

© 2017 Qian Cheng

PREDICTIVE MODELING OF HEALTH STATUS USING MOTION
ANALYSIS FROM MOBILE PHONES

BY

QIAN CHENG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Professor Bruce Schatz, Chair
Professor Jiawei Han
Associate Professor Paris Smaragdis
Dr. Christian Konig, Microsoft Research

ABSTRACT

It is unknown what physiological functions can be monitored at clinical quality with a normal smartphone, which is ubiquitous. There are standard measures like walk speed, pulmonary function and oxygen saturation variation for health status of cardiopulmonary patients. The dissertation is to summarize my studies of using sensor data from regular smartphones to accurately measure walking patterns, in order to monitoring health status for cardiopulmonary patients.

Fifty five pulmonary patients were participated in the study. The sensor data for their walk test and free walk are collected and stored by a novel designed Android smartphone application. Different machine learning techniques are applied and compared to predict gait speed, pulmonary function and oxygen saturation.

The result shows that walking patterns are highly correlated with health status. The trained models can predict health status accurately for each patient. Initial testing indicates the same high accuracy as with active monitors, for patients in hospitals during walk tests. The ultimate goal is that patients can simply carry their phones during everyday living, while models support automatic prediction of pulmonary function for health monitoring.

To My Parents
Thank You For Your Love and Support.
To Me
You Deserve It.

ACKNOWLEDGMENTS

Dr. Bruce Schatz as my primary supervisor helps me a lot for the accomplishment. Dr. Jiawei Han, Dr. Paris Smaragdis and Dr. Christian Konig serve as the Ph.D. committee and offer extra help. Dr. Chengxiang Zhai, Dr. Saurabh Sinha and other faculties and staffs in University of Illinois help me with academic and logistic issues.

At the University of Illinois at Urbana-Champaign, Department of Computer Science and the Institute for Genomic Biology provided facilities for software development and data analysis.

At Carle Foundation Hospital in Urbana Illinois, the Biomedical Research Institute helped test the patients. Dustie Mitchell, RN, provided medical supervision during patient testing. Debbie Brusveen, RN, provided access to the Pulmonary Rehabilitation facility, where the tests were run. Sujay Bangarulingam, MD, was the pulmonary physician responsible for IRB 497221.

At NorthShore University HealthSystem, the Department of Respiratory Therapy at Evanston Hospital carried out patient testing and phone recording. All clinical experiments were protected by NorthShore IRB EH15-025 entitled Mobile phone software MoveSense versus traditional exercise testing to assess health status, including extensions for passive monitors. We acknowledge Brian Edwards and Catherine Zhu for managing and anonymizing phone data and medical records using REDCap archiving system.

TABLE OF CONTENTS

LIST OF TABLES	vii
LIST OF FIGURES	ix
CHAPTER 1 INTRODUCTION	1
CHAPTER 2 MOTIVATION	4
2.1 Chronic Disease	4
2.2 Smartphones as Ultimate Solution	5
2.3 Best Intersection: Chronic Cardio-pulmonary Disease	6
CHAPTER 3 BACKGROUND	7
3.1 Gait Analysis	7
3.2 Wearable Devices and Sensors	8
3.3 Artificial Intelligence for Medical World	9
CHAPTER 4 RELATED WORK	10
4.1 Gait Parameters from Trunk Acceleration	10
4.2 Smartphone Application for Tracking Arthritis	11
4.3 Activity Recognition Studies	11
4.4 Gait and COPD: Daily Monitoring by Patient Input	12
CHAPTER 5 PROBLEMS AND GOALS	13
5.1 Defining the Problem	13
5.2 Blue Print of the Free-living Health Monitor	14
CHAPTER 6 EXPERIMENTS DESIGN	19
6.1 Technical Foundations	20
6.2 Medical Applications	21
CHAPTER 7 SUBJECT AND DATA COLLECTION	25
7.1 Pilot: Collaboration with University of Illinois Hospital & Health Sciences System	25
7.2 Collaboration with Carle Foundation Hospital	26
7.3 Collaboration with NorthShore University HealthSystem	26
7.4 Dataset in Main Study	27

CHAPTER 8	TECHNICAL APPROACHES	29
8.1	System Design	29
8.2	Gait Parameters	32
8.3	Learning Techniques	35
8.4	DPClass: Discriminative Patterns-based Classification	40
8.5	Model Evaluation	44
CHAPTER 9	PREDICTION OF PULMONARY FUNCTION	45
9.1	Pilot Study I: Severity Can Be detected for Patient in the Same Demographic Cohort	45
9.2	Pilot Study II: An Accurate Universal Model Fits All De- mographic Cohorts	47
9.3	Results	51
9.4	Main Study: Predicting Health Status for Healthy and COPD Subjects	52
CHAPTER 10	PREDICTION OF OXYGEN SATURATION	66
10.1	Pilot: Predicting Raw Pulse Oximetry Values	66
10.2	Predicting Transitions in Oxygen Saturation	72
CHAPTER 11	GAIT SPEED PREDICTION	81
11.1	Predicting Speed for Six Minute Walk Test	82
11.2	Healthy Controls with Fitness Devices	84
11.3	Rethink the Correlation between Speed and Status	85
CHAPTER 12	LIMITATIONS	88
12.1	Device and Sampling	88
12.2	Good Walking Detection	89
12.3	Feature Extraction	90
12.4	Model Training	91
12.5	Medical Applications	91
CHAPTER 13	POTENTIAL EXTENSIONS	92
13.1	Data Quality Improvement	92
13.2	Length of Good Walking for Health Status Prediction	93
13.3	Good Walking Detection	94
13.4	Passive Monitoring of Human Health From Daily Activities	99
CHAPTER 14	LESSONS LEARNED	102
14.1	Resolution	102
14.2	Sensitivity	104
14.3	Equivalence	105
14.4	Deviance	105
REFERENCES		107

LIST OF TABLES

5.1	Global Initiative for Chronic Obstructive Lung Disease (GOLD) Definition. There are four GOLD stages based on FEV1% predicted.	17
6.1	Medical Applications of Motion Data Analysis.	21
6.2	Planned experiments on the combination of technical found- ations and medical applications.	23
6.3	Planned experiments on the combination of technical found- ations and medical applications.	24
9.1	Demographic Information of Female Chronic Obstructive Pulmonary Disease Patients	46
9.2	Support Vector Machine Classification Results with Demo- graphic Cohorts	46
9.3	Demographic information of each group of patients by GOLD levels. Age, height and weight are in average (minimum - maximum) format. These demographics plus gender are used to adjust the model.	47
9.4	Self-validation results for RBFSVM model and linear SVM model. We compare the accuracy of classification, which is all correct classified samples divided by the total number of samples. This is by ten-second interval, not by patient. . . .	51
9.5	Demographic information of each group by GOLD levels. Patient [Female] shows the number of patient [the number of female patient] in each group. Age, height and weight are in median [minimum - maximum] format.	53
9.6	Model Validation.	56
9.7	Proportion of Rules Containing Each Feature. We count the number of appearance of a specific feature in all top 30 rules.	58
9.8	Discriminative patterns in DPClass model.	60

10.1	Validation result of feature selection approaches in absolute error rate. The average r shows that Support vector regression obtains a lower average r than Gaussian process regression. In 10-fold cross validation, selecting 50 features obtains a lower average r than selecting 20 features. In cross validation over subjects, selecting 50 features obtains a higher average r than selecting 20 features.	68
10.2	Validation result of component extraction approaches in absolute error rate. The result shows that Gaussian process regression obtains lower average r than support vector regression. In 10-fold cross validation, components extracted from magnitude values obtains lower average r than components extracted from three-dimensional values. In cross validation over subjects, components extracted from three-dimensional values obtains lower average r than components extracted from magnitude values.	68
10.3	Validation of regression models on cohorts. The result indicates that training on cohorts obtains a lower absolute error rate than training universally.	71
10.4	Free-walking evaluation of oxygen saturation prediction models. The average error rate of cohort models is lower than the universal model.	72
10.5	Demographic Information of Tested Subjects. The age, height and weight values are in the format of mean (min - max).	75
10.6	Oxygen Saturation Pattern Model Validation. I validate eight different predictive models by 10-fold cross validation. The first four rows represents the FSA results and the last four rows represent the CEA results. Each column represents the prediction accuracy for samples in the corresponding category. The accuracy is calculated by the portion of number of correctly predicted samples in the total number of samples in this category.	79
11.1	Subject Demographics	81
11.2	Error Rates for Speed Prediction for both FSA and CEA input features	83

LIST OF FIGURES

3.1	Gait analysis: motion of human body. Even one 3-D acceleration sensor at lower-back (L3) position will generate plenty of information to describe gait.	8
5.1	The overall design of the free-living health monitor system. . .	15
5.2	Correlation between 6MWD and mortality. The shorter distance a subject walks in a six-minute walk test, the higher mortality this subject suffers.	16
6.1	Overview of the dissertation study procedures.	20
8.1	CPU usage in percentage while sampling frequency increases for three common smartphones	30
8.2	Guarenteed sampling frequency of Devices measuring gait. Human walking is usually in 3-5Hz which means to measure stepping a monitor of 10Hz sampling is enough.	31
8.3	Overview of DPClass.	42
9.1	Sample result of walking detection algorithm. Green lines represent raw acceleration and blue dash lines represent walking status: higher level means this period is detected as walking and lower level means period is detected as stationary.	49
9.2	Evaluation of GOLD classification for labeled 6MWT sessions. The bar-plot shows the portion of predicted GOLD stages for all labeled walking samples of each patient. Except for Patient 7, 9 and 17, all other patients are 100% correctly predicted within every interval. After voting, all patients are predicted as their actual GOLD stages.	52
9.3	Brief Decision Tree Structure. Each Node represents a specific feature dominating decision making on this step. Decision trees making binary decision trees on each node.	59

