiv\ \Sigma_{i=1}^{n}\ ((P_{1i} * \alpha) + ((P_{2i} * (1-\alpha))))\ \mod (1) \\ Child_2\ \equiv\ \Sigma_{i=1}^{n}\ ((P_{2i} * \alpha) + ((P_{1i} * (1-\alpha))))\ \mod (1) \end{cases} \tag{4.3}$$



patient	Hr	slope	ca	(cp)	oldpeak	thal	chol	rest bps	Age	exang	sex	restecg	fbs	class
p1	0.67	0.5	0.3	1	0.4	0	0	0.5	0.2	1	1	0	0	un-healthy
p2	0.88	0.5	0.3	0	0.9	1	0.3	0.5	0.2	1	1	0	1	un-healthy

y1= $\sum_{i=1}^{n} (P1_i * \alpha + (P2_i * (1-\alpha)) \bmod (1)$
y2= $\sum_{i=1}^{n} (P2_i * \alpha + (P1_i * (1-\alpha)) \bmod (1)$

$\alpha = 0.3$, n = number of Features

Y1	0.81	0.5	0.3	0.3	0.8	0.7	0.2	0.5	0.2	0	0	0	0.7	SVM evaluation
Y2	0.73	0.5	0.3	0.7	0.6	0.3	0.1	0.5	0.2	0	0	0	0.3	SVM evaluation

Figure 4.6: A quick view of crossover operation

4.5.3 Gaussian Mutation

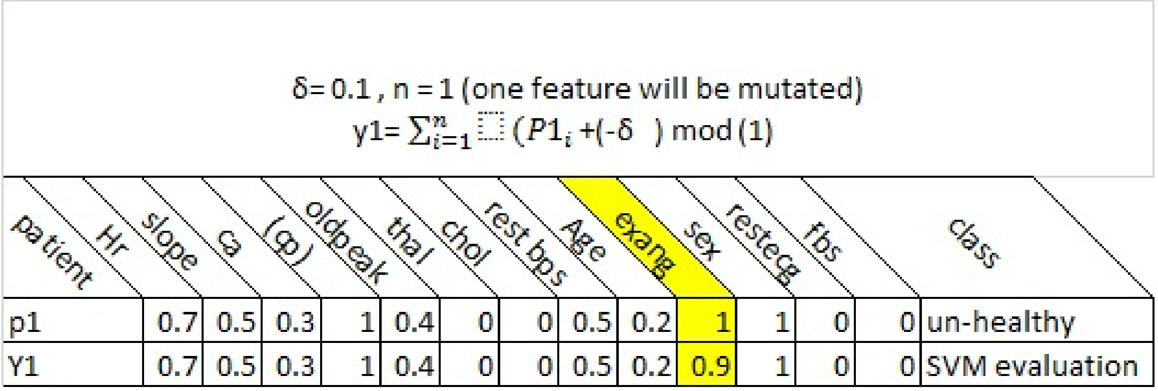
GA can generate new children using the mutation function as well. In contrast to the crossover function, the mutant children will be generated from a single parent. These new genotypes (children generated with a mutation function) will have the exact features (genotype) as their single parents except in μ percentage of genomes.

μ is a positive number between zero and one, which is denoted as a percentage of features which will have been affected by the mutation function. (i) is a random positive integer between 1 and the number of features, which shows the index of infected features, (in this thesis only one random gene is affected by mutation).

A selected random gene will be subtracted with σ unit when σ is a random positive number between 0 and 1. A random feature will be subtracted with σ and bounded with mod (1) to have a new child exactly the same as its single parent. The new features in both the crossover and mutation will be bonded with mode 1 when the dataset has been normalized between 0 and 1 and any negative number or any number greater than 1 should be bonded in this range. The new children might have a better health position in the graph.

If new children have a better phenotype than their parents or the other population's members, they will be selected for the next generation. In this thesis, 0 represents the best possible features, and 1 is the worst possible case. Equation 4.4 shows the methodology of Gaussian mutation. This function is visualized in Figure 4.7.

$$GaussianMutation = \begin{cases} P_x = \ x_{11}, x_{12}, x_{13}, ..., x_{1n}, n = \ number\ of\ features \\ i \rightarrow i \in \{N\}\ and 1 \leq i \leq n, n = number\ of\ features \\ \mu\ =\ presantage\ of\ mutant\ genum\ \Rightarrow 0 \leq \mu \leq 1 \\ N\mu = \ ([(n * \mu)]),\ number\ of\ infected\ featuers \\ Child_y\ \equiv\ P_x\ \rightarrow\ \Sigma_{J=1}^{N\mu},\ y_i = \ (x_i + (-\sigma)) \mod (1) \end{cases} \tag{4.4}$$



$\delta = 0.1$, $n = 1$ (one feature will be mutated)

$y1= \sum_{i=1}^{n} \square \, (P1_i +(-\delta \ ) \bmod (1)$

patient	Hr	slope	ca	(cp)	oldpeak	thal	chol	rest bps	Age	exang	sex	restecg	fbs	class
p1	0.7	0.5	0.3	1	0.4	0	0	0.5	0.2	1	1	0	0	un-healthy
Y1	0.7	0.5	0.3	1	0.4	0	0	0.5	0.2	0.9	1	0	0	SVM evaluation

Figure 4.7: A quick view of mutation operation

4.5.4 Evaluation

After each generation, new children have to be evaluated for class labeling. In our study, 2 evaluations have been used, a feature evaluation and class label (phenotype).

4.5.4.1 Feature Evaluation

Features evaluation means evaluating to make sure the new feature is in the dataset range. As can be seen in the equations 4.3 and 4.4 all new features have been bounded with mod 1. New features will have a correct value using this limitation. Moreover, some features have more restrictions such as gender (this feature can only accept 0 or 1), equation 4.5 shows the control for this feature. Using this control, any number between 0 and 0.499 will be treated as 0, and larger values will be replaced with 1.

$$FeatureControl = \begin{cases} x_i = \ \lfloor \ x_i * 2 \ \rfloor \\ if \ x_i \ > \ 1, \ x_i \ = 1 \end{cases} \tag{4.5}$$

4.5.4.2 Class label (Health Status) Evaluation

For the class label evaluation, the class label of the new offspring must be verified with the SVM algorithm as well. Using the SVM, it is required to introduce a training dataset to learn and a test dataset to evaluate. In our case, the training set is the raw dataset which is the ground truth, and the simulated data is evaluated as the test set. The optimal boundary then will be the set of points which have been labeled as unhealthy and dominate all other unhealthy points.

In this thesis, we have three different classification steps:

- **Dataset classification** within this step the accuracy of the raw data set has been calculated. The data set has been analyzed with the SVM. The results are presented in Table 3.2.
- **Evaluate the Pareto Front (optimal boundary curve)** In this part the generated members are evaluated one by one. First, a random set of data is selected. The accuracy of this set is 100%, which means we know that each member of this set belongs to either the healthy or unhealthy class. Next, a newly generated health position must be classified. The predefined class for these new positions is unhealthy. After we add a new position to our set, this new group will be our test set and will be evaluated with SVM. If the accuracy is still 100% the new member is in fact from the unhealthy class and will be added to the optimal boundary; otherwise, it will be re-labeled as healthy.
- **Tolerance threshold evaluation:** To explore the potential health tolerance threshold, a set of 3 random patients are presented in Table 5.1. The steps are

similar to the previous one, but this time the new generated points belong to a particular patient (in the healthy class) by increasing their dynamic features. All new health positions for this patient are labeled as unhealthy in the first step and these points are evaluated and relabeled if their accuracy result is less than 100%. The first point which gets the accuracy of 100% (which means this point belongs to the unhealthy class) will be the tolerance threshold for this patient.

Table 4.9: Tolerance threshold evaluation for P1 using SVM

Hr	Slope	Ca	Cp	Old Peak	Thal	Chol	Rest BPS	Age	Exang	Sex	Rest ECG	FSB	Given Class	Verified Class	TP Rate	FP Rate	Precision
0.02	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.04	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.06	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.08	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.1	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.12	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.14	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.16	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.18	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.2	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976

0.22	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.24	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.26	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.28	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.3	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.32	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.34	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.36	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.38	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.4	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.42	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.44	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.46	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.48	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976
0.5	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Healthy	1	0.022	0.976

0.52	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
0.54	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
0.56	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
0.58	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
0.6	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
0.62	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
0.64	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
0.66	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
0.68	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
0.7	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
0.72	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
0.74	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
0.76	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
0.78	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
0.8	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1

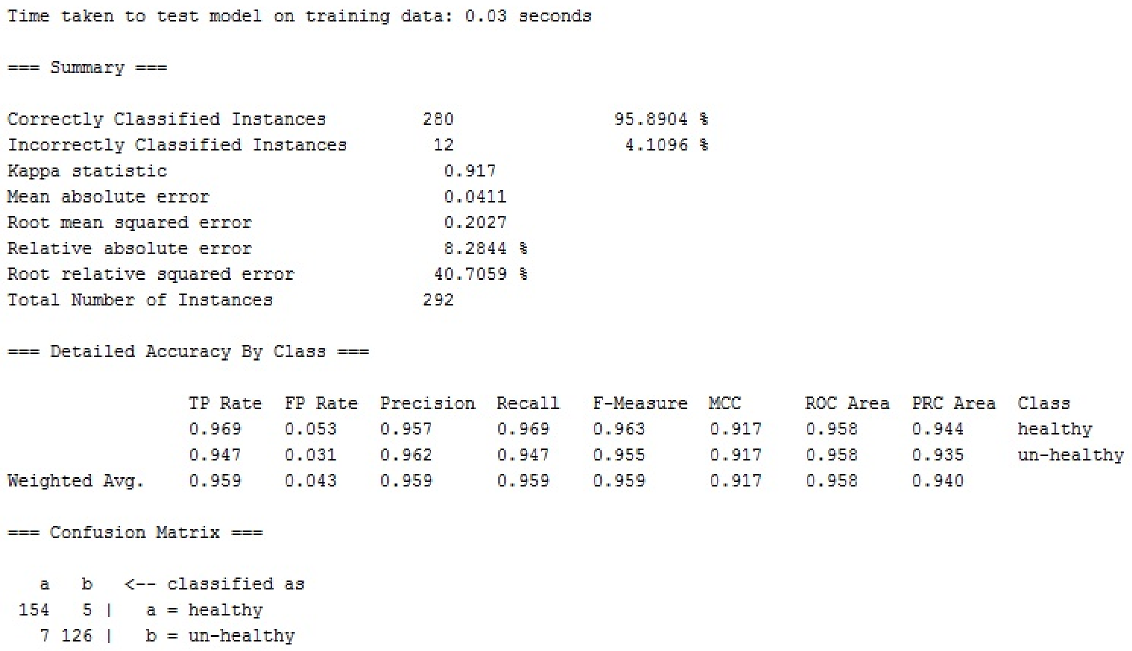
0.82	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
0.84	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
0.86	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
0.88	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
0.9	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
0.92	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
0.94	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
0.96	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
0.98	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1
1	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	Unhealthy	Unhealthy	1	1	1

The class label (classification) evaluation step mentioned in Table 4.9 has been shown in the following figures (4.8, 4.9, 4.10, 4.11 and 4.12). As is shown in Figure 4.8, the entire Cleveland dataset has been analyzed with the SVM as a Training set. SVM has created its own pattern for classifying the test data using this dataset. Following the evaluation processes, a set of patients is selected and named the safe-set. The accuracy of this set is presented in Figure 4.9. The safe-set has a random set of selected data with an accuracy of 100%, which means we know that each member of this set belongs to either the healthy or unhealthy class.

Next, a newly generated health position must be evaluated. In this step, all new health positions for this patient (here is P1, see Table 5.1) are labeled as unhealthy in the first step and these points are evaluated and relabeled if their accuracy result is less than 100%. Figure 4.10 shows the first new health position point (hr=0.02, BPS=0.2). As mentioned in the methodology, this point is added to the safe-set with the label of unhealthy, and then is evaluated with the SVM as a test set. Since the accuracy of this safe-set has dropped to 0.976%, the class label given to this new point was incorrect. Thus, this point is relabeled as healthy class and added to the safe set. All the new points have been evaluated with the same process.