10.1	Comparison of absolute error rate (r) obtained by 10-fold cross validation. Support vector regression (SVR) versus Gaussian process regression, feature selection approach (FSA) versus component extraction approach (CEA) and magnitude of acceleration versus raw values of acceleration. . .	69
10.2	Bland-Altman plots for oxygen saturation prediction results. Mean= $(pred + obs)/2$; Difference= $pred - obs$. Confidence interval (CI) is $p < 0.05$. (a) Comparison between two regression models: support vector regression (SVR) and Gaussian process regression (GPR). (b) Comparison between two parameter generating approaches: feature selection approach (FSA) and component extraction approach (CEA). (c) Comparison between two acceleration processing strategies: magnitude of acceleration (Mag) and raw value of three-dimensional acceleration (3D)	70
10.3	Bland-Altman plot for free walking evaluation. Mean= $(pred + obs)/2$; Difference= $pred - obs$. Confidence interval (CI) is $p < 0.05$. Most of the outliers are generated by the universal prediction model. Prediction model of Cohort I does not generate any outliers.	73
10.4	Mean oxygen saturation (SpO ₂) for GOLD1, GOLD1T and GOLD2 in the 6MWT	76
10.5	Mean and standard deviation (SD) of oxygen saturation for each subject.	77
11.1	Error in Distance Estimation for 6MWT for the Full and Cross Validation Models in Healthy and Pulmonary Patients. .	84
11.2	Box Plot of Actual versus Predicted Distance for All Devices.	85
11.3	Comparison of mobility prediction between MoveSense and popular fitness devices.	86
11.4	Figure. Bland-Altman plot of 6MWD comparison. Blue dots represent 6MWDs GOLD 2 patient sessions. Red squares represent 6MWDs of GOLD 3 patient sessions. GOLD severity levels are not well separated by 6MWD. . . .	87
13.1	Classification Accuracy with Different Walk Length in Training. The accuracy in ten-fold cross validation reaches 100% after two-minute (120-second) length.	94
13.2	Overall Framework of GaitTrack	97
13.3	Comparison of traditional COPD diagnosis approach and proposed approach. Passive monitor will enhance case finding capacity.	101
14.1	Main factors affect model quality.	103

CHAPTER 1

INTRODUCTION

A revolution in health monitoring is coming, due to widespread mobile devices. Smart health techniques provide new solutions to hardcore clinic care problems. Maintaining smart health requires predicting situation that deviate from the norm, to detect significant health status differences. Today, mobile devices are widely spread, including cellphones, music players.

Mobile phones are nearly ubiquitous, with the investigation of Pew Research Center showing above 90% ownership as of January 2015, and keep increasing. More inspiring, the ownership of smartphones increased to 68% at fall of 2015 from 35% in spring of 2011[1][2]. At the same time, market of innovative wearable devices, such as fitness bands, smart watches and other health care related gadgets, is expanding rapidly.

Determining abnormal status requires information input from a particular individual and the particular time/place. The “normal” of health varies over time, across populations with differing lifestyles[3]. Certain behaviors increase normal, such as proper diet and exercise, while other behaviors decrease normal, such as unusual stress and strain. Health determinants of “abnormal” are relative, rather than absolute. The status of “health” needs to be continuously measured and deviation from the norm detected as a statistical outlier[4].

As work with Dr. Bruce Schatz for my Ph.D. study, my fellows and I dedicate ourselves in exploring a sufficient approach to monitor human health with simple smartphones. For populational study, individual cost is always one of a high priority factors that should be concerned. If the ultimate goal is to perform a function on millions or even billions of people, utilizing a ubiquitous device, like a smartphone, that most of the target population suppose to own is the only solution. We demonstrate the ability of smart phones to operate as medical quality health monitors by making the following contributions. We confirm the quality of smart phone hardware by comparing the

sensor chips of smartphones to chips in validated medical devices and health fitness devices. We design two versions of sensor pipelines, GaitTrack and MoveSense, to overcome smartphone firmware limitations. We confirm the output of our software running on three smartphones is compatible to output of medical devices by directly analyzing the waveforms. We collect walking speed data from pulmonary patients running 6MWTs and free-walking tests. We train different machine learning models to predict natural walking speed. We validate the speed prediction within 6MWT sessions, the standard medical measurement of pulmonary patients, by computing the predicted distance to the real six-minute walk distance (6MWD) in the patients medical records.

After confirming the smartphones is a sufficient device to measure gait, the problem is how to apply gait analysis knowledge to indicate health status. Chronic Obstructive Pulmonary Disease (COPD) is selected as the target chronic disease. COPD is a major cardiopulmonary disease affecting tens of millions of patients. It is usually measured by pulmonary function tests (PFTs)[5]. Pulmonary function is measured with a spirometer in clinical condition. The patient breathes into the spirometer and the airflow is measured for such action. The forced expiratory volume in 1 second (FEV1) has been calibrated by forced vital capacity (FVC) and normalized by subject's demographic information to provide a standard measures function, called predicted FEV1%. The Global Initiative for Chronic Obstructive Lung Disease (GOLD) defines four different stages (GOLD 1 to GOLD 4) representing mild, moderate, severe and more severe status of COPD. In motion analysis, the more severe COPD (GOLD 4) is eliminated from the scope of study due to their lack of mobility. In addition, healthy subjects are defined as "GOLD 0" and introduced in the evaluating system.

The dissertation demonstrates an Android smartphone with its integrated sensors could be turned to work as a health monitor for COPD patients. It contains three major types of predictive models: predicting gait speed, predicting pulmonary functions and predicting oxygen saturation. Predicting gait speed is an intermediate step for predicting health status[6][7][8]. Also it provides a gold standard to evaluating the technique, if the predictive model cannot predict speed accurately, it is highly possible that it cannot predicting anything. Predicting pulmonary function is actually the major task for predicting health. Here we regard the GOLD stages as our gold standard for classifying different health status levels. Predicting oxygen saturation is

a step beyond predicting current status, but predicting “risk”. Continuous SpO₂ values are recorded for the full 6MWT, which is a novel approach for COPD diagnosis. The observation shows clearly there is a transition group between standard GOLD 1 and GOLD 2, which can be distinguished from stable GOLD 1 and GOLD 2 status by the patient’s oxygen saturation pattern during the 6MWT.

CHAPTER 2

MOTIVATION

“The very first step to change the world is to choose which part of the world you are willing, and able to change.”

My Ph.D. study started from the question: If I am interested in the medical and healthcare area, what could be the most valuable problem I can try to solve with the knowledge I have learned, and will be learning in the nearer future? The answer is: chronic disease.

2.1 Chronic Disease

Personally, I am interested in topics about chronic diseases because of my family. Both my patients have the major chronic diseases, and so do lot of my relatives and friends. I can imagine that their health status will get worse as their age increase, but currently we cannot do anything about it but just follow the general medicine treatments, controlling the status from getting worse too quickly, in a way that may not be sufficient to them.

In the scope of large population, there are chronic diseases related to all major parts of human body, such as cancer, diabetes, heart and lung problems. Every second people die because of their mortality and body functions are totally destroyed by chronic disease, after fighting to survive for years or even decades. People want to fight with chronic disease in a better way, as the technique improvements provide possibilities to realize it.

In acute care context, patient will notice when they need to go to the hospital. If you get cough, fever, headache, etc. and cannot bear it, just go to the hospital. If you are injured and bleeding, you go to the hospital. However, in chronic context, it is hard to decide when to go to the hospital for screening and examination. Always it is too late for early treatment when the chronic status is confirmed by diagnosis. Even when a patient is diagnosed, it

is still too hard to track the status change and adjust the treatment in time. Rehospitalization frequency is based on the intuition of the physician and natural feeling of the patient, neither of which are sensitive enough. Thus, for better status tracking, the patients need to know more about their status, not from the clinical diagnosis, but from the information of their body.

Of course hospital provides fancy devices, which could technically track all detail information the physician needs to know to inform whether your status is changing or not. However, in real world it is impossible to apply this procedure in the whole population of chronic diseases. The hospitals don't have enough resources, and the patients are not rich enough. Thus, a possible solution dealing with large population can only be applying ubiquitous hardware that everyone has already owned and machine decision making approaches requires least human interrupt.

2.2 Smartphones as Ultimate Solution

As I mentioned before, smartphones are now widely spread among the whole population. The ownership of smartphones in the United States is 68% in 2015 and it will climb quickly as expected in the following years[1].

To support the basic functionality, all smartphones have a set of motion tracking sensors: accelerometer, gyroscope, GPS, etc. It means no matter how cheap it is, if it is a Android or iOS smartphone, it must have these sensors. Currently high-end smartphones also have co-processor specifically for continuous motion tracking.

The smartphone itself is a powerful computer, which makes running machine learning models on the phone to make prediction possible.

Also, currently almost all smartphones share two major operating systems, iOS and Android. It provides an easy what to extend the tracking software with more functions in a standard way, and all functions are suppose to run on every device running the same operating system without any gaps.

As a phone, the device must be able to make phone calls and support data roaming. It remind us that a phone is always a device for communication, even when it is used to track health. It is a plus when compare to other motion trackers. You can always imagine the situation that whenever the phone detects problems, it will try to communicate with the patient to con-

firm, and it can help the patient to reach out the physician intelligently in all conditions.

Another important reason to use smartphones as the solution is: it is very cheap, when compare to the medical related cost. A top high-end smartphone cost less than \$1,000, and the cheapest option could cost as low as \$30 and everyone can buy one from any local Walmart stores.

2.3 Best Intersection: Chronic Cardio-pulmonary Disease

Chronic Cardio-pulmonary disease is a term describing all heart and lung function related chronic disease, including Chronic Obstructive Pulmonary Disease (COPD)[9] and Congestive Heart Failure (CHF)[10] are the major causes of hospitalization for senior patients, and thus the major costs for Medicare[11]. Such chronic diseases cause shortness of breath, which limits walking speed. Mental limitations, such as sadness caused by stress, also limit gait speed[12], similar to those from physical limitation. Gait analysis has great potential to evaluate severity of anxiety/depression as well as COPD/CHF[13]. Recovery from physical injury such as trauma or arthritis[14] can also be detected by walking patterns.

Gait Speed is the individual's walking pattern; any abnormal gait is then a deviation from normal walking[15]. Harrison's Principles of Internal Medicine states "Watching a patient walk is the most important part of the neurological examination" [16]. Normal gait requires that many systems, including strength, sensation and coordination, function in an integrated fashion. Many severe problems in the nervous system and musculoskeletal system will show in the way a person walks, such as cerebral palsy and stroke paralysis[17]. Gait speed is recognized as a significant vital sign for monitoring human health[6][7].

In the dissertation I demonstrated how our lab build a software that will be able to make a general smartphone into a device that can perform gait analysis. Based on the gait information collected from phones, it is possible to predict health status for chronic cardio-pulmonary diseases.