The point with (hr=0.52 and BPS=0.2) is called the p1 health tolerance threshold at a maximum blood pressure of 0.2 BPS, since the class label given to this point is verified with the SVM. The evaluation of this point is shown in Figure 4.10. This step is repeated for all possible health positions for BPS=0.2. Figure 4.12, shows the evaluation of the last member of this line, and all the steps are presented in Figure

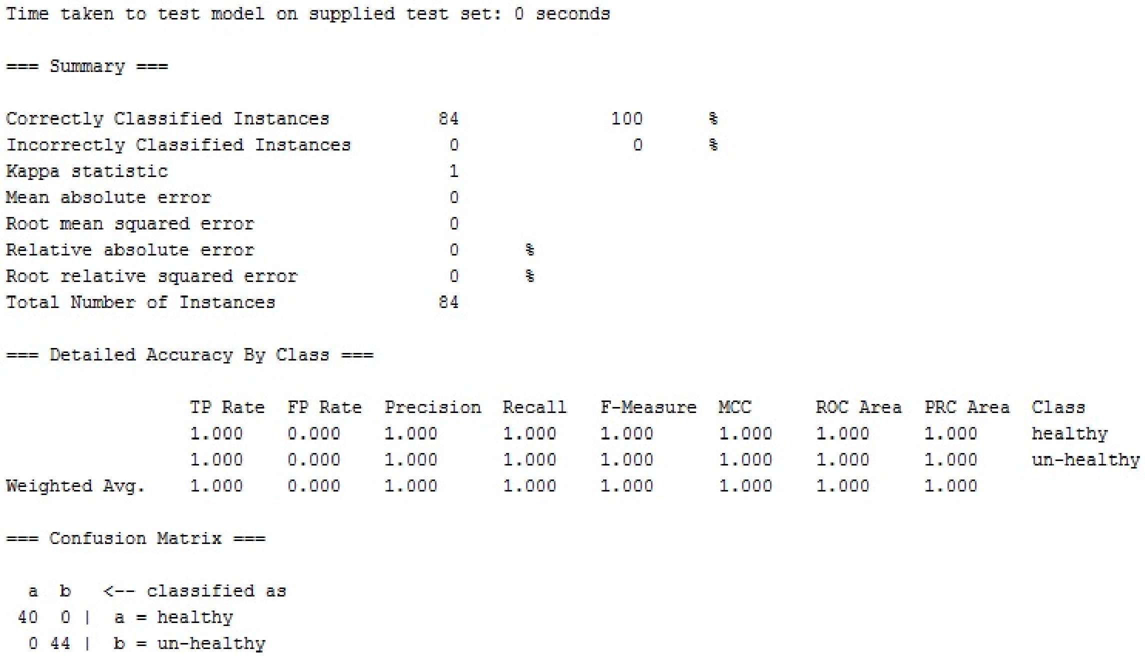
1.4. Next, the BPS for this patient is replaced by 0.3 and resimulated again to the new health tolerance threshold for the new BPS explored as well. All the possible health tolerance thresholds for this particular patient (P1) are denoted in Table 5.9.



```
Time taken to test model on training data: 0.03 seconds

=== Summary ===

Correctly Classified Instances         280               95.8904 %
Incorrectly Classified Instances        12                4.1096 %
Kappa statistic                          0.917
Mean absolute error                      0.0411
Root mean squared error                  0.2027
Relative absolute error                  8.2844 %
Root relative squared error             40.7059 %
Total Number of Instances              292

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.969    0.053    0.957      0.969    0.963      0.917    0.958     0.944     healthy
                 0.947    0.031    0.962      0.947    0.955      0.917    0.958     0.935     un-healthy
Weighted Avg.    0.959    0.043    0.959      0.959    0.959      0.917    0.958     0.940

=== Confusion Matrix ===

   a   b   <-- classified as
 154   5 |   a = healthy
   7 126 |   b = un-healthy
```

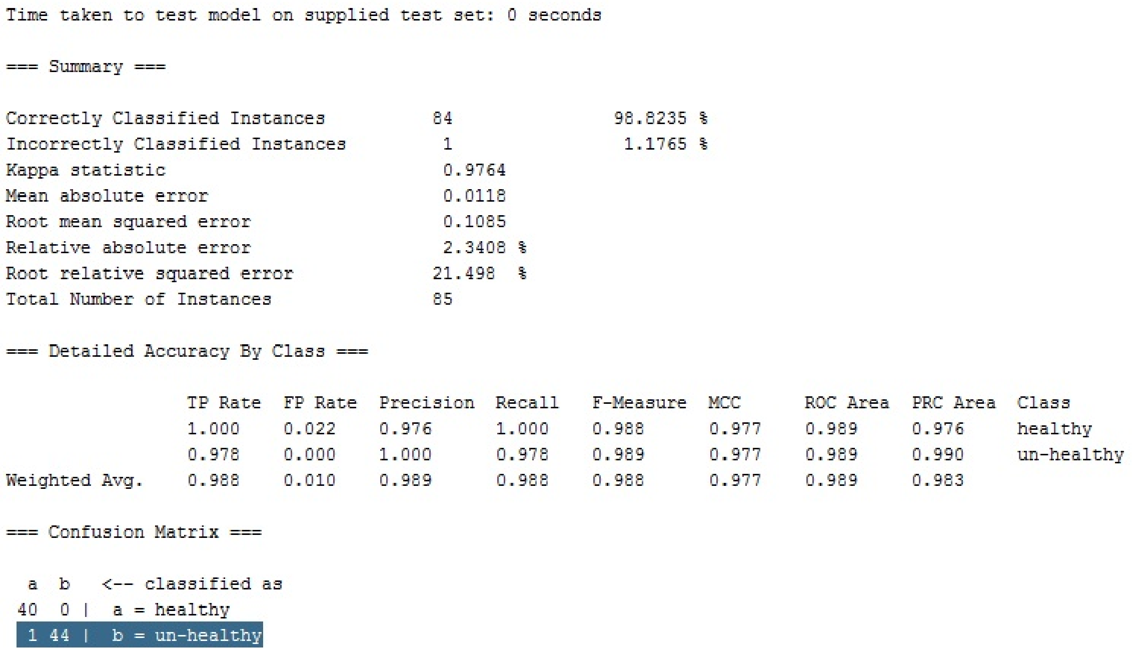
Figure 4.8: Dataset classification using SVM as a Train set

Figure 4.13 shows how a set of tolerance thresholds will be explored for each patient using a different range of dynamic features.



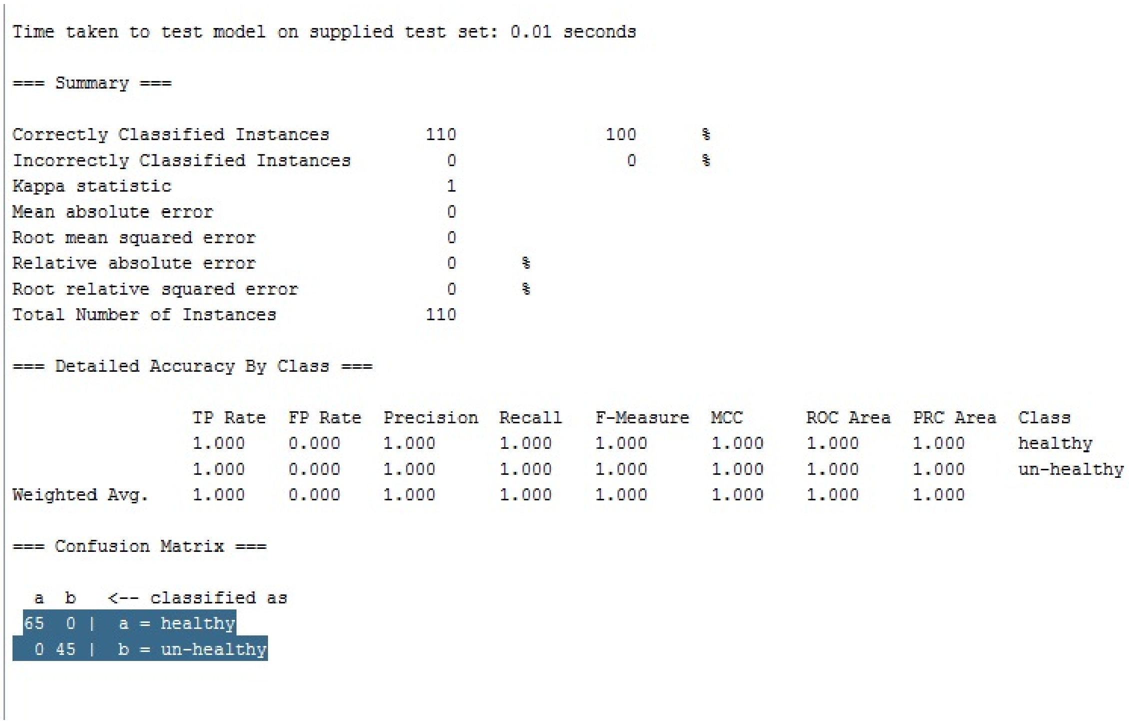
```
Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances          84              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances               84

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     healthy
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     un-healthy
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 40  0 |  a = healthy
  0 44 |  b = un-healthy
```

Figure 4.9: Safe-Sate classification using SVM as a Test set



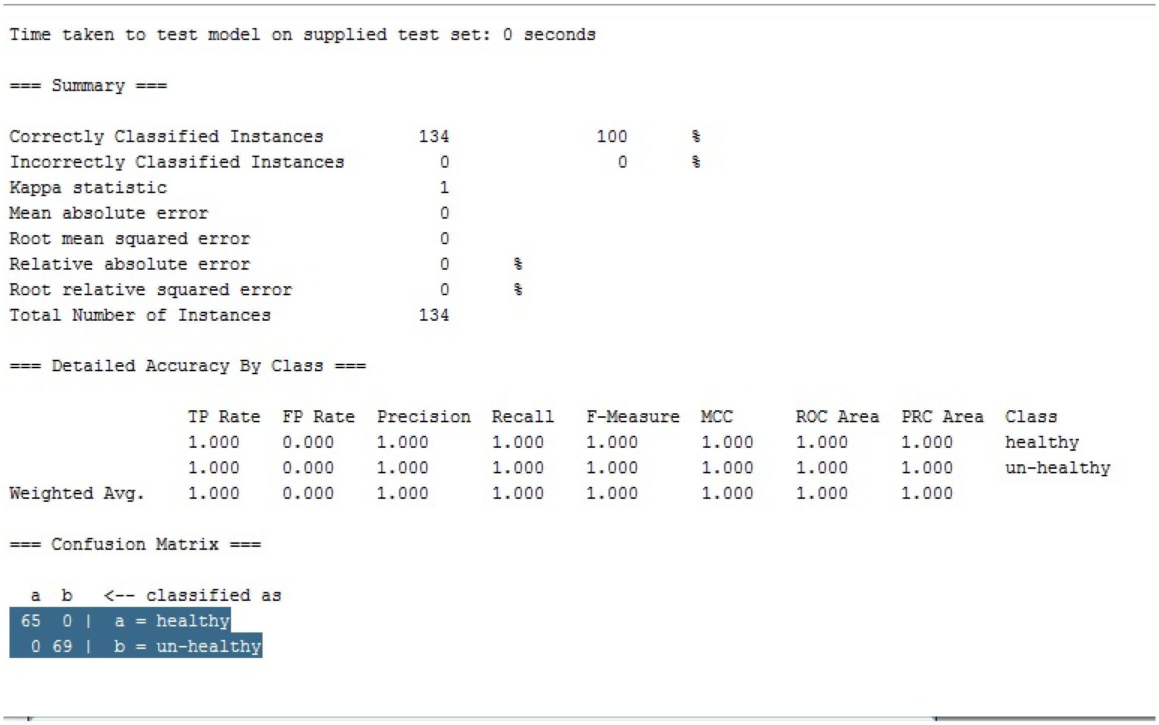
```
Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances          84               98.8235 %
Incorrectly Classified Instances         1                1.1765 %
Kappa statistic                          0.9764
Mean absolute error                      0.0118
Root mean squared error                  0.1085
Relative absolute error                  2.3408 %
Root relative squared error             21.498  %
Total Number of Instances               85

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.022    0.976      1.000    0.988      0.977    0.989     0.976     healthy
                 0.978    0.000    1.000      0.978    0.989      0.977    0.989     0.990     un-healthy
Weighted Avg.    0.988    0.010    0.989      0.988    0.988      0.977    0.989     0.983

=== Confusion Matrix ===

  a  b   <-- classified as
 40  0 |  a = healthy
  1 44 |  b = un-healthy
```

Figure 4.10: Evaluation for first health position's point for P1 using SVM as a Test set



```
Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

Correctly Classified Instances         110              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances              110

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     healthy
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     un-healthy
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 65  0 |  a = healthy
  0 45 |  b = un-healthy
```

Figure 4.11: Evaluation for Tolerance Threshold point using SVM as a Test set



```
Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances         134              100      %
Incorrectly Classified Instances         0                0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances              134

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     healthy
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     un-healthy
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 65  0 |  a = healthy
  0 69 |  b = un-healthy
```

Figure 4.12: Evaluation for last health position's points for P1 using SVM as a Test set

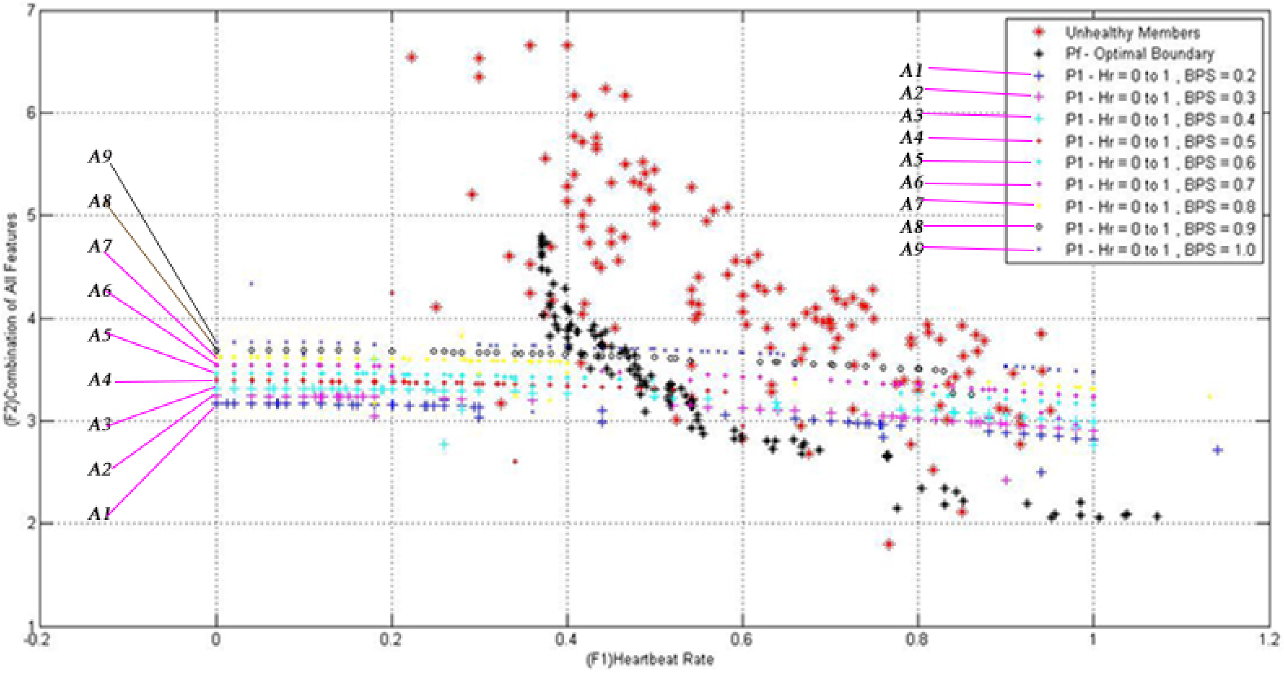


Figure 4.13: Tolerance threshold prediction for P1 using 13 features

This is a two-dimensional graph. The X-axis has been used for heart rate, and the Y-axis is a trade-off between all features, including static and dynamic features. In this graph, the red * shows the unhealthy members from the dataset. The black * line is the optimal boundary line which is generated by NSGA-II and evaluated with the SVM. All lines which start from the left side of the graph and continue to the right side are new health positions created with NSGA-II based on the new possible features (Heart Rate and Blood Pressure).

In each line of the Figure 4.13, only the HR feature is increased and as can be seen the patient's health might be at risk if their heart rate approaches the PF-optimal boundary line. For example, the P1 in the first line (A1 BPS=0.2 means 11 BPS) will experience chest pain if their heart rate increases from 0 to 0.52 (133 bpm); see Table 4.1. The changing variable between each line is the blood pressure feature. The next lines represent the same patient status when the resting blood pressure is increased. For example, Line A1 has a BPS equal to 0.2 and line A2 has a BPS equal to 0.3 and in line A3 the BPS is equal to 1. Note that if the maximum blood pressure is increased to 1.0 BPS, and P1 can accept a maximum bpm of 0.38 (114 bpm).

Chapter 5

Findings and Discussions

In this thesis, a new data processing method to predict patients' actual health status, named Self-Coach, is presented. This application can predict patients' health status immediately from their activity in linear time. Using an optimization technique allows pre-processing for each patient individually. By monitoring the dynamic data, patients' health conditions will be monitored before their maximum tolerance thresholds are reached. The goal of proposing this application is to discover the health tolerance threshold for each patient. An excellent example to show the patient's tolerance threshold is to imagine the person's health condition as an empty glass. In an example where there are two empty glasses with different capacity, the capacity represents patients' healthy phenotype and the overflow point for each one is their tolerance threshold point.

Figure 5.1 shows two patients with different tolerance thresholds. These capacities are assumed to be their maximum acceptance of tolerance limits for the dynamic data (heart rate, blood pressure, and the stress factor). These maximum points have been

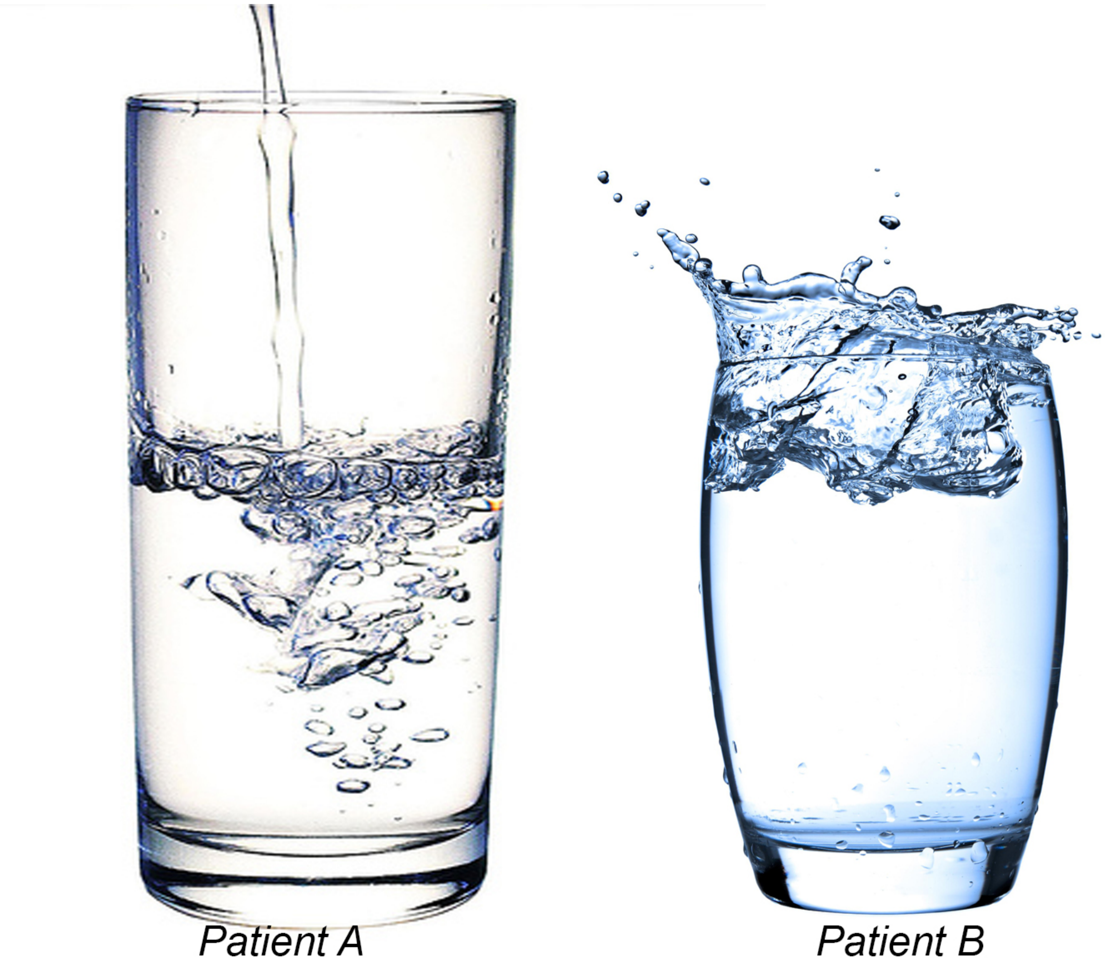


Figure 5.1: Accepting heart rate tolerance threshold

created based on static data (see section 4.1 to 4.5). In this scenario, both patients have a healthy phenotype and the same activities. As can be seen, while patient B overflows with 150 cc of water (here, the heart rate) Patient A still has a healthy phenotype (i.e can manage a higher heart rate). A real story about C.L; a 19 year old US soldier is a very good example of this objective.

C.L's Life Story

C.L's Arrhythmia during Army Boot Camp [13] The subject was a 19-year-old United States' soldier, who successfully graduated from army school. However, he had a problem with long runs wearing full gear while he passed this processed during his training. His heart was starting to race with shortness of breath, and he would become dizzy and pass out when he had to run with full gear. He had two incidents when the results of the infirmary exam were unclear. After his second incident, C.L. went home on sick leave for additional follow-up.

The results of his chest examination after a month using a Holter monitor showed that he may have had an abnormal heartbeat when he did not have any difficulties during normal activities. C.L's story demonstrates the different capacities for various healthy classes. He was passing out during his duty because his body could not accept that much activity.

In this chapter, the same phenotypes (same classes) show various acceptable heartbeat thresholds in different situations. When a normal resting heart rate for a patient in the healthy class ranges between 60 to 100 bpm, considering the body's features such as fitness level, activity level, air temperature, emotions, body position (standing up or lying down), medications and body size, this can cause a different acceptable tolerance of the heart rate (100 to 180). The results of this simulation for exploring the patient's acceptance tolerance thresholds of dynamic features have been presented in the following sections. In these sections, the trade-off between the static and dynamic data and their effect on the patient's health classes are presented.

As mentioned in CL's Life Story, patients' statuses are subject to change based on new conditions. So under specific condition, any healthy patients can experience critical situations and can be re-categorized in the unhealthy class. A tolerance threshold, which has been discussed in this thesis, is a new health position for the patient representing the new health status, which would be dominated by at least one point in the optimal boundary. A change in any feature results in a new health position.

Technically, only some features can be modified based on daily conditions such as activity, using medicine, stress or emotional level. The features that do not change suddenly such as age, sex, and weight have been categorized as static features. On the other side, we have dynamic features which include features with flexible curves due to the patient's daily activity. Heart rate, blood pressure, and blood sugar are categorized in this class.

The tolerance threshold points for each patient are points which show that the patient's health status could be categorized in the unhealthy class based on the patient's new conditions. Indeed, the patient might experience chest pain or heart failure if its health position is dominated by one of these non-dominated points in the optimal boundary curve. In other words, by increasing the patient's dynamic features, their health position will tend to the right side of the graph. The first point which is dominated with the optimal boundary points will be recorded as the patient's tolerance threshold.

Following the dataset limitation, we did not have enough features for the study, and some features such as skin temperature, stress, respiration rate, sweating rate the patient's activity, and their medical usage have not been mentioned in our dataset. Thus, the only available dynamic data are heart rate and blood pressure. These features have been used in NSGA-II to generate new possible health positions for each particular patient. These two factors are not only subject to sudden changes but have also been found to be critical features for heart disease problems [23]. The effect of these features on heart disease has been analyzed in [66]. The author also mentioned the importance of reducing blood pressure in risky situations, regardless of the use of anti-hypertensive medication [23].

In the real case, heart rate and blood pressure are not the only features that can lead to heart failure, but as mentioned in [23], blood pressure is a critical feature in cases of heart failure prediction. In this reference, blood pressure, smoking, total cholesterol (TC), LDL-C, HDL-C, and diabetes have been categorized as the most common risk factors of Coronary Heart Disease (CHD). Based on their experimental results, blood pressure has been categorized in 5 phases, and based on each condition (the trade-off between the other mentioned features) this factor was the key to actual health status.

1. optimal (systolic,120 mm Hg and diastolic ,80 mm Hg)

2. normal blood pressure (systolic 120 to 129 mm Hg or diastolic 80 to 84 mm Hg)

3. high normal blood pressure (systolic 130 to 139 mm Hg or diastolic 85 to 89 mm Hg)

4. hypertension stage I (systolic 140 to 159 mm Hg or diastolic 90 to 99 mm Hg)

5. and hypertension stage IIIV (systolic 160 or diastolic 100 mm Hg).

based on their experimental results, when systolic and diastolic pressures fell into different categories, the higher category was selected for the purposes of classification. As can be seen even in the medical prescription, blood pressure categorization was made regardless of the use of anti-hypertensive medication. However, in [23] the effect of blood pressure on CHD was explored, but they did not mention any method to use to predict or pre-analyze the actual tolerance threshold for a particular patient. The Self-Coach application which has been proposed in this thesis addresses this problem.

5.1 Simulation Results

The results of these simulations have been presented in the following sections. Each part includes the following phases:

- To evaluate the accuracy of the dataset, the SVM algorithm is used for classification. The result of each classification has been shown in the corresponding table, entitled the classification table, using SVM and including the related steps.