CHAPTER 3

BACKGROUND

The dissertation study is applying signal processing and machine learning techniques to solve practical problems in bio-medical and healthcare domain. Beforehand, it is necessary to briefly introduce the background knowledge of techniques and the application.

3.1 Gait Analysis

In Kinesiology, gait analysis is a term representing analyzing the mechanism of human motion. The Whittle book: *Gait Analysis: An Introduction*, describes in detail how we can measure gait, the motion pattern of human being[15]. The book starts from the mechanism of human move, talking about the skeleton and muscle interaction while people walk. It also introduces a series of parameters which can potentially indicates the gait status, such as gait speed, cadence, *etc.* Besides, it also points out that with capturing the acceleration signals during human walking, there are even more indicators can be retrieved out for gait. The most common method is to put an accelerometer on the lower-back (L3) position, which is the most stable position on human body when he is moving, show in Figure 3.1. Even one single accelerometer put on L3 could generate plenty of information for human health[18][19][20].

Gait analysis is a major area in kinesiology and it helps our understanding of human motion in lots of ways, improving sports training, rescue training and other areas that for healthy people. However, is gait analysis ready for clinical use? Studenski briefly summarizes that it would be possible [7]. Even simple gait speed will unveil lots of information about human health.

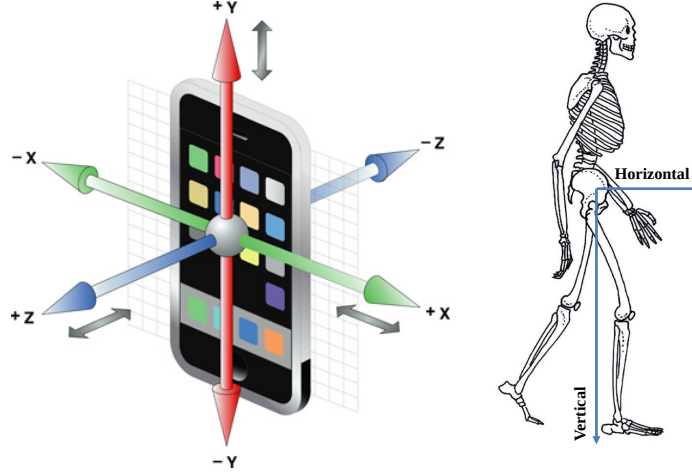


Figure 3.1: Gait analysis: motion of human body. Even one 3-D acceleration sensor at lower-back (L3) position will generate plenty of information to describe gait.

3.2 Wearable Devices and Sensors

Most sensors studies require specialized devices to measure gait for personal diagnosis[21]. Such devices are placed on the person's body and measure motion, in order to assess health. For chronic lung disease, a medical accelerometer is strapped to the lower back (L3) position[22]. The device is used to evaluate severity of COPD[23]·[24] and CHF[25]. Despite how helpful the methodology is, the expense and inconvenience raise the threshold and limit the number of patients who can be monitored.

Years ago, it might be thought that phone sensors would not be adequate for medical tasks and phones are used as hubs for transmitting data[26]·[27]. However, the underlying physical MEMS accelerometer chips[28] are often the same or have very similar operating characteristics in phones and in medical devices[29].

First and foremost, the difference between phones and dedicated medical devices lies in the firmware of the device itself. While medical devices are designed to solely read the accelerometer sensor at a high frequency signed to solely read the accelerometer sensor at a high frequency and record the results, phones are designed to conduct a myriad of tasks including reading sensors, taking phone calls, playing games, etc. and must handle multitasking each task in a fair manner. Thus, medical accelerometers excel at talking raw data at high frequencies but lack the capability to do computational tasks

on the device while phones must take readings at lower frequencies to handle the multitasking. However, compared to any type of medical accelerometers, phones, especially smartphones, have much higher computational power to execute sensor data processing, which makes real-time gait analysis possible. Recently, major phone manufactures published their high-end smartphones, such as Samsung Galaxy S6 and iPhone 6S, with a standalone process specifically handling continuously monitoring. Moreover, these high-end smartphones also have software kits to support converting the continuously collected sensor data into health and fitness indicators such as step counts, sleeping time, etc.. From the perspective of developers, raw sensor data is also available to execute advance analysis as demand. Thus, as the improvement of phone hardware, there will be less restriction to use phones to replace medical accelerometers.

3.3 Artificial Intelligence for Medical World

It has been decades since electronic medical recording (EMR) system was utilized in health systems. After more and more patient data being stored in the database, more complicated data driven analysis is required to retrieve insights. As computing power improves, artificial intelligence techniques obtain capabilities to overcome hardcore challenges in lots of areas.

In medical domain, lots of machine learning techniques has been applied on analyzing the EMRs and assisting medical diagnosis. The trends will keep and develop even faster in the next years. It is a good time to start seeking for possible health monitoring solutions with machine learning techniques. In 2012, I have a co-author publication on AMIA which talks about classifiers of medical terms that can help better classify online posts and help people to find useful information easier[30].

Personalized medicine leverages the need of artificial intelligence techniques. Health monitoring system needs to read and learn from the patient's medical history so that it can be better adjusted and achieve higher monitoring reliability.

CHAPTER 4

RELATED WORK

Monitoring human health with wearable devices, including smartphones, is being studied for years. The original concept came from gait analysis, a standard research area in Kinesiology. Later on, as the development of hardware and computing power, more and more electronic technologies are applied in this area, such as signal processing, information retrieval, image processing and data mining.

4.1 Gait Parameters from Trunk Acceleration

Zijstra *et al.* first published a series of research outcomes on the feasibility of analysis of spatio-temporal gait parameters based upon accelerometry. They recruited healthy subjects walking on treadmills, putting an accelerometer on the L3 position of their body. They demonstrated that the spatio-temporal gait parameters they compute could be correlated with existing gait features such as gait speed. They also provide a bunch of ways to calculate different gait parameters.

In this study, they use medical accelerometers which usually cost thousands of dollars and the subjects are all wearing regular shoes and cloths. It demonstrates there are proper ways to measure gait with accelerometry in casual situation, which make it possible to track human motion in everyday life without attaching special sensors on the body of the subject.

Yang *et al.* study the feasibility of using a smartphone with an embedded accelerometer in gait pattern monitoring. They performed sophisticated experiments to assess the quality of such smartphone based gait pattern monitoring. On one hand, they proposed over thirty potential gait parameters, combining with the domain knowledge in signal processing. On the other hand, they sampled acceleration data in the frequency range from 100 hz to

as low as 4 hz to compare the quality of major gait parameters and drew the performance chart.

4.2 Smartphone Application for Tracking Arthritis

A research team at Kyoto University, Japan, published a series of study results which uses smartphone software to track arthritis [31] [14] [32]. This series of publications talk about technical details about how to implement the phone application to collect sensor data and compute spatio-temporal gait parameters to describe human motion, and real medical application on how to use the gait parameters to measure status of arthritis.

In their study, the computed peak frequency (PF), autocorrelation peak (AC), and coefficient of variance (CV) of each acceleration peak intervals. They first compare the gait parameters between the positive group and negative group. The statistical results show that the different arthritis status group could be distinguished by gait features. Then based on these outcomes, the created a system to measure arthritis levels by gait, which can be computed with a simple smartphone application.

4.3 Activity Recognition Studies

Machine learning techniques are applied in recognizing human activities. However, most of the accurate activity recognition techniques are based on imaging or multi-sensing. Simply using one accelerometer to collect enough information to train an activity recognition model is still challenging. The current situation is it has been well-studied but poorly applied.

Lu *et al.* presented a high accuracy activity classification algorithm based on Nokia phone sensors. The classifier is able to distinguish ten different activities in four categories, including stationary, walking, uneven moves and other heavier moves. The classifier was trained only by on persons data so it is certain that it won't be able to apply on a population level.

WISDM Laboratory at Fordham University initiated their activity recognition study with collecting student's motion data [18]. They recruited several university students to perform the required activities and label each samples.

Then use supervised-learning methods to train the classifier. The validation results shows a specific known activity could be recognized from the pool of known activities, such as walking can be detected from a sample set only contains walk, run, stay data, which all contribute to model training.

Park *et al.* study on activity recognition with similar ideas [33]. This study applies support vector machine and the trained model can recognize the desire activities very accurately. Besides, the study tries to recognize the position of the accelerometer being put on human body, and predict the speed level when human walking on a treadmill.

4.4 Gait and COPD: Daily Monitoring by Patient Input

Physical activity level is an important clinical marker of disease status in COPD but the correlation between them are still not clear. Moy *et al.* operated a population level study to initially unveil the relationship between physical activity level and COPD severity. This study is to explore the daily step counts, a directly measured physical activity on 171 people with stable COPD, with a ankle-worn step counter named StepWatch Activity Monitor. Patients need to self-report their daily step counting values on a web-portal. The collected data are used to compare with the patient's performance of 6MWT in clinical condition.

This study is one of the very first projects that aims to make daily health monitor for COPD, which applies the one dimensional gait feature input, the step counts, to obtain status variation.

CHAPTER 5

PROBLEMS AND GOALS

An essential problem to implement a health monitor is to interpret the sensor data so the information collected from phones can be correlated to health status. The goal of my Ph.D. research contribution is to utilize the sensor data collected from phones and apply regression and classification techniques to build predictive models for three medical applications which have diagnosis power for measuring COPD severity: 1) gait speed and six-minute walk distance, 2) pulmonary function tests and GOLD stages and 3) SpO_2 and oxygen saturation pattern.

5.1 Defining the Problem

Monitoring health is an interdisciplinary topic, requiring an effective combination of knowledge from different domains. This research project originally starts at 2011, along with the booming spread of smartphones and beyond the age of widely spread of popular fitness devices. Other research groups has made contributions to this area. Motion tracking analysis has been improved a lot recently. Back to that time, artificial intelligent is able to handle problems such as recognizing different activity motions, recognizing different device positions on human body and detecting abnormality of gait. Such techniques has been applied on general fitness cases like if the person is health or sick based on the variation of his/her gait. Meanwhile, some groups has been working on monitoring special diseases, such as arthritis, with motion sensing techniques. They choose arthritis because it is a physical disease which obviously will change the patient's gait. The measurement will capture the differences and obtain a perfect reaction on arthritis status. However, the medical value of such type of work is quite limited, simply because it duplicates the observation and does not require any machine learning

and information retrieval step to return a “perfect” decision.