- To visualize the patient's position in the 2-D graph, the healthy and unhealthy classes are simulated separately. In all graphs, healthy and unhealthy patients have been illustrated with green and red, respectively.

- The black * line shows the PF set of the unhealthy class after 50 iterations, which is named the optimal boundary. This line represents the best possible positions for this class. These positions have been evaluated with the SVM algorithm. The results of this evaluation are shown in Tables 5.3, 5.4, 5.6, 5.7, 5.8 and 5.9.

- Curves with a + sign show new positions for individual patients. In these lines, the dynamic data for selected individuals are increased. These lines visualize the possible positions for the particular patient due to new features.

- Finally, the tolerance threshold for each individual (blue line curves) is evaluated in the threshold evaluation table.

In the following sections, three phases of our simulations are presented, which have been mentioned in the NECEC IEEE Conference (using 6 features) [19], the Aldrich Conference at Memorial University of Newfoundland (using 10 features) [20], and finally the presentation for the E-health conference in Vancouver on June 8, 2016 (using 13 features) [21]. As can be seen, with time, the accuracy of this prediction has improved due to using more related features. For example, in the first simulation ($NECEC - 2015$) six features were used, which are mentioned in [3]. Ten features were used for the Aldrich Conference, and subsequently the fitness function was modified to support more features. As a result of this modification, the accuracy of these curves and patients' positions have been improved. To have a comprehensive comparison between these three steps, groups of three patients have been chosen randomly. Table 5.1 shows these members and their features.

Table 5.1: Tolerance threshold extraction

Patient	Hr	Ca	Cp	oldpeak	exang	BPS	slope	thal	chol	Age	restecg	sex	fbs	Extracted HR for	Extracted HR for	Extracted HR for
														NECEC	Aldrich	E-health
P1	0	0.33	0.66	0	0	0.2	0.5	0	0.5	0.3	0	0	0	0.45	0.44	0.5
P2	0.25	0	0.66	0.1	1	0.4	0	0	0.25	0.3	0	1	0	0.43	0.44	0.52
P3	0.33	0	0.33	0	0	0.6	0	0	0.5	0.9	0	1	0	0.35	0.50	0.46

5.1.1 Simulation With 6 Features

All metrics and the dataset have been presented in Table 5.2. Using $ZDT4$ as the cost function requires applying some normalization to the real dataset. Due to this requirement, all data have been normalized between 0 and 1. In this simulation a heart-disease dataset (Cleveland) is utilized; this database contains 13 attributes which are extracted from a larger set of 75 attributes [4]. It is based on [3], six features which have the most correlations with the class label have been chosen for this simulation.

Table 5.2: Simulation setup

Simulator software	Matlab 2011
Dataset	cleveland
Normal cases	150
Abnormal cases	120
Mutation ratio	0.2
Parent selection	binary tournament
Number of generations	50
Selected population	100

Table 5.4 shows the accuracy of the selected features. For this classification,the SVM algorithm in WEKA is used.

Figure 5.2 shows patients' positions, the new tolerance threshold for these selected individuals and the non-dominated set using $NSGA-II$ with these selected features.

A black * curve (PF-optimal boundary) shows, the new offspring have been sorted in the first Pareto front. These are not real positions but they present the best combinations of human features that can exist for the unhealthy class. Table 5.4 shows the evaluation of these points. As expected, the accuracy of these points (PF-optimal boundary) is very close to 97%. The results of this section are presented in [19].

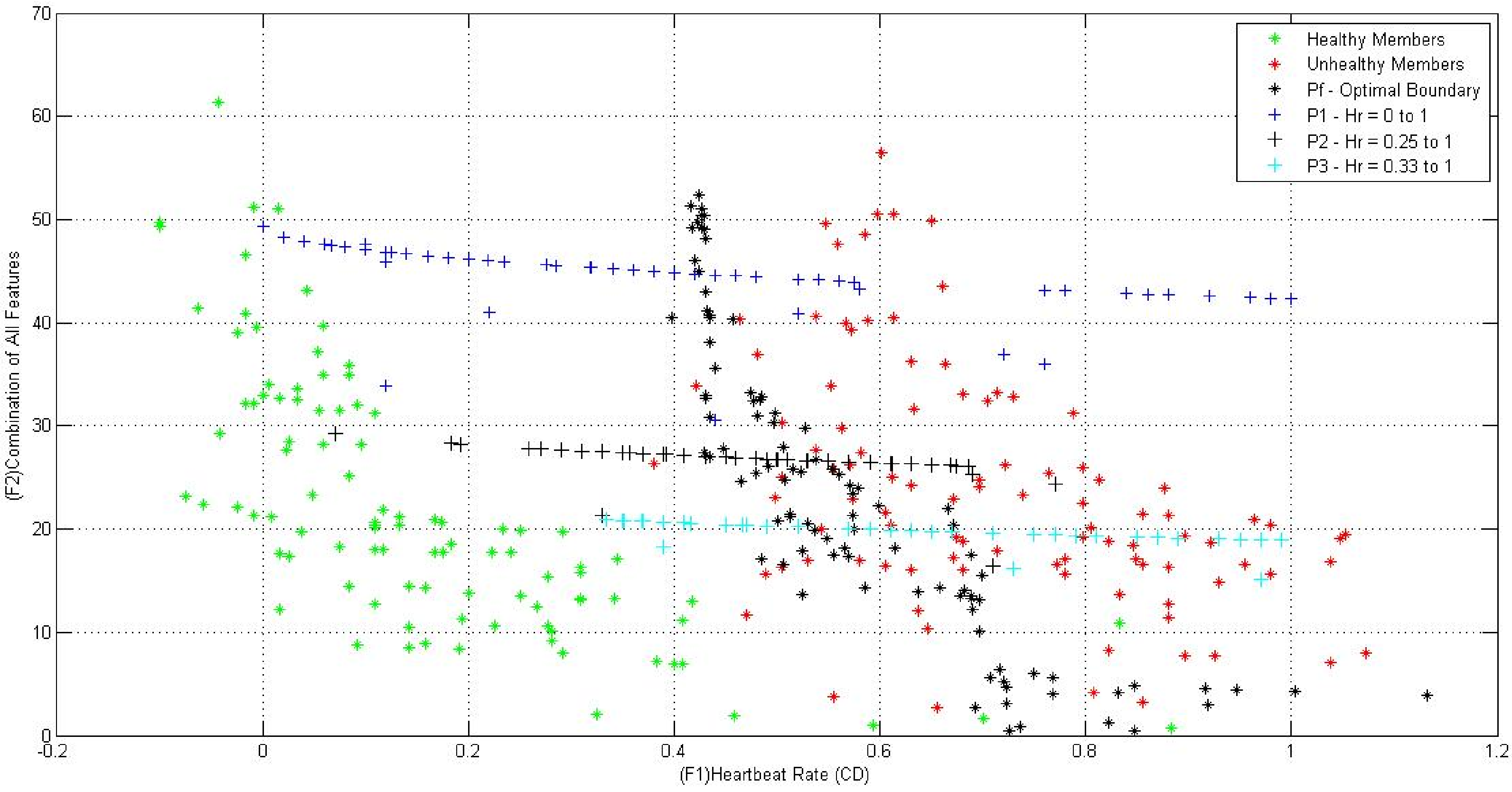


Figure 5.2: Patients' position using 6 features

Table 5.3: Tolerance threshold for selected patients using 6 features

Title	Normal HR	Predicted Tolerance HR Using NSGA-II	Predicted Tolerance HR Verified with SVM
P1	0	0.45	0.43
P2	0.25	0.43	0.42
P3	0.33	0.50	0.52

As mentioned above, healthy patients are shown with a green color and unhealthy phenotypes in red. The black + is the best PF after six iterations. This line shows that a healthy phenotype can be turned into an unhealthy case by crossing this line. The explored HR tolerance thresholds for these patients are shown in Table 5.3.

Table 5.4: Classification table Using SVM for 6 features

Title	True Positive Rate	False Positive Rate	Precision
Whole features	0.953	0.049	0.954
All members using 6 features	0.927	0.073	0.927
F1 members iteration 1 to 6	0.909	0.09	0.909

5.1.2 Simulation With 10 Features

For this simulation, metrics and patients have not changed from the previous section. Due to some comments, received during the previous conference[19], the number of selected features has increased for this simulation and some neglected features such as Age and Thal (a dataset's feature) have been added to the parents' features to increase the accuracy of the prediction. In this simulation 10 features have been selected because of the limitation of using the ZDT4 as a fitness function. To solve this step, features have been analyzed with a feature selection technique (SVMAttributeEval). As a result of this feature analysis, the top ten features which have shown most correlations with the dataset class have been selected.

Table 5.5 shows the attributes' ranks using this feature selection technique. The top ten selected features for this simulation are: Heart Rate (HR), Slope, Number of Colored Vessels (CA), Chest Pain (CP), Old Peak, Thal, Cholesterol, Age, BPS, and Exercise Induced Angina (Exang).

Table 5.5: Feature selection ranking table using SVMAttributeEval algorithm

Ranked	Feature ID	Attributes
14	1	Heart Rate (HR)
13	11	Slope
12	12	Number of Colored Vessels (CA)
11	3	Chest Pain (CP)
10	10	Old Peak
9	14	Thal
8	6	Cholesterol
7	2	Age
6	13	Sex
5	9	Exercise Induced Angina (Exang)
4	4	Resting Blood Pressure (Rest BPS)
3	8	Rest ECG
2	5	Col-normal
1	7	Fasting blood sugar (FSB)

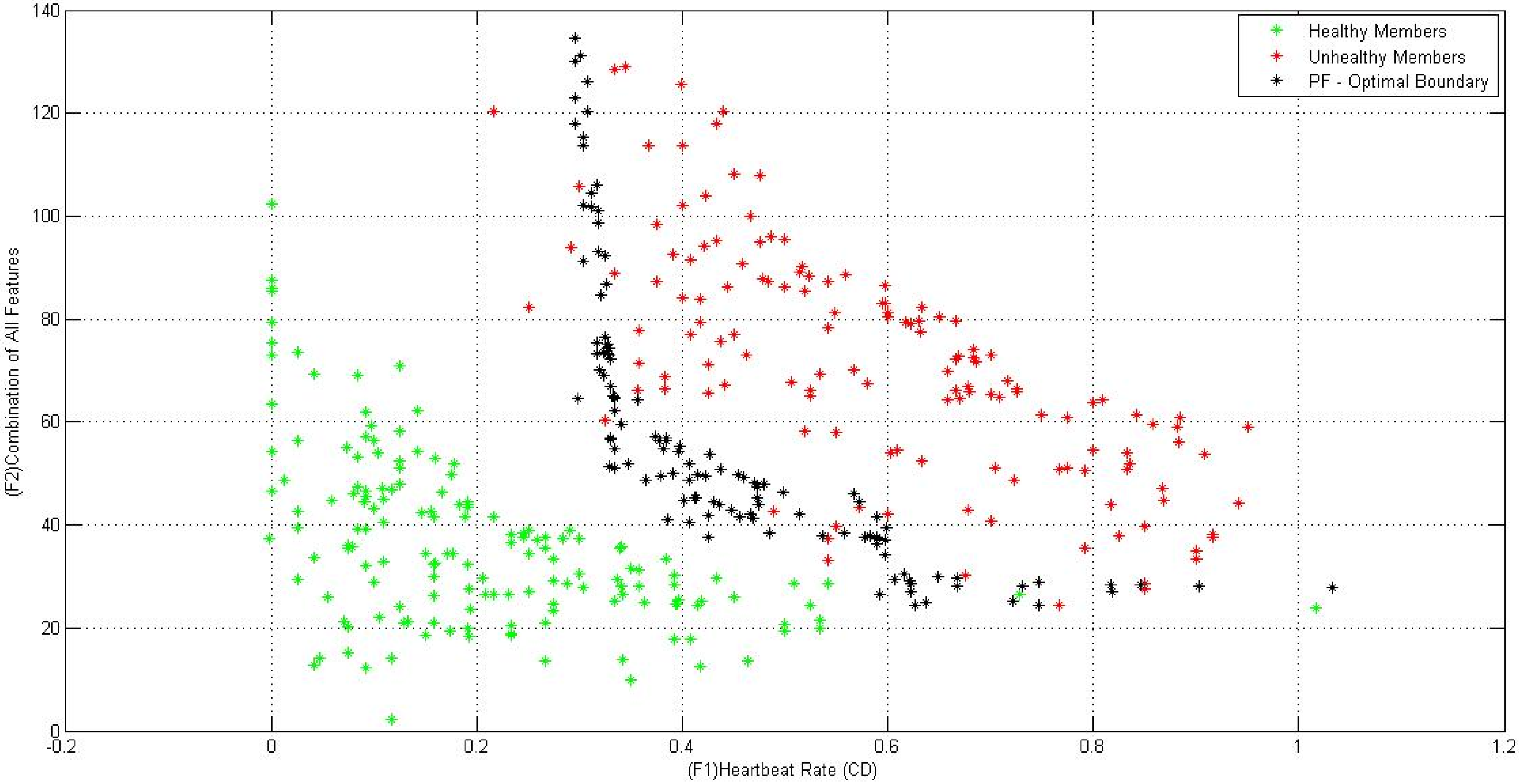


Figure 5.3: Patients' positions using 10 Features

Table 5.6 shows the accuracy of the selected features using the SVMAttributeEval algorithm. The accuracy of the new acceptable threshold for selected patients have been evaluated in Table 5.7. Figure 5.3 shows the patients' positions using 10 features.

Table 5.6: Classification table Using SVM for 10 features

Title	TP Rate	FP Rate	Precision
Whole features	0.953	0.049	0.954
10 selected features	0.93	0.07	0.93
Unhealthy PF	0.931	0.071	0.931

Tolerance Threshold positions for New Random Patients:

To predict the tolerance threshold for each healthy patient the following steps are required:

- Select the random patient from the healthy class.
- Increase the Hr feature to 1.0.
- Simulate the new dataset with the NSGA-II.
- Extract the intersection position with the unhealthy class.
- Evaluate the new point with the SVM.

Figure 5.4 demonstrates these points for the selected patients. For example, the health status for the P1 has been changed at point A (0.44,46.56). Signifying that this user might have a chest pain experience when the heartbeat approaches 130 bpm.

To evaluate the predicted point for this particular patient, all new records have been added to the dataset and re-analyzed with the SVM algorithm. Table 5.6 shows the accuracy of this prediction. The results of this simulation have been presented in [20].

Table 5.7: Tolerance threshold for selected patients using 10 features

Title	Normal HR	Predicted Tolerance HR Using NSGA-II	Predicted Tolerance HR Verified with SVM
P1	0	0.44	0.42
P2	0.25	0.45	0.43
P3	.33	0.52	0.51

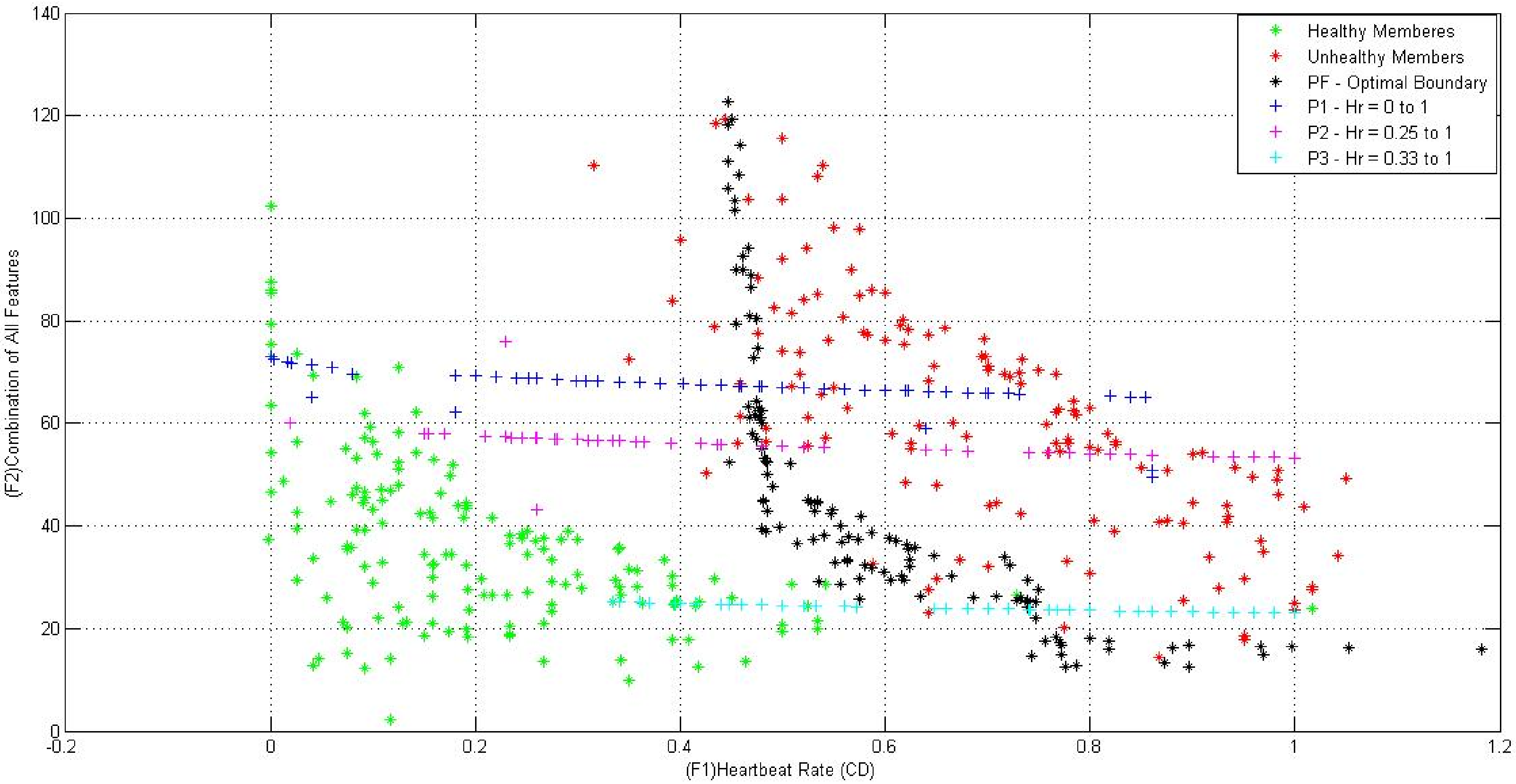


Figure 5.4: Tolerance threshold prediction for selected patients using 10 Features

5.1.3 Simulation With 13 Features

Since patients' features are limited to ten because of the use of the ZDT4 as a fitness function, to increase the accuracy of prediction and to have more functionality, ZDT2 is selected instead of ZDT4. This function also provides the conventional curves as accurate as ZDT4 and the results of this algorithm can also be used as the PF curve. As the previous simulations, all metrics and the dataset have been presented in Table 5.2. Using ZDT2 as the cost function requires applying some normalization to the dataset. Due to this requirement all data have been normalized between 0 to 1. In this simulation, a heart disease dataset (Cleveland) is utilized; this database contains 13 attributes, which have been extracted from a larger set of 75 attributes [4].

Table 5.8: Classification table using SVM for 13 features

Title	TP Rate	FP Rate	Precision
Whole features	0.953	0.049	0.954
Unhealthy PF	0.951	0.051	0.951

Figures 5.5 to 5.8 show the patients' positions using NSGA-II . In this dataset, two sets of features for each patient are monitored along the X and Y axis:

- The static data include chest pain type, resting blood pressure, exercise-induced angina, old peak, the number of colored vessels and age. The Y-axis shows the trade-off between these values.
- The dynamic data refers to the heart rate, which has been shown by the X-axis.

While using classification methods only determines whether or not the patient is in a healthy state, by applying an optimization approach the tolerance threshold of each patient is explored in linear time as well. Figure 5.5 shows the patients' positions using ZDT2 as the fitness function. In this figure, the green color represents a healthy status. Unhealthy phenotypes have been shown in red color and the black points are the unhealthy PF, which represents a potential tolerance threshold for a healthy patient. For this simulation, all selected samples have two dynamic features: HR and resting blood pressure (BPS). Figures 5.6, 5.7 and 5.8 show the results of this simulation. In each of these graphs, a patient has been simulated with a range of data, with Hr up to 1 and BPS also up to 1. The explored potential tolerance thresholds for these patients are evaluated in Table 5.9.

In each line only the HR feature is increased and as can be seen, the patients' health might be at risk if their heartbeat approaches the PF-optimal boundary line. For example, the P1 in the first line (BPS=0.2 means 11 BPS) will have an experience of the chest pain if the heart rate increases from 0 to 0.52 (133 bpm); see table **??**. Next lines represent the same patient status when the resting blood pressure is increased. Note that if the maximum blood pressure increased to 1.0 BPS, P1 can accept a maximum bpm of 0.38 (114 bpm). The results of this simulation have been presented in [21].

Table 5.9: Tolerance threshold for selected patients using 13 features

Patient	HR	BPS	Predicted Tolerance HR Using NSGA-II	Predicted Tolerance HR Verified with SVM
P1	0	0.2	0.52	0.52
P1	0	0.3	0.51	0.51
P1	0	0.4	0.47	0.47
P1	0	0.5	0.46	0.46
P1	0	0.6	0.43	0.42
P1	0	0.7	0.42	0.42
P1	0	0.8	0.40	0.40
P1	0	0.9	0.39	0.38
P1	0	1.0	0.38	0.38
P2	0.25	0.4	0.52	0.50
P2	0.25	0.5	0.49	0.49
P2	0.25	0.6	0.47	0.46
P2	0.25	0.7	0.46	0.44
P2	0.25	0.8	0.44	0.43
P2	0.25	0.9	0.42	0.42
P2	0.25	01.0	0.40	0.40
P3	0.33	0.6	0.53	0.52
P3	0.33	0.7	0.51	0.51
P3	0.33	0.8	0.49	0.49
P3	0.33	0.9	0.48	0.47
P3	0.33	1.0	0.46	0.44