As described in Chapter 2, (COPD) Chronic Obstructive Pulmonary Disease is selected as the target disease, after considering seriously about the medical value. COPD is a pulmonary disease related to patient’s lung function. It means that the causality of this disease is not directly related to motion, like bone and joint problem. The standard test for diagnosis of COPD is pulmonary function tests, also not related to body motion. However, kenisiological study consistantly shows that COPD will change human gait. Six-minute walk test (6MWT) is a documented test for COPD diagnosis even though it is not a necessary step for the lab test. At the same time, experienced pulmonary nurses, who have been watching COPD patient moving for decades in rehab center, are able to have an accurate sense of which GOLD level the patients are in while watching a patient walk for a short while.

5.2 Blue Print of the Free-living Health Monitor

Our lab proved that the accelerometer in a regular smartphone can perform equally good as a medical accelerometer in collecting human walking data. In the dissertation study I design a multi-level detection system to make a smartphone as a health monitor for free-living condition, shown in Figure 5.1.

The first step is to address the energy problems present while monitoring health continuously. I initially proposed a low-power screen off passive monitoring strategy to sample potential walking data in daily manner. The low-energy cycle reduce the battery usage significantly, meanwhile it ensures that we have a high chance to capture at least several good walking sample data in each day if the subject walks.

After the data being collected, both activity recognition and position recognition will be performed so that we can extract “good walk”, the term representing walking acceleration samples that has predicting power for health status. The system can then compute the necessary spatio-temporal gait parameters for health status prediction.

The gait speed, which cannot directly be obtained from raw acceleration, requires to be predicted too. This step is for validating the predictive model

in a way that matches the medical gold standard. The distance in 6MWT is proved to be correlated to health status. To obtain the distance information we need to predict speed first since we have the time information recorded. Predicting a medical approved indicator for COPD status help us to demonstrate that the system can actually predict health status directly.

Then the Free-living monitoring step is to obtain the health status information, including the current severity level of COPD and the risk level of the status change.

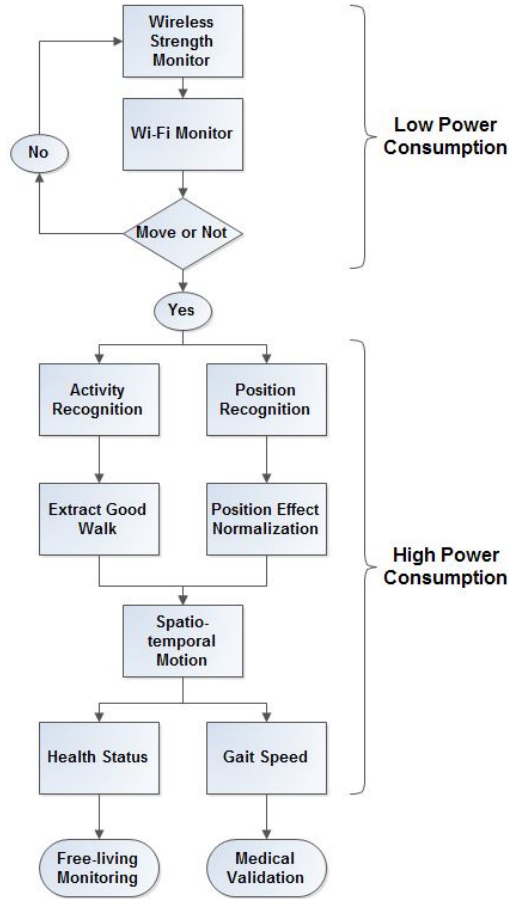


Figure 5.1: The overall design of the free-living health monitor system.

5.2.1 Gait Speed

As our lab proposed in our system design, after pre-processing the system obtains cleaned spatio-temporal motion data collected by phones. Then the major task is to apply proper techniques to make predictive models.

Human mobility can be used to continuously monitor movement and predict health status for responsive treatment. Medical literature shows the distance of 6MWT (6MWD) has the diagnosis power for COPD[34]. The observation result shows that there is significant correlation between mortality and 6MWD, shown in Figure 5.2. The shorter distance a subject walks in a six-minute walk test, the higher mortality this subject suffers. The patients can be group up by 6MWD in diagnosis of their COPD severity. Gait speed and six-minute walk distance excellently predicts poor ($\leq 350\text{m}$) and very poor ($\leq 200\text{m}$) in 6MWT, which indicates the severity of COPD[35].

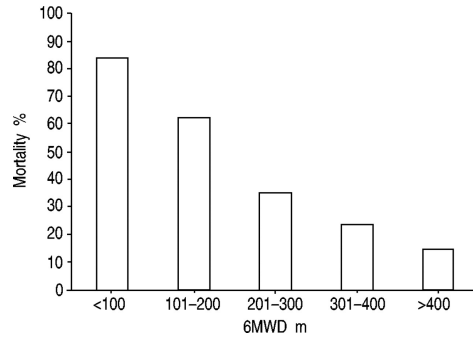


Figure 5.2: Correlation between 6MWD and mortality. The shorter distance a subject walks in a six-minute walk test, the higher mortality this subject suffers.

To better calculate the six-minute walk distance (6MWD), we need to obtain a predictive model for gait speed. The two alternative options here to implement 6MWD prediction models for COPD diagnosis: the first and most straight forward one is to predict gait speed and calculate actual 6MWD values by the product of the length of time interval and predicted speed within this interval; another approach is grouping up COPD patients into different severity groups by their 6MWD and obtain a model to predict which group a specific patient belongs to. Obviously, the first strategy requires regression techniques to train the predictive model, and the second strategy requires a typical classification model. The regression model returns the value of 6MWD help validating our model quality with patients medical records since 6MWT will be operated to lots of COLD patients in their clinical diagnosis and the distance must be recorded. However, the classification model, directly telling which category the patient belongs to, could be more practical and useful in actual COPD diagnosis procedures.

5.2.2 Pulmonary Function

Results of pulmonary function tests (PFTs) are also medically proved have significant predicting power of COPD severity. Pulmonary function tests (PFTs) are performed with a spirometer in clinic conditions to measure severity level of pulmonary diseases. In obstructive diseases, such as COPD, the ratio of Forced Expiratory Volume in 1 second (FEV1) and Forced Vital Capacity (FVC) is used as an effective indicator. The FEV1% predicted is a normalized version of FEV1%, which is defined as the ratio of the patients FEV1% divided by the average FEV1% in the population who are in the same demographic range. The Global Initiative for Chronic Obstructive Lung Disease (GOLD) defines a standard for COPD severity levels based on the predicted FEV1% values. There are four stages GOLD1, GOLD2, GOLD3 and GOLD4 representing four severity levels: mild COPD, moderate COPD, severe COPD and very Severe COPD[36]. The detail of GOLD definition is in Table 5.1.

GOLD stage	COPD Severity	FEV1%
Stage I	Mild	>80
Stage II	Moderate	50-79
Stage III	Severe	30-49
Stage IV	Very Severe	<30

Table 5.1: Global Initiative for Chronic Obstructive Lung Disease (GOLD) Definition. There are four GOLD stages based on FEV1% predicted.

Similar to 6MWD and severity, FEV1% predicted values and GOLD stages could also be regarded as two alternative outcomes of the predictive model. FEV1% predicted value prediction requires a regression model and is proper for medical validation since FEV1% predicted will be recorded in the patients EMR, while GOLD stage prediction requires a classification model and help actual COPD diagnosis procedures.

5.2.3 Oxygen Saturation

Besides, walk distance and PFTs, oxygen desaturation is also a significant indicator of COPD severity[37][38]. To perform walk tests with chronic patients, medical safety requires that the test be performed while the patients are wearing a pulse oximeter[39]. This is a standard medical device, which

measures heart rate and especially oxygen saturation. Oxygen saturation is good overall measure of problems, simultaneous heart and lung, the two primary organ systems circulating throughout the body. The saturation is a good measure of the oxygen within the blood, which is pumped throughout the body by the heart after the oxygen is mixed into the blood within the lungs. Spot checks of oxygen saturation, called SpO_2 , are standard practice for any procedure performed in a clinic or a hospital[40], as this provides a single number of patient status. If the patient desaturates, then the procedure must be stopped, to avoid harm to the patient.

Our software records spot SpO_2 values while measuring it during walk tests, which leaves us a chance to analysis the pattern of oxygen saturation change over time[41]. By observing the SpO_2 variations during 6MWT, we explored correlation between SpO_2 and GOLD levels. Generally, an SpO_2 pattern for GOLD1 will yield higher SpO_2 values during 6MWT. However, by manual observation, there are several subjects observed as a “transition” group between GOLD1 and GOLD2.

CHAPTER 6

EXPERIMENTS DESIGN

The Ph.D. dissertation consists of systematic argumentation on the impact of my studies in the area of telemedicine and e-health. Previous studies brought solutions to medical applications with issues being unsolved. In the left period of my Ph.D. study, the planned work is trying to solve the issues and organize the general knowledge base for this research topic.

Previous studies were all done with a subset of our subjects data because they belong to different research projects. For my dissertation, an integrated study with the whole dataset is preformed. Applying all methods in previous study with the larger dataset is necessary. On one hand, small training set may lead to over-fitting of the model and when the same technique is applied on a larger dataset, the predicting accuracy will decrease. If there is no over-fitting issue, the quality of model could be steadily increasing while the size of training set increases. On the other hand, a larger dataset involves more unique samples, all these samples will contributes to model validation so that it is persuasive that the model covers a larger variation of sampling space.

The dissertation is about all analytic works in this health monitoring system, from the data dataset preparation step to the technical foundation of predictive models, then to the real medical applications, shown in Figure 6.1.

Measuring health status by assessing human motion includes two aspects of technical foundations: feature computation and model training. Feature computation is a crucial step to capture significant gait features from raw sensor data from phones, and machine learning provides a way to manipulate gait to health status. In my Ph.D. dissertation, three different applications are discussed. I discussed about feature selection approaches (FSA) and component extraction approaches (CEA) for computing gait features. In FSA, I applied medical and kinesiological knowledge base to calculate the gait features representing human motions. These features includes descriptions of the acceleration in time domain such as mean, standard deviation

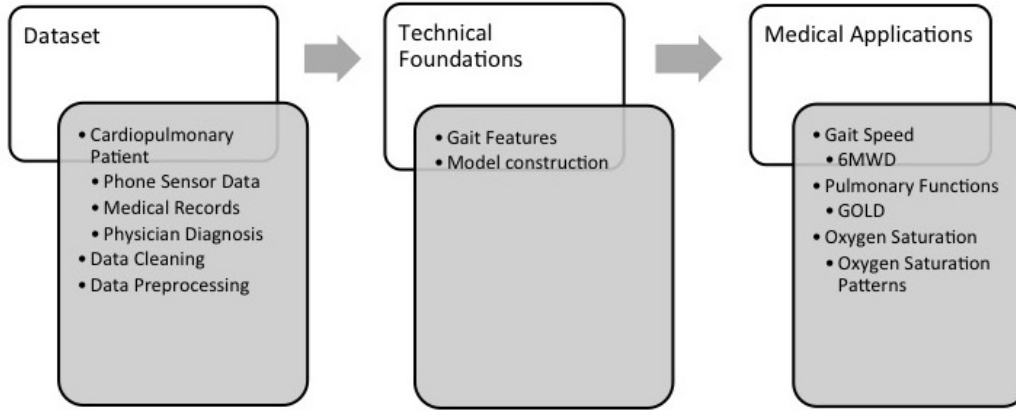


Figure 6.1: Overview of the dissertation study procedures.

and mean crossing rate, and descriptions in frequency domain represents the periodical patterns of the acceleration. In CEA, I generalize the acceleration information of human motion into a type of signals, and signal-processing techniques are applied to select the principal components representing the most significant information in raw acceleration. Our previous work applied various learning methods in previous work to train regression and classification models. The regression techniques I apply are support vector regression, Gaussian process regression. The classification technique I applied are support vector machine classification, and decision tree. In the planned work for completing my Ph.D. dissertation, I will systematically evaluate the proposed technique foundations on the three medical applications.

6.1 Technical Foundations

I proposed two alternative approaches to retrieve features from raw motion data as input for training the models: FSA and CEA. FSA is an intuitive approach based on background knowledge in relevant literature. Theoretically I can calculate all features describing the time-series motion data and then select the most significant subset of these features which lead to an optimized performance of the predictive model. However, in practical I always start from the features has already been used in relevant researches and proved to be efficient. New features should be included gradually until the performance of the model satisfies the requirement. CEA proposes an alternative way to

obtain the features that without any interference of domain knowledge, but from the data itself. I pick PCA as the typical method to realize CEA.

The dataset is assumed to be cleaned and well labeled. So I will pick up supervised learning techniques to train the predictive models. To select appropriate learning models, I consider comparing the state-of-art learning methods with different mechanisms. Decision Trees utilizes tree structure for decision making; linear models are widely applied in regression and classification applications and have a large variation of algorithms to adjust the models to achieve better performance; Support Vector Machine is also proved to be efficient in a variety of classification and regression tasks when the input vector is multi-dimensional.

6.2 Medical Applications

Previously, I talked about predicting gait speed, to medically validate our system. I also trained models to predict GOLD levels of COPD patients, which is used to measure health status for this population. Besides these standard measures of pulmonary patients, I also found that during the 6MWT, patients in stable status and transition status have different oxygen saturation patterns, and such difference can be measured by the spatio-temporal motion of the patients. Different oxygen saturation patterns indicate different risk of transition into a more severe health status. Based on the previous study, I consider three medical applications that motion data can be utilized in health monitoring: 1) gait speed, 2) pulmonary function and 3) oxygen saturation. For each application both the raw value prediction, which is a typical regression problem and the category prediction, which is a typical classification problem are valuable, shown in Table 6.1.

	Regression Models	Classification Models
Gait Speed	6MWD	Severity Levels by 6MWD
Pulmonary Function Test	PFT values	GOLD stages
Oxygen Saturation	SpO ₂ Value	Oxygen Saturation Patterns

Table 6.1: Medical Applications of Motion Data Analysis.

Gait speed. Gait speed is the most direct measure that spatio-temporal motion data can be transferred to. The previous study the input parameters are highly correlated to the speed of patients when they are in the walk

tests. At the same time, the six-minute walk distance is medically proved to have prediction power for COPD severity, which can simply be calculated if the real-time gait speed can be predicted during the walk test. For medical validation, a regression model is required to return real-time gait speed for calculating six-minute walk distance, which can be used to evaluate the model by comparing to the actual six-minute walk distance in the medical record. Meanwhile, a classification model is needed to return a predicted states regarding to our system design, which predicts the severity levels defined by walk distance, in Table 5.2.

Pulmonary function. Pulmonary function tests are a set of values helping diagnosis of COPD, and the predicted FEV1% is among the most crucial PFTs which defines the GOLD stages to describe severity. Similar as for gait speed, for the prediction of FEV1% and other PFTs values a regression model is required and the accuracy can be evaluated with the actual PFTs in medical records. While a classification model yields to a prediction of GOLD stages, which represents the current health status of the COPD patient.

Oxygen Saturation. Oxygen saturation represents the real-time cardiac pulmonary variation in human body[42]. Our previous work discovers that oxygen saturation can be used to predict GOLD stages. Additionally, specific oxygen saturation patterns indicates “intermediate” status in between two GOLD stages, which I defined as the transition stages. The study initially showed the transition stage between GOLD1 and GOLD2 can be detected by oxygen saturation pattern[43]. Oxygen saturation can be measured by pulse-oximeters, indicated as SpO_2 . Predicting raw SpO_2 value requires a regression model. However the classification model predicting categories of oxygen saturation patterns could be rather useful. It is not only because the regression model predicting raw SpO_2 introduces high error rate, but also because the instantaneous SpO_2 value, just as other crucial vital signs, heart rate, temperature, etc., is less relevant to the health status or risk[44]. Instead, detecting oxygen saturation pattern, tracking a comparatively long period of the SpO_2 values, correlates more with medical diagnosis from real physicians[43].

6.2.1 Experiments

After defining the technical foundations and medical applications, I have created the road-map of experiments design for my dissertation. Totally there are six variation of techniques after combining the feature selection and model training processes, and all these techniques can be applied to train predictive model for each medical application. Thus, there will be thirty-six experiments to be operated in total, showed in Table 6.2.

Applications	Foundation 1	Foundation 2	Foundation 3
6MWD	Experiment 1	Experiment 2	Experiment 3
Severity by 6MWD	Experiment 4	Experiment 5	Experiment 6
PFTs	Experiment 7	Experiment 8	Experiment 9
GOLD stages	Experiment 10	Experiment 11	Experiment 12
SpO ₂	Experiment 13	Experiment 14	Experiment 15
Oxygen Saturation Pattern	Experiment 16	Experiment 17	Experiment 18
	Foundation 4	Foundation 5	Foundation 6
6MWD	Experiment 19	Experiment 20	Experiment 21
Severity by 6MWD	Experiment 22	Experiment 23	Experiment 24
PFTs	Experiment 25	Experiment 26	Experiment 27
GOLD stages	Experiment 28	Experiment 29	Experiment 30
SpO ₂	Experiment 31	Experiment 32	Experiment 33
Oxygen Saturation Pattern	Experiment 34	Experiment 35	Experiment 36

Table 6.2: Planned experiments on the combination of technical foundations and medical applications.

Each medical application direction contains two problems: regression problem to predict raw values and classification problem to categorize status. For gait speed, since the goal is to match the real measured speed for medical validation, the regression problem is selected. Then the pulmonary function and oxygen saturation studies are about predicting real status, so the classification problems are selected as the main problems to be solved, show in Table 6.3.

Applications	Foundation 1	Foundation 2	Foundation 3
6MWD	Experiment 1	Experiment 2	Experiment 3
Severity by 6MWD	Experiment 4	Experiment 5	Experiment 6
PFTs	Experiment 7	Experiment 8	Experiment 9
GOLD stages	Experiment 10	Experiment 11	Experiment 12
SpO ₂	Experiment 13	Experiment 14	Experiment 15
Oxygen Saturation Pattern	Experiment 16	Experiment 17	Experiment 18
	Foundation 4	Foundation 5	Foundation 6
6MWD	Experiment 19	Experiment 20	Experiment 21
Severity by 6MWD	Experiment 22	Experiment 23	Experiment 24
PFTs	Experiment 25	Experiment 26	Experiment 27
GOLD stages	Experiment 28	Experiment 29	Experiment 30
SpO ₂	Experiment 31	Experiment 32	Experiment 33
Oxygen Saturation Pattern	Experiment 34	Experiment 35	Experiment 36

Table 6.3: Planned experiments on the combination of technical foundations and medical applications.

CHAPTER 7

SUBJECT AND DATA COLLECTION

The data collection process was initiated back to 2012, collaborating with three different research health systems:

University of Illinois Hospital & Health Sciences System;

Research Institute of Carle Hospital at Urbana;

NorthShore University Health System.

The data collection protocol varies for each health system but generally all data collection procedure contains at least one of the two parts: a well executed six-minute walk test[45] and a free-walk with random activities. Totally, our lab have collected data from pulmonary patients with IRB approval. All data are de-identified for analysis.

7.1 Pilot: Collaboration with University of Illinois Hospital & Health Sciences System

Our lab collaborated with University of Illinois Hospital & Health Sciences System at Chicago since 2013. In about six months we operated 6MWT on 29 patients in clinical settings with a 15-meter corridor. However, since the recruited patients are not all pulmonary and we didn't restrict walking conditions (the phone not fixed on the L3 position, patient walking was walkers, etc.) and the alpha version of our software crashed on several patient tests. The dataset is also extremely biased on gender, containing only five male subjects. The operation procedure did not force all patients to wear pulse oximeter so we barely have oxygen saturation values recorded for these patients. For better data quality control, we may eliminate the entire dataset from University of Illinois Hospital & Health Sciences System.