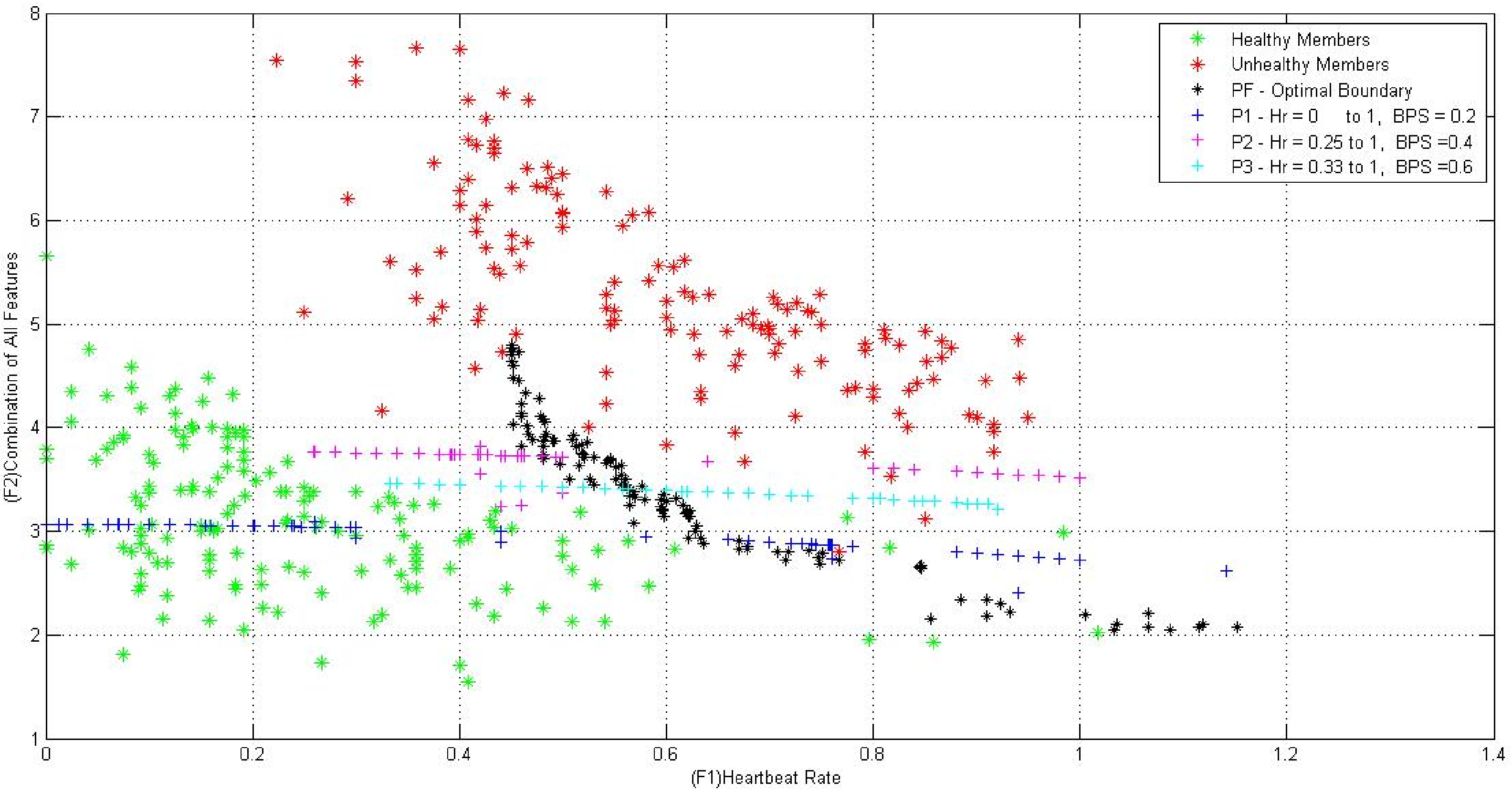


Figure 5.5: Patients' positions using 13 features

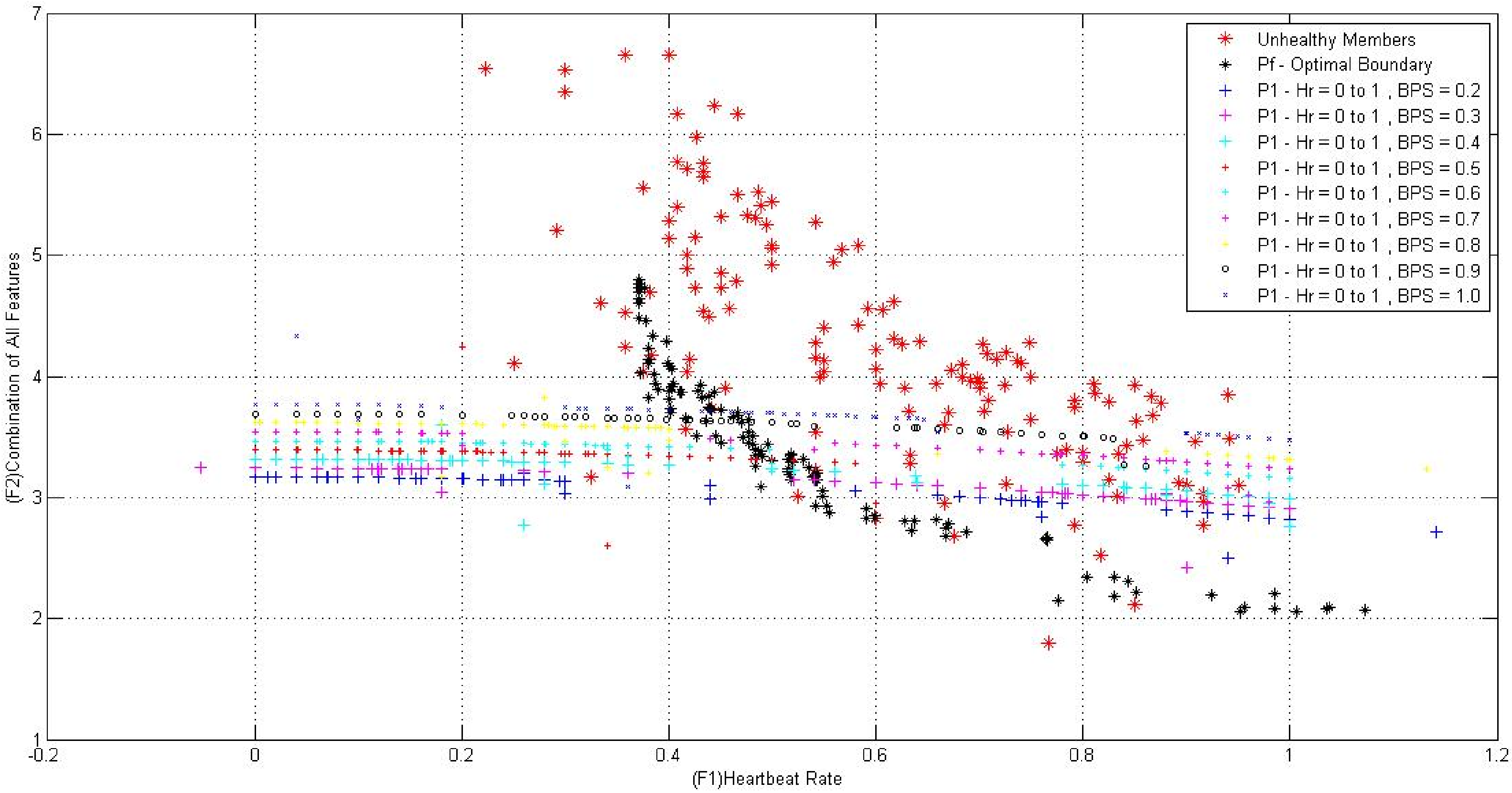


Figure 5.6: Tolerance threshold prediction for P1 using 13 features

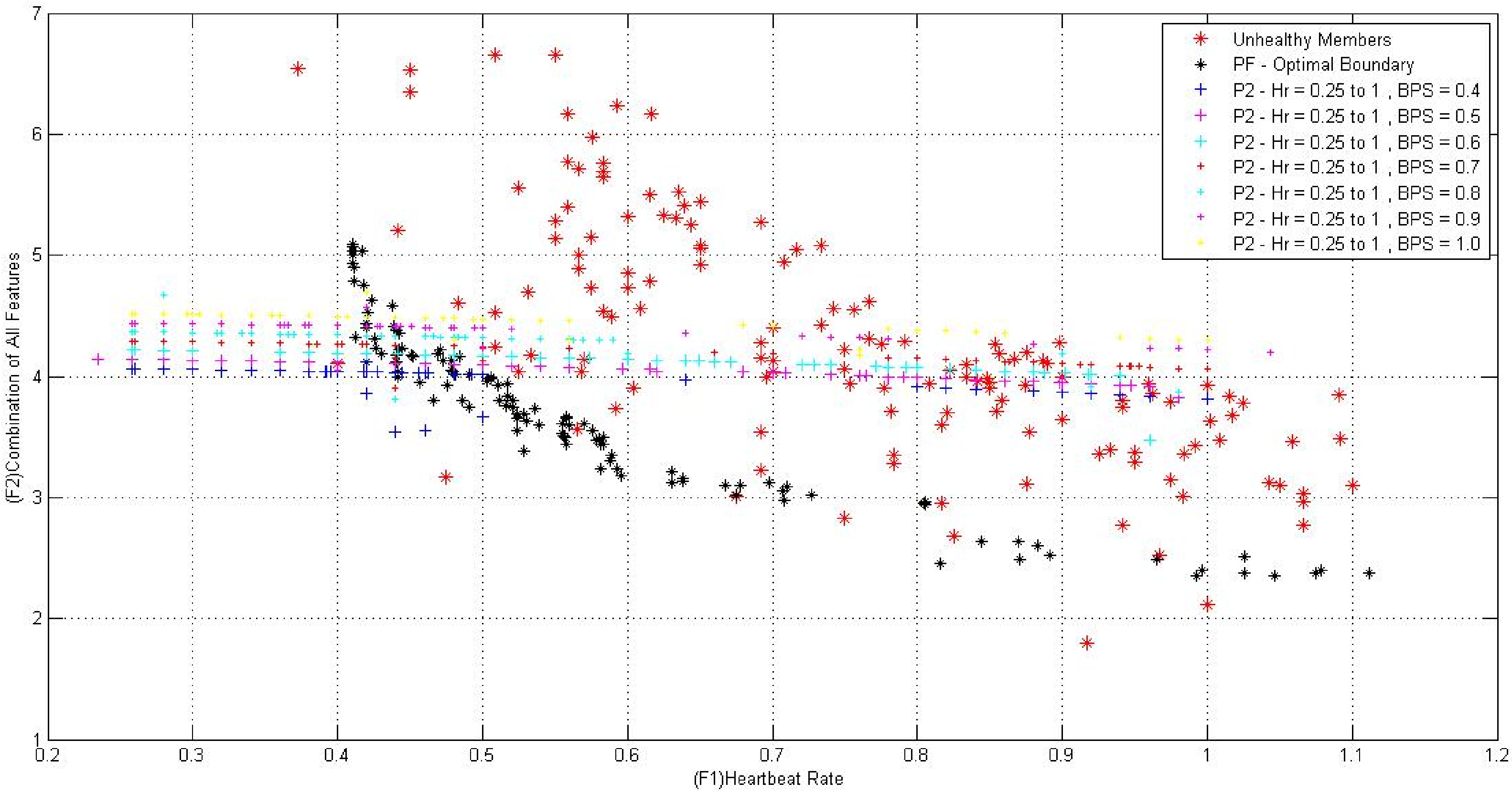


Figure 5.7: Tolerance threshold prediction for P2 using 13 features

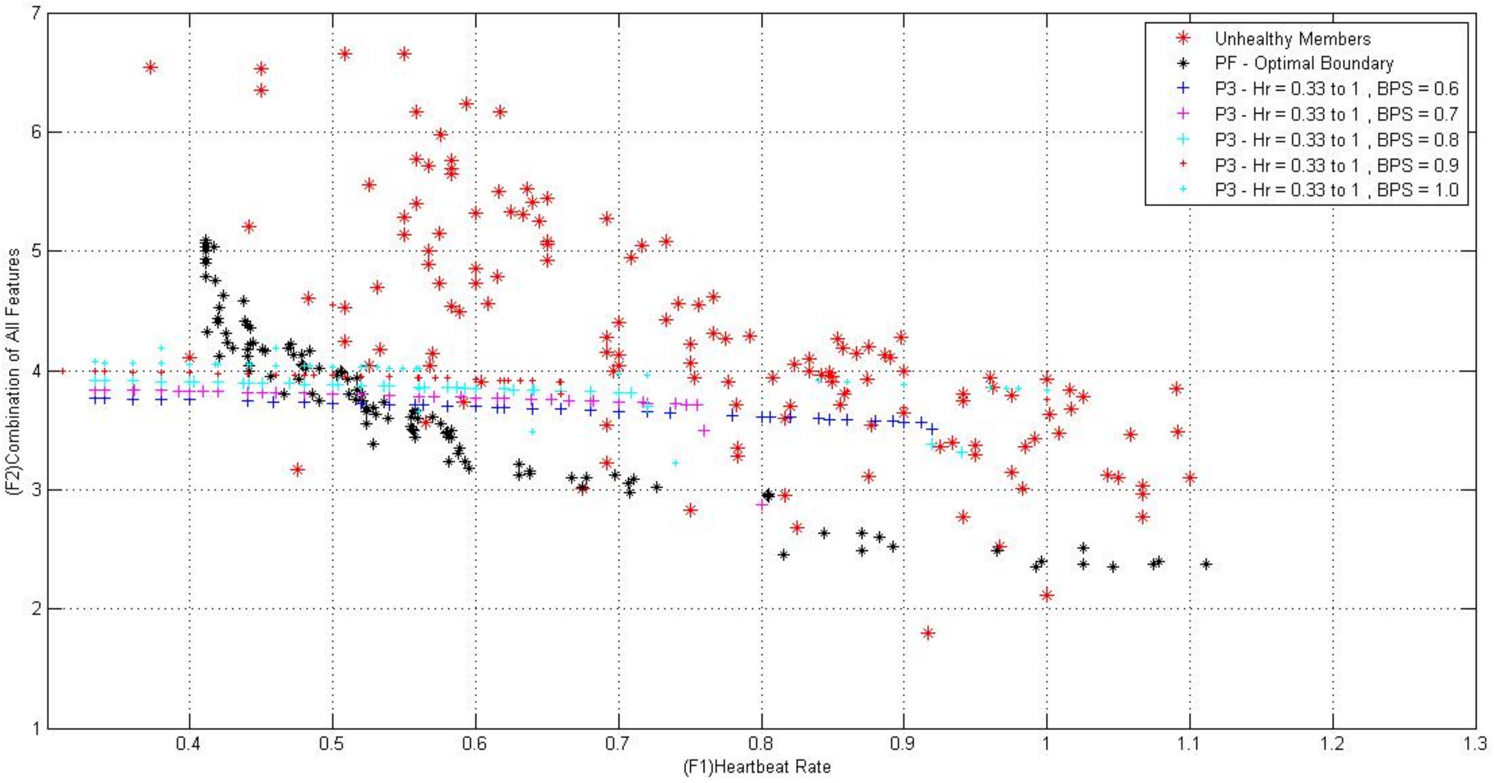


Figure 5.8: Tolerance threshold prediction for P3 using 13 features

Table 5.10: Comparison between the explored points for selected patients

Title	True Positive Rate	False Positive Rate	Precision
6 features Pure Member	0.917	0.083	0.917
6 features predicted points	0.921	0.079	0.921
10 features Pure Member	0.947	0.056	0.947
10 features predicted points	0.947	0.055	0.948
13 features Pure Member	0.953	0.049	0.954
13 features predicted points	0.951	0.051	0.951

5.2 Conclusion

The results of these simulations show that a tolerance threshold for each patient is related to their features. Table 5.10 shows a comparison of these results. However, the explored points are evaluated with SVM. These points have been changed based on the number of features included in each simulation. Indeed, in terms of prediction and real-time health monitoring, the use of feature selection is not allowed. As shown in Table 5.1, the actual tolerance threshold for each patient has been changed based on the number of features which have been included in the simulation. This signifies that the trade-off among all the human features related to heart status can show the exact tolerance thresholds for each person.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

A rapid increase in the world population of elderly people has motivated researchers to improve the healthcare system [54]. Since current healthcare systems are not organized to provide adequate services for an aging population [62], the use of a telemedicine system was proposed in the 19^{th} century [55]. The aim of this service is to reduce the cost and increase the efficient utilization of physician skills, providing remote access to patients for continuous monitoring and real-time analysis of the patients' feedback based on Remote Patient Monitoring (RPM) [54]. To increase the efficiency of these applications, reduce the cost of healthcare and improve the efficiency of the monitoring system [68], RPM has been introduced as a section of telemedicine for better and easier treatment of distant patients [37].