7.2 Collaboration with Carle Foundation Hospital

In 2014, we started our collaboration with Carle Foundation Hospital and operated walk test on 28 patients and 10 healthy subjects. Patients and healthy subjects were requested to perform a full six-minute walk test, and additionally some patients performed a short period of free-walking after the 6MWT. None of the healthy subjects performed free walking. For patients, we obtain their basic demographic information and PFT results from their EMR. For healthy subjects, we obtain their basic demographic information. In the oxygen saturation analysis, we focus on ambulatory patients who are the most mobile as the best candidates for healthy monitors in daily living. So we select data of GOLD1 and GOLD2 patients and eliminate the 2 GOLD3 patients. We also eliminated 2 subjects who did not complete the walk test and 2 subjects for whom the pulse oximeter did not record. We also eliminated 2 subjects whose medical conditions caused the pulmonary function for validation have changed significantly. The PFT no longer reflected the function being measured by the 6MWT, due to time delay between tests. The first had a diagnosis of bronchiectasis, a lung condition where function changes within short periods. The second had change of medication between PFT and 6MWT, with our model correctly showing a marked increase in function. All subjects were permitted to use their inhalers before testing, but she changed from albuterol (for asthma) to tiotropium (for COPD). Of initial 28 patients tested, this left 26 patients and 10 healthy subjects for gait speed analysis and 20 patients for oxygen saturation analysis.

7.3 Collaboration with NorthShore University HealthSystem

In 2015, we initiated our current collaboration with NorthShore University HealthSystem and tested 27 patients. In the protocol, each patient will perform multiple 6MWTs in several months, with a full PFT at first time. 10 of the 27 patients performed at least twice walk tests (7 twice and 3 three times).

Overall, I will combine the datasets from Carle Foundation Hospital and NorthShore University Health System for the analysis in my dissertation.

That is over 60 walk test samples for 50 real pulmonary patients.

Pulmonary function tests are performed with a spirometer in clinical conditions [5]. In obstructive diseases, such as COPD, the ratio of forced expiratory volume in one second (FEV1) as compared to an age, gender, race, and height adjusted expected value is used as a sufficient indicator to measure severity level of the disease, called predicted FEV1%. Based on the predicted FEV1% values, the Global Initiative for Chronic Obstructive Lung Disease (GOLD) defines a standard measure for cardiopulmonary diseases, evaluating the severity levels from mild (GOLD 1, FEV1%: ≤ 80), moderate (GOLD 2 FEV1%: 50 - 79), severe (GOLD 3, FEV1%: 30 - 49) and more severe (GOLD 4, FEV1%: < 30). In this study, all healthy subjects are labeled as GOLD 0, indicating that they do not have cardiopulmonary diseases. Since the more severe (GOLD 4) patients have significant lack of mobility, measuring their health status is a difficult with motion sensors. So we only focus on GOLD 0 to GOLD 3 in this study.

Our phone application MoveSense [46] was installed on the smartphones (a high-end Samsung Galaxy S5 cost \$600 and a low-end LG Optimus Zone2 cost \$60) for motion data collection. Previous hardware experiment showed that the accelerometers in high-end and low-end smartphones are identical for human motion capturing, and both of them are equally performing as a high-quality medical accelerometer. We primarily collect data from high-end smartphone for analysis and keep the low-end smartphone data for backup. Once the high-end phone failed, we will use the low-end phone data instead since we previously showed that the low-end phone keeps equivalent quality of sensor data as high-end phones [46]. After de-identification, the phone sensor data, together with other demographic information and test results, are stored into an archival database.

7.4 Dataset in Main Study

In the dissertation study, I combined the datasets from two of the above: one is from the collaboration with Carle Foundation Hospital and the other is from the collaboration with NorthShore University Health System. 28 patient participated in the experiment in Carle Hondation Hospital, each of the was operated a standard six-minute walk test under the supervision of

nurses in the rehabilitation center. The NorthShore experiment currently has recruited 27 patients, each of them complete at least one standard six-minute walk test in clinical settings. As the description in the approved IRB, we expect that most of them will have a second six-minute walk test operated in the following months. Thus, overall the dataset will contain walking data from at least 55 pulmonary patients. That will be totally 82 six-minute walk test sessions, containing 324,720 sample points of good walking motion since the sampling rate is set as 60Hz.

CHAPTER 8

TECHNICAL APPROACHES

This study requires techniques from hardware level of sensor data retrieval to analytic level of health status analysis. In this section I will talk about the proposed technical approaches for monitoring human health with smart phone sensors.

8.1 System Design

At hardware level, testify of phone sensor eligibility to be compatible to medical devices matters. Our lab have developed a continuous sensing middleware designed to overcome hardware and firmware limitations in current smartphones, by converting them into medical quality monitors. We have both GaitTrack [47] [48] and MoveSense [29] [46] [43] implemented with the same middle-ware design, to adjust the phone software into different medical applications.

Such middleware framework can be installed on smartphones without purchasing extra sensors or gaining special access or “rooting” the phone. We have implemented both GaitTrack and MoveSense on Android with future plans to port to iOS.

8.1.1 Hardware

There might be concerns on that phone sensors would not be comparable to medical accelerometers; however, the accelerometers in smartphones is proved to be similar or better than the accelerometers used in medical devices in this specific experimental conditions[18]. Accelerometers measure the relative acceleration of the devices spatio-temporal motion. Modern accelerometers are all micro-electromechanical systems (MEMS) with relatively

similar operating characteristics are used to evaluate the capability of the accelerometer: range, sensitivity and sampling frequency. We postulate that cell phone hardware is capable of medical quality readings by comparing the capabilities of sensors in smart phones to the Actigraph GT3X, the most widely used monitoring device in clinical kinesiology trials, the Zephyr Bio-Harness, a popular medical accelerometer used to monitor firefighters during rescues, and the Fitbit Flex and Jawbone Up24, two popular fitness devices.

The analysis indicates that most phones in a specific generation tend to have similar accelerometers with roughly 16g sensitivities. All chips were capable of reading at higher frequencies than required for human motion. While accelerometer chips in phones have similar operating characteristics, medical devices tend to have more variation. The slow regulatory process in the medical community often causes these devices to have older sensor technology. More interestingly, current fitness devices (Fitbit Flex and Jawbone Up24) have same MEMS that are found in smartphones. On framework level, we test three different price range phones (Low: Zone, Middle: Ace, High: Mini) and all three phones maintain 100Hz sampling frequency, while keep the CPU usage lower than 50%, in Figure 8.1.

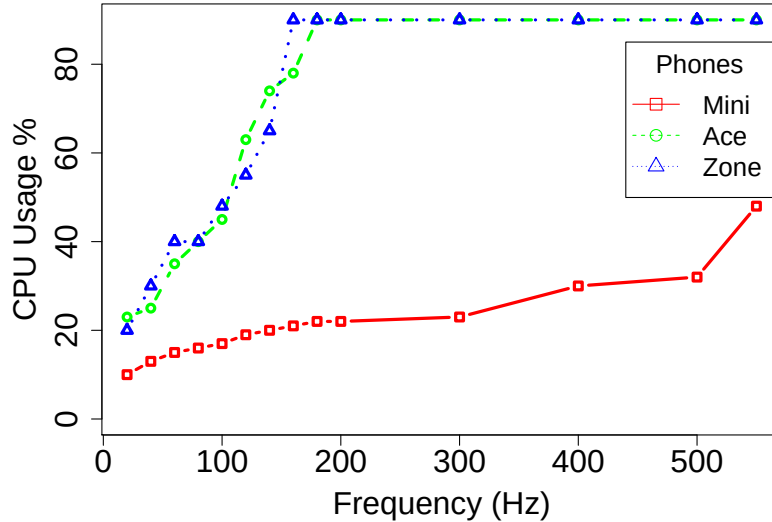


Figure 8.1: CPU usage in percentage while sampling frequency increases for three common smartphones

A medical health monitor must sample at a frequency sufficient to capture

a patients movements accurately. While the upper-end of medically validated devices choose to sample at 100Hz, medical studies tend to put the range of human motion much lower at 3-5Hz[22][49], show in Figure 8.2. If it is simply measuring stepping, against the Nyquist rate, a monitor sampling in 10Hz is enough.



Figure 8.2: Guaranteed sampling frequency of Devices measuring gait. Human walking is usually in 3-5Hz which means to measure stepping a monitor of 10Hz sampling is enough.

8.1.2 Software

Our lab build the health monitor on Android smartphones. While android 4.4 puts increased emphasis on battery efficient continuous monitoring, the system is not designed to implement a medical quality monitor which requires fixed frequency readings. Android leaves it up to developers to implement drivers to access sensors leaving wide variation in maximum polling frequency and sensor behavior among devices. Thus the sensor chips themselves may be capable of sampling up to 800Hz and beyond, but the sampling frequency may be substantially lowered by the phones processing capability and driver implementations.

The software overcomes these limitations by implementing its own sensor reading queue. The Android sensor manager continuously queries the sensors to access the raw readings and dispatches `OnSensorChanged` messages to the software end. The software spawns a high priority thread to process the accelerometer sensor readings, record the time-stamp and magnetometer and insert this information into a first in first out queue. We implements a data handler thread to continuously poll the queue for new sensor readings. The data handler pops sensor readings and calculates the interval to the next readings assuming a fixed sampling frequency. If the readings are coming in

at greater than the fixed frequency, the algorithm will average all readings in the current interval. If the readings are coming in at less than the fixed frequency, the algorithm will extrapolate the missing values using a linear approximation algorithm. The final output from the accelerometer pipeline is a fixed frequency sampling from both the accelerometer sensor. Thus, the software uses the processing power in the phone to get an accurate fixed frequency stream of sensor data at the expense of processor cycles. We finally note that our algorithms could easily be implemented in dedicated hardware in line with the industry trend to use dedicated hardware to decrease CPU and power usage.

To test the performance of our firmware, three phones from different vendors of various quality are applied, including a high-end Motorola Droid Mini, a mid-range last generation Samsung Galaxy Ace and a lowend LG Optimus Zone. We are interested in validating the ability of a cell phone to operate at the same accuracy as the Zephyr BioHarness, a medical grade device.

8.2 Gait Parameters

The smart phones recorded three-dimensional acceleration from the accelerometer integrated in a Motorola Droid Mini during 6MWTs. The raw acceleration was sampled at 60Hz. Previous work has demonstrated that down sampling from 100Hz to 10Hz does not impact gait parameters [50] [51]. Thus, a low-pass-filter at 10Hz was set up to eliminate high frequency noise. After filtering, we compare two different approaches for retrieving spatio-temporal gait parameters [12] from the phone acceleration data including the feature selection approach (FSA) and the component extraction approach (CEA).