Wearable Health Monitoring Systems (WHMS) is a branch of the RPM applications which takes advantage of wearable biosensors to reduce the cost of health care and improve the efficiency of the monitoring system [68]. For the last two decades, using WHMS has been widely accepted in the global healthcare system.

To address this demand, WBAN has been proposed as a continuous health system to provide real-time feedback[51, 8]. This application is a collection of wireless medical sensor nodes and mobile applications placed around or in a human body to capture, compute and deliver the body signals information to medical centers [54]. The significance of using WBAN can be explored in three different areas [22]:

1. Ease of data collection with a comfortable interface
2. Scalability to support numerous users
3. Real-time monitoring, processing and estimation to improve physician assessments.

In the last decade, a number of WBAN frameworks have been proposed to improve the efficiency of the remote health status monitoring system. However, the earlier devices such as AMAC [33] did not have the flexibility for communication between the patient and physician. In the latest application, Live-Net [62], the device will transfer the data to the correspondent medical center automatically. It is a flexible wearable platform, which has been produced for long-term health monitoring with semi-real-time data analysis. Based on the MIT Wearable Computing Group's distributed mobile system architecture, Live-Net is able to continuously monitor a wide range of physiological signals.

To provide embedded data computing and real-time data analysis, Live-Net takes advantage of the lightweight machine learning algorithms. However using the context engine, a light-weight machine learning algorithm allows systems to classify and identify the variety of user-state contexts, but does not provide a real-time health monitoring system because of the time complexity of the Bayesian Network algorithm which is used in the Live-Net application. While real-time data processing has been proposed to increase WBAN classification's accuracy, using conventional data classification methods such as the Bayesian Network, SVM and decision tree are still highly time-consuming.

To overcome the issue mentioned above, this thesis proposes the use of the Non-Dominated Sorting Genetic Algorithm (NSGA-II) for real-time monitoring systems. The proposed system would be able to classify patients' health status based on their medical records, current activity, blood pressure and heart rate.

Classification and data processing would be satisfied in polynomial time with the order of 3 $O(m(n)^3)$, where (m) represents the number of objectives and (n) is the number of records [3]. While (m) has been reduced from 13 features (Heart-disease Dataset) [4] to 6 using genetic algorithms [16] to speed up the processing time, using standard methods like the decision tree, SVM and bayesian network for data classification still requires polynomial time [68]. Due to this amount of time complexity, the WBAN system has not been used as a prediction system [62]. Applying feature selection speeds up the classification process, but at the same time, the prediction's accuracy would be decreased due to removing some critical features such as age, gender, and current health status. Consequently, since WBAN has been proposed as an autonomous healthcare system for monitoring aging people as well as remote patients, a decrease in the WBAN detection results with the removal some features such as age and gender will have an indirect effect on the final health phenotype [51].

In this work, a new approach based on NSGA-II has been presented for providing real-time heart disease prediction. The Self-Coach application, which is proposed in this research, provides a real-time heart disease prediction and monitoring system using, the specific configuration for each individual patient. Unlike existing classification methods based on NSGA-II, in this study an estimation of risk ratio for each phenotype has been presented; i.e., the likelihood of closeness of a patient, which is categorized in a specific class, to the border of other class (unhealthy PF curve). Furthermore, using this method it is possible to provide initial simulations for patients to find the healthy boundary for their features.

In this thesis, the Cleveland dataset (heart disease) [4] is used. This database contains 75 attributes; 13 features out of 75, which have shown the most correlation with the heart disease problem, are used. To increase the accuracy of the simulation results and provide a real-time prediction function, patients' features are separated into two categories; dynamic and static attributes. The features that do not change suddenly such as age, sex, weight and other medical records are the static features, and some features such as heart rate and blood pressure, that have very flexible curves due to the patient's health status, are categorized as the dynamic features.

The results of these simulations are presented with a (2-D) graph, where the X-axis represents the (Hr) and the Y-axis denotes the trade-off between these dynamic and static features.

These features' taxonomies are used to simulate the new phenotypes (position) for patients regarding their new heart rate and possibly increased blood pressure. Using this method, a set of potential tolerance threshold has been predicted for each patient. The results of this research proved our objective to have a real-time prediction in linear time.

6.2 Future works

Self-Coach will be implemented for IOS and Android operating systems. It would connect to existing tracking heartbeat devices to provide a continues monitoring system. The accuracy of this application will be improved utilizing more dynamic features such as skin temperature, breathing passages and sweating. These features have not been analyzed in this research because they are not included in the Cleveland dataset. The accuracy of detection could be higher if there were a specific dataset with precise features, such as analysis the Hr and BPS for resting time as well as daily activity, stress and emotion.

The proposed method is an open classification application which is suitable for a different study, for example, a daily diet program or the severity of cancer. Furthermore, this approach can be applied to any study in terms of the pattern recognition such as network security and denial of service detection.

Bibliography

[1] Wireless measurement devices,tele-station units. 2004.

[2] M. Al-Noukari and W. Al-Hussan. Using data mining techniques for predicting future car market demand; dcx case study. In *Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on*, pages 1–5. IEEE, 2008.

[3] M. Anbarasi, E. Anupriya, and N. Iyengar. Enhanced prediction of heart disease with feature subset selection using genetic algorithm. *International Journal of Engineering Science and Technology*, 2(10):5370–5376, 2010.

[4] M. P. V. M. C. L. B. Andras Janosi, William Steinbrun, K. Cleveland Clinic Foundation:Robert Detrano, and M. Lichman. Uci machine learning repository, cleveland, hungary, switzerland, and the va long beach. 1988.

[5] U. Anliker, J. A. Ward, P. Lukowicz, G. Troster, F. Dolveck, M. Baer, F. Keita, E. B. Schenker, F. Catarsi, L. Coluccini, et al. Amon: a wearable multiparameter medical monitoring and alert system. *IEEE Transactions on information technology in Biomedicine*, 8(4):415–427, 2004.

[6] C. Baffaut and J. Delleur. Expert system for calibrating swmm. *Journal of Water Resources Planning and Management*, 115(3):278–298, 1989.

[7] J. Bi. Multi-objective programming in svms. In *Proceedings of the Twentieth International Conference on Machine Learning, ICML*, pages 35–42, 2003.

[8] P. Bonato. Advances in wearable technology and applications in physical medicine and rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, 2(1):1, 2005.

[9] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[10] H. Chen, R. H. Chiang, and V. C. Storey. Business intelligence and analytics: From big data to big impact. *MIS quarterly*, 36(4):1165–1188, 2012.

[11] C. A. C. Coello. A comprehensive survey of evolutionary-based multiobjective optimization techniques. *Knowledge and Information systems*, 1(3):269–308, 1999.

[12] C. A. C. Coello, D. A. Van Veldhuizen, and G. B. Lamont. *Evolutionary algorithms for solving multi-objective problems*, volume 242. Springer, 2002.

[13] B. J. Cohen and A. DePetris. *Medical terminology: an illustrated guide.* Lippincott Williams & Wilkins, 2013.

[14] J. Craig and V. Petterson. Introduction to the practice of telemedicine. *Journal of Telemedicine and Telecare*, 11(1):3–9, 2005.

[15] G. V. Crosby, T. Ghosh, R. Murimi, and C. A. Chin. Wireless body area networks for healthcare: a survey. *International Journal of Ad Hoc, Sensor & Ubiquitous Computing*, 3(3):1, 2012.

[16] K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii. In *International Conference on Parallel Problem Solving From Nature*, pages 849–858. Springer, 2000.

[17] K. Deb and T. Goel. Controlled elitist non-dominated sorting genetic algorithms for better convergence. In *International Conference on Evolutionary Multi-Criterion Optimization*, pages 67–81. Springer, 2001.

[18] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2):182–197, 2002.

[19] B. Emami A and S. Samet. An intelligent telemedicine system based on non-dominated genetic algorithm (its-nsga). volume 24. Newfoundland Electrical and Computer Engineering Conference (NECEC, 2015), 2015.

[20] B. Emami A and S. Samet. An intelligent telemedicine system based on non-dominated genetic algorithm (its-wban). volume 18. Annual Aldrich Multidisciplinary Graduate Research Conference, Memorial University of Newfoundland, NL.Canada, 2016.

[21] B. Emami A and S. Samet. An intelligent wban system for heart disease prediction using non-dominated genetic algorithm (its-wban). E-Health Conference 2016Jun 5 - 8, 2016 - Vancouver, BC, 2016.

[22] T. O. J. V. a. L. O.-M. Esteban J. Pino, Dorothy.Curtis. Wireless patient monitoring in a clinical setting. In M. R. Yuce and J. Khan, editors, *Wireless body area networks: Technology, implementation, and applications*, pages 18–37. Pan Stanford., 2012.

[23] E. S. Ford and W. H. Giles.

[24] A. Gaspar-Cunha. Feature selection using multi-objective evolutionary algorithms: application to cardiac spect diagnosis. In *Advances in Bioinformatics*, pages 85–92. Springer, 2010.

[25] A. Gaspar-Cunha and J. A. Covas. Rpsgaereduced pareto set genetic algorithm: Application to polymer extrusion. In *Metaheuristics for Multiobjective Optimisation*, pages 221–249. Springer, 2004.

[26] I. Graham, D. Atar, K. Borch-Johnsen, G. Boysen, G. Burell, R. Cifkova, J. Dallongeville, G. De Backer, S. Ebrahim, B. Gjelsvik, et al. European guidelines on cardiovascular disease prevention in clinical practice: executive summary. *European heart journal*, 2007.

[27] J. Habetha. The myheart project-fighting cardiovascular diseases by prevention and early diagnosis. In *Conf Proc IEEE Eng Med Biol Soc*, pages 6746–6749. Citeseer, 2006.

[28] C. d. v. Haiying.Zhou, Kun-Mean.Hou and Jian.Li. Real-time cardiac arrhythmias monitoring for pervasive health care. In M. R. Yuce and J. Khan, editors, *Wireless body area networks: Technology, implementation, and applications*, pages 41–73. Pan Stanford., 2012.

[29] T. M. Hamdani, J.-M. Won, A. M. Alimi, and F. Karray. Multi-objective feature selection with nsga ii. In *International Conference on Adaptive and Natural Computing Algorithms*, pages 240–247. Springer, 2007.

[30] M. Harris and J. Habetha. The myheart project: a framework for personal health care applications. In *2007 Computers in Cardiology*, pages 137–140. IEEE, 2007.

[31] H. Ishibuchi, N. Tsukamoto, Y. Hitotsuyanagi, and Y. Nojima. Effectiveness of scalability improvement attempts on the performance of nsga-ii for many-objective problems. In *Proceedings of the 10th annual conference on Genetic and evolutionary computation*, pages 649–656. ACM, 2008.

[32] R. S. Istepanian, E. Jovanov, and Y. Zhang. Guest editorial introduction to the special section on m-health: Beyond seamless mobility and global wireless health-care connectivity. *IEEE Transactions on information technology in biomedicine*, 8(4):405–414, 2004.

[33] J. S. Jacobson, A. Lieblein, A. H. Fierman, E. R. Fishkin, V. E. Hutchinson, L. Rodriguez, D. Serebrisky, M. Chau, and A. Saperstein. Randomized trial of an electronic asthma monitoring system among new york city children. *The American journal of managed care*, 15(11):809–814, 2009.

[34] N. Jozefowiez, F. Glover, and M. Laguna. Multi-objective meta-heuristics for the traveling salesman problem with profits. *Journal of Mathematical Modelling and Algorithms*, 7(2):177–195, 2008.

[35] V. Khatibi and G. A. Montazer. A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment. *Expert Systems with Applications*, 37(12):8536–8542, 2010.

[36] J. Kittilsen. Detecting malicious pdf documents. 2011.

[37] S. Kotsiantis. Supervised machine learning: A review of classification techniques. *Informatica*, 31:249–268, 2007.