For preprocessing, I manually labeled the starting and stopping point of 6MWT for each session to clean the input data. All walking acceleration data are split into ten-second pieces, with five-second of overlap between each interval to avoid information loss.

8.2.1 Feature Selection Approach

The Feature selection approach (FSA) is based on empirical knowledge. A primary parameter to measure gait is the cadence, which is the number of

strides of a unit of time[15]. Previous studies in motion tracking and activity recognition by wearable accelerometers extract a series of features from raw acceleration data[20] [52] [31] [53] [14] [54]. In the thesis research I defined fifteen features in the time and frequency domains to track motion.

In the time domain, I selected the mean and standard deviation of acceleration which describe the general distribution of the acceleration sample. Additionally, the mean crossing rate (MCR), root mean square (RMS), autocorrelation coefficient (AC) and coefficient of variance (CV) were calculated from time-series acceleration data. The mean crossing rate (MCR) represents the ratio of above and below acceleration.

$$MCR = \frac{1}{N} * \sum_{t=1}^N I_t \{I_t = 1 | (a_{t-1} - \bar{a})(a_t - \bar{a}) < 0\} \quad (8.1)$$

The root mean square (RMS) is a statistical measure on the variation of signal magnitude.

$$RMS = \sqrt{\frac{1}{N} * \sum_{t=0}^N a_t^2} \quad (8.2)$$

The Autocorrelation coefficient (AC) measures periodical similarity in time domain.

$$AC = \max \{ \forall_{t \in N, i < t} (\sum_{t=0}^N a_t \bar{a}_{t-i}) \} \quad (8.3)$$

The coefficient of variance (CV) is a normalized measure for dispersion of discrete samples.

$$CV = \frac{\sigma_a}{\mu_a} \quad (8.4)$$

In the frequency domain, I computed the peak frequency (PF), Shannon entropy, subband energy and subband energy ratio. The peak frequency (PF) represents the frequency of peak magnitude in spectrum.

$$P_f = \arg \max_f \{x_f | f = 0, \dots, N\} \quad (8.5)$$

I calculate the Shannon entropy [55] to describe the information contained

in the acceleration spectrum.

$$H = - \sum_{f=0}^N x_f * \log x_f \quad (8.6)$$

The subband energy and ratio represents the proportional information of frequency spectra. Energy was calculated for four different bands in the frequency domain including $B1(0, 1]$, $B2(1, 4]$, $B3(4, 10]$, $B4(10, 30]$ and the subband energy ratios $B1/B2$, $B3/B4$ and $(B1 \cup B2)/(B3 \cup B4)$. In total, fifteen features were computed from the raw acceleration data.

There are more features can be potentially retrieved from raw acceleration data and applied in medical applications [50]. Following works will expand the feature space if it is necessary to obtain higher performance.

8.2.2 Component Extraction Approach

Alternatively, I am planning to select input features with a novel components extraction approach (CEA), extracting signal features from both time-domain and frequency domain by optimization [56]. The intuitive feature selection approach applies the magnitude of three-dimensional acceleration to calculate gait parameters. However, for signal processing, individual axis of three dimensional acceleration contains unique signal information. CEA is aimed to retrieve sufficient information for health status prediction.

Independent component analysis (ICA) [57] can be applied to formalize the three-dimensional acceleration so that the phone axis will be transformed to a fixed 3-d axis [58]. After formalizing the raw acceleration, dynamic linear models (DLM) can be applied for data augmentation and retrieving significant information on temporal space [59]. Primarily, DLM can be applied to make use of information of current time t to predict the next time $t + 1$, with the equation

$$y_t = F_t x_t + v_t, v_t \sim N(0, V_t) \quad x_t = G_t x_{t-1} + w_t, w_t \sim N(0, W_t). \quad (8.7)$$

y_t are observations at time t and vector x_t contains the unobserved states that are assumed to contribute in time according to a linear operator G_t [60]. DLM returns a predicting parameter which contains the information whether

the signal are periodically stable or not. While stability of walking strongly correlated with the severity of COPD, DLM is potentially able to extract sufficient temporal information on the raw three-dimensional acceleration.

In frequency domain, I apply principal component analysis (PCA) to extract the principal components. To transform the temporal signals in frequency domain, I applied fast Fourier's transform (FFT) since the ten-second piece of walking can be regarded as extendable and periodical. Principal component analysis is helpful to find significant components from massive data [19]. PCA is widely applied in dimensionality reduction, which is an important pre-processing progress for learning with high dimensional data. In this study, the frequency spectra of walking acceleration pieces are high dimensional data. I apply principal component analysis to reduce dimensionality of raw dataset before training. Assume matrix $X = \{\vec{x}_i | i = 1, \dots, N\}$ is set of frequency spectra, It is possible to calculate the covariance matrix Σ for X . Then, I calculate the matrix of eigenvectors W and eigenvalues $\{\lambda_i | i = 1, \dots, N\}$ for the covariance matrix Σ . Then, I order W by $\{\lambda_i\}$ to obtain all principal components of X , and I have

$$T = X W^T \quad (8.8)$$

To reduce dimensionality, I only select first- l qualified eigenvectors W_l and generate the first- l principal components by

$$T_l = X W_l^T \quad (8.9)$$

I applied the Kaiser's stopping rule for selecting principal components. Kaiser's stopping rule states that only the number of factors with eigenvalues $\lambda_i \geq 1.0$ should be considered [61]. After selecting k principal components by Kaiser's stopping rule, I iteratively calculate the validation scores by increasing l from 1 to k and find the local optimal l which maximizes the validation score. l is set to k when no local optimal score is obtained.

8.3 Learning Techniques

After obtaining the gait parameters as input, there are several learning techniques can be applied to train the predictive models. As I discussed in the

previous section, both classification models and regression models could be applied for different perspectives[62]. I considered various classification and prediction models and evaluate their capability of predicting health status measures for each medical application.

8.3.1 Classification Models

Classification models are designed to predict what class a group of input data belongs to or indicates[63]. Each model is trained with a set of input parameters and target values. The model then determines cutoff points in the data to later classify unknown input as one of the previously trained classes. Then I evaluate various classification models in their ability to predict classes including phone identification, user identification, gender, activity, and COPD status.

Linear Model for Classification

A liner classification model makes classification decisions by the value of a linear combination of the input parameters. If the input feature vector to the classifier is $\vec{x}$ then the output score is as Equation (8.10),

$$y = f(\vec{w} * \vec{x}) = f\left(\sum_j w_j x_j\right) \quad (8.10)$$

where $\vec{w}$ is a real vector of weights and f is the adjusted function by model training which split input points in different classes. Generalized linear model (GLM) allows error distribution other than normal distribution for response variables[64]. It introduces variation of the model kernel to handle a larger variation of classification problems.

$$P(y|x_1, x_2, \dots, x_N) = \frac{P(y)P(x_1, x_2, \dots, x_N|y)}{P(x_1, x_2, \dots, x_N)} \quad (8.11)$$

Support Vector Machine (SVM)

Support vector machines are a class of machine learning algorithms which allow classification by remapping training data into a higher dimensional

space and determining sets of hyperplanes that separate classes[65]. The hyperplanes are called “support vectors”, optimizing the separation between data points in the given space. This yields the classification equation (8.12) which requires a training set to simply find the variables c_i to determine $f(x)$ given the kernel function and inputs.

$$f(x) = \sum_{i=1}^N c_i * K(x_i, x) \quad (8.12)$$

$$k(x_i, x_j) = e^{(-\tau \|x_i - x_j\|^2)} \text{ where } \tau = 1/2\sigma^2 \quad (8.13)$$

From Equation (8.12) we know that different kernel functions affect the overall performance of the trained model, previously I applied simple linear kernel in model training and obtain the technical baseline[47],[46], shown as Equation (8.13). In future study I will also applied RBFSVM, using the radial basis function kernel to train the SVM model, as Equation (??)[66]. The hyperparameter $\gamma = 1/D$ is set based on D , the number of dimensions in training samples. x_i and x_j are two arbitrary samples in the training set. In our dataset patients status are labeled as one of the three GOLD stages (GOLD1, GOLD2 and GOLD3). One-against-all strategy is applied to train this three-class classification model. The tolerance of termination criterion is set as 0.001 and the insensitive-loss variable ϵ is set as 0.1.

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma = 1/D. \quad (8.14)$$

Tree-based Classification Model

Decision trees applying a tree structure for searching proper decisions based on the given conditions[67]. At each node a binary decision is made for the input vector. For each potential split, the algorithm calculates the fraction of targets which are contained in each potentially created leaf (8.15). Random forests is constructing a multitude of decision tree at training time and outputting the overall optimized class as the final decision, which provides a more robust result if there are a large training set and corrects the problem

of overfitting [68].

$$\begin{aligned}
f_i &= \frac{1}{N} \sum_{k \in N} I(y_k = i) \\
H(f) &= \sum_{n=1}^N f_n(1 - f_n) = \sum_{n=1}^N (f_n - f_n^2) \\
&= \sum_{n=1}^N f_n - \sum_{n=1}^N f_n^2 = 1 - \sum_{n=1}^N f_n^2
\end{aligned} \tag{8.15}$$

Comparing to SVM, traditional decision tree or random forest does not lead to better performance but the tree structure model provide more interpretative information for how the predictions are made. Such hierarchical decision structure helps better understanding the medical mechanism behind the predictive model.

8.3.2 Regression Models

Regression models predict the value of a continuous variable based on a given input vector. I consider applying multiple types of regression models to predict values of the medical measurements in three applications: gait speed and six-minute walk distance, FEV1% predicted and SpO₂.

Gaussian Process Regression (GPR)

GPR applies statistical distributions in a basic linear regression. In Gaussian process, every point in some input space is associated with a normally distributed random variable[69]. GPR was found to be suitable for treadmill walking by Vathsangam et al.[19].