[38] A. Lahsasna, R. N. Ainon, R. Zainuddin, and A. Bulgiba. Design of a fuzzy-based decision support system for coronary heart disease diagnosis. *Journal of Medical Systems*, 36(5):3293–3306, 2012.

[39] D. T. Larose. *Discovering knowledge in data: an introduction to data mining.* John Wiley & Sons, 2014.

[40] H. Li and Q. Zhang. Multiobjective optimization problems with complicated pareto sets, moea/d and nsga-ii. *IEEE Transactions on Evolutionary Computation*, 13(2):284–302, 2009.

[41] K. Maalel and W. Huber. Swmm calibration using continuous and multiple event simulation. In *Proc., 3rd Int. Conf. on Urban Storm Drainage*, volume 4, pages 1564–1577. Chalmers Univ., 1984.

[42] S. Mirza. Overview of the u.s. diabetes remote patient monitoring devices market. telemedicine and remote patient monitoring, 2004.

[43] M. Mitchell. *An introduction to genetic algorithms*. MIT press, 1998.

[44] T. M. Mitchell. Machine learning and data mining. *Communications of the ACM*, 42(11):30–36, 1999.

[45] J. Muhlsteff, O. Such, R. Schmidt, M. Perkuhn, H. Reiter, J. Lauter, J. Thijs, G. Musch, and M. Harris. Wearable approach for continuous ecg-and activity patient-monitoring. In *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, volume 1, pages 2184–2187. IEEE, 2004.

[46] T. Murata and H. Ishibuchi. Moga: Multi-objective genetic algorithms. In *Evolutionary Computation, 1995., IEEE International Conference on*, volume 1, page 289. IEEE, 1995.

[47] S. K. Murthy. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data mining and knowledge discovery*, 2(4):345–389, 1998.

[48] L. S. Oliveira, M. Morita, and R. Sabourin. Feature selection for ensembles using the multi-objective optimization approach. In *Multi-Objective Machine Learning*, pages 49–74. Springer, 2006.

[49] S. Panda. Multi-objective pid controller tuning for a facts-based damping stabilizer using non-dominated sorting genetic algorithm-ii. *International Journal of Electrical Power & Energy Systems*, 33(7):1296–1308, 2011.

[50] A. Pantelopoulos and N. Bourbakis. A survey on wearable biosensor systems for health monitoring. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4887–4890. IEEE, 2008.

[51] A. Pantelopoulos and N. G. Bourbakis. A survey on wearable sensor-based systems for health monitoring and prognosis. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(1):1–12, 2010.

[52] A. Petcu and B. Faltings. A Scalable Method for Multiagent Constraint Optimization". 2005.

[53] T. N. Phyu. Survey of classification techniques in data mining. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1, pages 18–20, 2009.

[54] A. Rehman, M. Mustafa, I. Israr, and M. Yaqoob. Survey of wearable sensors with comparative study of noise reduction ecg filters. *Int. J. Com. Net. Tech*, 1(1):61–81, 2013.

[55] E. Rinde, I. Nordrum, and B. J. Nymo. Telemedicine in rural norway. In *World health forum*, volume 14, pages 71–71. World Health Organization, 1993.

[56] L. Rokach and O. Maimon. *Decision Trees*, pages 165–192. Springer US, Boston, MA, 2005.

[57] S. Ryu. History of telemedicine: evolution, context, and transformation. *Healthcare Informatics Research*, 16(1):65–66, 2010.

[58] R. Saadatdoost. *Knowledge discovery for large databases in education institutes.* PhD thesis, Universiti Teknologi Malaysia, Faculty of Computer Science and Information System, 2011.

[59] W. Stadler. Fundamentals of multicriteria optimization. In *Multicriteria Optimization in Engineering and in the Sciences*, pages 1–25. Springer, 1988.

[60] P. Stone and M. Veloso.

[61] B. Suman and P. Kumar. A survey of simulated annealing as a tool for single and multiobjective optimization. *Journal of the operational research society*, 57(10):1143–1160, 2006.

[62] M. Sung, C. Marci, and A. Pentland. Wearable feedback systems for rehabilitation. *Journal of neuroengineering and rehabilitation*, 2(1):1, 2005.

[63] S. Tachakra, X. Wang, R. S. Istepanian, and Y. Song. Mobile e-health: the unwired evolution of telemedicine. *Telemedicine Journal and E-health*, 9(3):247–257, 2003.

[64] C. von Lücken, B. Barán, and C. Brizuela. A survey on multi-objective evolutionary algorithms for many-objective problems. *Computational Optimization and Applications*, 58(3):707–756, 2014.

[65] F. Voznika and L. Viana. Data mining classification. 2007.

[66] P. W. Wilson, R. B. DAgostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, 1998.

[67] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[68] M. R. Yuce and J. Khan. Introduction to wireless body area network. In M. R. Yuce and J. Khan, editors, *Wireless body area networks: Technology, implementation, and applications*, pages 1–17. Pan Stanford., 2012.

[69] M. H. Zangooei, J. Habibi, and R. Alizadehsani. Disease diagnosis with a hybrid method {SVR} using nsga-ii. *Neurocomputing*, 136:14 – 29, 2014.

[70] X. Zhang and F. Xu. Survey of research on big data storage. In *Distributed Computing and Applications to Business, Engineering & Science (DCABES), 2013 12th International Symposium on*, pages 76–80. IEEE, 2013.

Chapter 7

Dataset Information

Appendix-A

Publication Request:

*** This file describes the contents of the heart-disease directory.

This directory contains 4 databases concerning heart disease diagnosis. All attributes are numeric-valued. The data was collected from the four following locations:

- Cleveland Clinic Foundation (cleveland.data)
- Hungarian Institute of Cardiology, Budapest (hungarian.data)
- V.A. Medical Center, Long Beach, CA (long-beach-va.data)
- University Hospital, Zurich, Switzerland (switzerland.data)

Each database has the same instance format. While the databases have 76 raw attributes, only 14 of them are actually used. Thus I've taken the liberty of making

2 copies of each database: one with all the attributes and 1 with the 14 attributes actually used in past experiments.

The authors of the databases have requested:

Any publications resulting from the use of the data include the names of the principal investigator responsible for the data collection at each institution. They would be:

- Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
- University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
- University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
- V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

¿¿¿

- Title: Heart Disease Databases
- Source Information: (a) Creators:

  — 1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

  — 2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.

  — 3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.

  — 4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

(b) Donor: David W. Aha (aha@ics.uci.edu) (714) 856-8779 (c) Date: July, 1988

- Past Usage:

  - Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). *International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology*, *64*,304–310.

    – International Probability Analysis

    – Address: Robert Detrano, M.D.

    Cardiology 111-C

    V.A. Medical Center

    5901 E. 7th Street

    Long Beach, CA 90028

    – Results in percent accuracy: (for 0.5 probability threshold)

    Data Name: CDF CADENZA

    – Hungarian 77 74

    Long beach 79 77

    Swiss 81 81

    – Approximately a 77

    logistic-regression-derived discriminant function

  - David W. Aha & Dennis Kibler

    – Instance-based prediction of heart-disease presence with the

Cleveland database

– NTgrowth: 77.0 % accuracy

– C4: 74.8 % accuracy

- John Gennari

  – Gennari, J. H., Langley, P, & Fisher, D. (1989). Models of incremental concept formation. *Artificial Intelligence, 40*, 11–61.

  – Results:

  – The CLASSIT conceptual clustering system achieved a 78.9% accuracy on the Cleveland database.

- Relevant Information:

  This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The ”goal” field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

  The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

  One file has been ”processed”, that one containing the Cleveland database. All four unprocessed files also exist in this directory.

- Number of Instances:

  Database: # of instances:

  Cleveland: 303

  Hungarian: 294

  Switzerland: 123

  Long Beach VA: 200

- Number of Attributes: 76 (including the predicted attribute)

- Attribute Information:

  - Only 14 used
  - 1. #3 (age)
  - 2. #4 (sex)
  - 3. #9 (cp)
  - 4. #10 (trestbps)
  - 5. #12 (chol)
  - 6. #16 (fbs)
  - 7. #19 (restecg)
  - 8. #32 (thalach)
  - 9. #38 (exang)
  - 10. #40 (oldpeak)

- 11. #41 (slope)
- 12. #44 (ca)
- 13. #51 (thal)
- 14. #58 (num) (the predicted attribute)

- Complete attribute documentation:

1. id: patient identification number
2. ccf: social security number (I replaced this with a dummy value of 0)
3. age: age in years
4. sex: sex (1 = male; 0 = female)
5. painloc: chest pain location (1 = substernal; 0 = otherwise)
6. painexer (1 = provoked by exertion; 0 = otherwise)
7. relrest (1 = relieved after rest; 0 = otherwise)
8. pncaden (sum of 5, 6, and 7)
9. cp: chest pain type
   - Value 1: typical angina
   - Value 2: atypical angina
   - Value 3: non-anginal pain
   - Value 4: asymptomatic

10. trestbps: resting blood pressure (in mm Hg on admission to the hospital)

11. htn

12. chol: serum cholestoral in mg/dl

13. smoke: I believe this is 1 = yes; 0 = no (is or is not a smoker)

14. cigs (cigarettes per day)

15. years (number of years as a smoker)

16. fbs: (fasting blood sugar ¿ 120 mg/dl) (1 = true; 0 = false)

17. dm (1 = history of diabetes; 0 = no such history)

18. famhist: family history of coronary artery disease (1 = yes; 0 = no)

19. restecg: resting electrocardiographic results

    – Value 0: normal

    – Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of ¿ 0.05 mV)

    – Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

20. ekgmo (month of exercise ECG reading)

21. ekgday(day of exercise ECG reading)

22. ekgyr (year of exercise ECG reading)

23. dig (digitalis used furing exercise ECG: 1 = yes; 0 = no)

24. prop (Beta blocker used during exercise ECG: 1 = yes; 0 = no)

25. nitr (nitrates used during exercise ECG: 1 = yes; 0 = no)

26. pro (calcium channel blocker used during exercise ECG: 1 = yes; 0 = no)

27. diuretic (diuretic used used during exercise ECG: 1 = yes; 0 = no)

28. proto: exercise protocol

    1 = Bruce

    2 = Kottus

    3 = McHenry

    4 = fast Balke

    5 = Balke

    6 = Noughton

    7 = bike 150 kpa min/min (Not sure if ”kpa min/min” is what was written!)

    8 = bike 125 kpa min/min

    9 = bike 100 kpa min/min

    10 = bike 75 kpa min/min

    11 = bike 50 kpa min/min

    12 = arm ergometer

29. thaldur: duration of exercise test in minutes

30. thaltime: time when ST measure depression was noted

31. met: mets achieved

32. thalach: maximum heart rate achieved

33. thalrest: resting heart rate

34. tpeakbps: peak exercise blood pressure (first of 2 parts)

35. tpeakbpd: peak exercise blood pressure (second of 2 parts)

36. dummy

37. trestbpd: resting blood pressure

38. exang: exercise induced angina (1 = yes; 0 = no)

39. xhypo: (1 = yes; 0 = no)

40. oldpeak = ST depression induced by exercise relative to rest

41. slope: the slope of the peak exercise ST segment

    – Value 1: upsloping

    – Value 2: flat

    – Value 3: downsloping

42. rldv5: height at rest

43. rldv5e: height at peak exercise

44. ca: number of major vessels (0-3) colored by flourosopy

45. restckm: irrelevant

46. exerckm: irrelevant

47. restef: rest raidonuclid (sp?) ejection fraction

48. restwm: rest wall (sp?) motion abnormality

    0 = none

    1 = mild or moderate

    2 = moderate or severe

3 = akinesis or dyskmem (sp?)

49. exeref: exercise radinalid (sp?) ejection fraction
50. exerwm: exercise wall (sp?) motion
51. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
52. thalsev: not used
53. thalpul: not used
54. earlobe: not used
55. cmo: month of cardiac cath (sp?) (perhaps ”call”)
56. cday: day of cardiac cath (sp?)
57. cyr: year of cardiac cath (sp?)
58. num: diagnosis of heart disease (angiographic disease status)
    - Value 0: ¡ 50% diameter narrowing
    - Value 1: ¿ 50% diameter narrowing

    (in any major vessel: attributes 59 through 68 are vessels)

59. lmt
60. ladprox
61. laddist
62. diag
63. cxmain

64. ramus
65. om1
66. om2
67. rcaprox
68. rcadist
69. lvx1: not used
70. lvx2: not used
71. lvx3: not used
72. lvx4: not used
73. lvf: not used
74. cathef: not used
75. junk: not used
76. name: last name of patient

- Missing Attribute Values: Several. Distinguished with value -9.0.

- Class Distribution:

  Database: 0 1 2 3 4 Total

  Cleveland: 164 55 36 35 13 303

  Hungarian: 188 37 26 28 15 294

  Switzerland: 8 48 32 30 5 123

  Long Beach VA: 51 56 41 42 10 200