I defined an estimator $\hat{y}$ to estimate the speed given a Gaussian function $h()$ to transform each input variable x to speed. $\hat{y}$ is also a Gaussian, as Equation (8.16),

$$\hat{y}_i = h(x_i) + \epsilon_i, i = 1, \dots, N \tag{8.16}$$

with noise ϵ giving a Gaussian $N(h(x_i), \epsilon_i)$ leaving that $h()$ is also Gaussian such that $N(0, k)$ where k is the specified kernel function estimating the covariance between x values. We defined the values X as input values

we have ground truth from training and X_* as the input values we have to test. Define $K(X, X_*)$ to be a m by m kernel matrix such that

$$(K(K, K_*))_{ij} = k(x_i, x_{*j}). \quad (8.17)$$

We can then define

$$\begin{bmatrix} h \\ h_* \end{bmatrix} | X, X_* \sim \mathcal{N}(0, \begin{bmatrix} K(X, X) & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}) \quad (8.18)$$

and

$$\begin{bmatrix} \epsilon \\ \epsilon_* \end{bmatrix} \sim \mathcal{N}(0, \begin{bmatrix} \sigma^2 I & 0 \\ 0^T & \sigma^2 I \end{bmatrix}) \quad (8.19)$$

which combined gives our estimator

$$\begin{bmatrix} \hat{y} \\ \hat{y}_* \end{bmatrix} | X, X_* = \begin{bmatrix} h \\ h_* \end{bmatrix} + \begin{bmatrix} \epsilon \\ \epsilon_* \end{bmatrix} \sim \mathcal{N}(0, \begin{bmatrix} K(X, X) + \sigma^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) + \sigma^2 I \end{bmatrix}). \quad (8.20)$$

Then Equation 8.21 follows directly from rules conditioning Gaussians. Therefore given an training set X and a set of corresponding kernel functions, we can obtain an estimate for the speed of $\hat{y}$.

$$\begin{aligned} P(\hat{y}_* | \hat{y}, X, X_*) &\sim \mathcal{N}(\mu_*, \sigma_*) \quad \text{where} \\ \mu_* &= K(X_*, X)(K(X, X) + \sigma^2 I)^{-1} \hat{y} \\ \sigma_* &= K(X_*, X_*) + \sigma^2 I - \\ &\quad K(X_*, X)(K(X, X) + \sigma^2 I)^{-1} K(X, X_*) \end{aligned} \quad (8.21)$$

Support Vector Regression

Part et al. evaluated the user of support vector regression (SVR)[70] to predict walking speed [33]. SVR model is built by using a standard Gaussian kernel (8.13). Features are mapped to higher dimensional space $g(x)$ with kernels $K(x, x)$ defined as the distance between the non-linear hyperplanes. This yields the classic equation (8.12) which requires a training set to simply find the variables c_i to determine $f(x)$ given the kernel function and inputs. The exactly same kernels in classification SVM models can be applied in regression SVM models.

Regression Tree

Construction of regression tree is similar to constructing a decision tree. Unlike the decision tree which split based on information scores, the regression tree minimize the mean square error on each node for each set of training data (8.22). The leaf node holds the average value seen in the training data for its respective branch in the tree. For a given input vector X the predicting value is the average value in the leaf node. Random forests can also be used to conquer over-fitting in regression tree.

$$H(f) = \frac{1}{N} \sum_{n \in N} (y_n - c_f)^2 c_f = \frac{1}{N} \sum_{n \in N} y_n \quad (8.22)$$

8.4 DPClass: Discriminative Patterns-based Classification

In previous section I discussed several state-of-art machine learning techniques that would be applied in this study. Besides, this study also applies a novel pattern mining approach, DPClass, which is designed by Jingbo Shang, one of my colleges in University of Illinois [71]. This section briefly talks about why I choose DPClass even the state-of-art methods. like SVM, achieve high prediction scores.

8.4.1 Disadvantages of State-of-art Models

Generalized linear classification models, such as support vector machine and logistic regression are usually have reasonably good performance. We applied several of these state-of-art classification models to obtain high prediction accuracy. Tree-based models, are designed by the nature of decision making process, which will lead to a clear interpretation of the model mechanism. Neural networks models nonlinear relationships from cognitive learning with well established training sets, providing a different thought for classification. However, each type of classifiers has its downside, which means, we have to compromise between the results and the models we selected, dealing with different problems.

In real world applications, linear models are preferred because the simplicity and efficiency. Also it is because people mostly just care about the outcome, but not how the model works. Deep learning and neural networks are also in the similar situation, if the problem is hardly being solved linearly, people will try to build a non-linear neural network structure and train the model with well selected and well labelled training data. Once again, the neural network models all work as a “black box” and we have no idea how the input and output of the model are correlated in each step when the model makes the decision. Comparatively, a tree-based model, such as simple decision tree and random forests, are simulating the natural steps of decision making, which makes the interpretation of model easy and straightforward. The down side is the tree models are not good at dealing with high dimensional dataset and the performance are not as good as other types of model. It will also be easily over-fitted if the training data are biased.

8.4.2 DPClass: a Hybrid Novel Approach Considering Model Accuracy and Interpretation

I applied DPClass [71], an effective but concise discriminative pattern-based classification framework, to obtain a predictive model for GOLD, where a discriminative pattern is a conjunctive clause of several conditions, for example, $(\text{Age} < 80.5) \text{ AND } (\text{Height} \geq 1.5)$. DPClass can automatically extract the top discriminative patterns in model training. The generated discriminative patterns, or discriminative rules reflect interactions of all input features.

Pattern-based classification, such as PatClass and DDPMine, was originally proposed to improve the prediction accuracy by constructing high-order interactions among features using hundreds or thousands of frequent patterns. Multi-tree based models, such as Random Forest, have demonstrated advantages of classification accuracy and handling both numerical and categorical features. DPClass takes the advantages of both concepts and proposes to resolve classification problems using discriminative patterns.

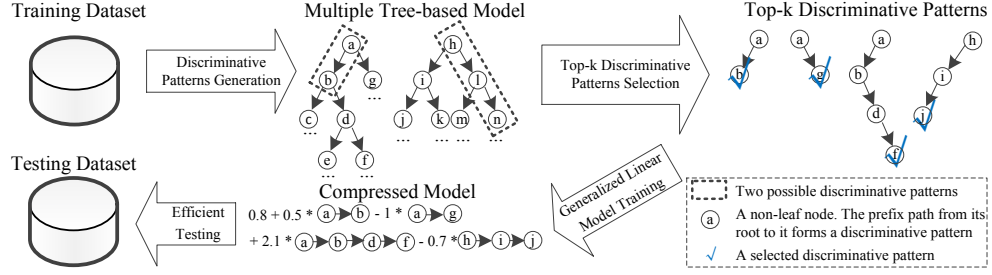


Figure 8.3: Overview of DPClass.

8.4.3 Candidate Discriminative Patterns

The first step of DPClass is to generate high quality discriminative patterns. As shown in Figure 8.3, DPClass starts from training data and train many (e.g., 100) random decision trees [72]. The random decision tree, introducing randomness via bootstrapping training data, selects features and assigns each feature to specific tree nodes. Every prefix path from the root of any tree is adopted as a candidate discriminative pattern. By the default setting, about 10 thousand candidate discriminative patterns are usually generated.

The length of discriminative patterns is restricted with a limited bar D , and no less than the length of δ . To fit these basic requirement, I define the prefix path starting from the root node all the way down to its non-leaf node as a discriminative pattern. Particularly, random forest is the technique we applied here to train the discriminative patterns, due to the nature of multiple trees constructed in model training. Also, by limiting the depth of trees, the model is able to build constraints on δ and D .

8.4.4 Pattern Space

Now, instances in their original feature space can be mapped into the pattern space using a set of discriminative patterns. For each discriminative pattern, there is one corresponding binary dimension describing whether the instances satisfy the pattern or not. Because the dimension of the pattern space is equal to the number of discriminative patterns which is a very large number after the generation phase, we need to further select a limited number of patterns and thus make the pattern space small and efficient.

8.4.5 Top-k Discriminative Patterns

After all potential discriminative patterns are generated, we need to select the top- k discriminative patterns to identify the most informative and interpretable patterns. To achieve the best performance and find complementary patterns, two effective strategies are proposed in DPClass: forward selection and LASSO, which makes decisions based on the effects of the pattern combinations instead of considering different patterns independently.

Forward Pattern Selection

It is a simple and efficient way to select discriminative patterns, one by one while each newly added discriminative pattern is the best choice at that time. Since we use a generalized linear classification model to combine the selected discriminative pattern, when we have fixed the first k' discriminative patterns. We empirically add one more discriminative pattern with the biggest improvement of the training classification accuracy based on $k' + 1$ discriminative patterns.

LASSO based Pattern Selection

L1 regularization (i.e., LASSO [73]) is designed to make the weight vector sparse by tuning a non-negative parameter λ . Since we are actually selecting features in the pattern space, for a given λ , we optimize the following loss function to get a subset of patterns which are most important.

$$\mathcal{L} = \sum_i^n l(\mathbf{x}'_i^T \mathbf{w}, y_i) + \lambda \cdot \|\mathbf{w}\|_1 \quad (8.23)$$

where, $\mathbf{x}'_i$ is the mapped binary feature representation in pattern space of i -th instance; y_i is the label of i -th instance; $\mathbf{w}$ is the weight vector in the generalized linear model; $l(\cdot, \cdot)$ is a general loss function such as logistic loss. To ensure there are at most k patterns having non-zero weights in the pattern space, the binary search technique can be utilized to automatically determine λ , assuming that there exists some hidden importance ranking among features and once the weight of a feature becomes non-zero in a given $\lambda = v$, it will also be non-zero for almost any smaller $\lambda < v$.

8.4.6 Final Multi-class Classification

After selecting top-30 discriminative patterns, a 30-dimensional pattern space is created and feed these features to train a multi-class classification model, where we adopt the implementation in GLMNET [74].

8.5 Model Evaluation

To avoid over-fitting problem, 10-fold cross validation is applied for self-validating the classification model [75]. That is, randomly split the training set into ten parts, and each time select one out of the ten subset as the test set and the other nine to train them model. After ten folds of model training and testing, we obtain the predicting results of all samples in the dataset with a model trained by other samples.

A classification model predicts GOLD stages for each ten-second walking samples. However, each six-minute walk test contains multiple ten-second walking samples. As a typical chronic disease, COPD status will not change within a single walk test. But a single walking sample may be affected by other factors and lead to false prediction. Majority voting is a most straightforward and robust rule to obtain single decision out of multiple sample results [76]. Majority voting works sufficiently in our previous study [48] [43]. We apply the same strategy to obtain predicted GOLD stages for each patient. Additionally, different portion of GOLD prediction for samples in one 6MWT session indicates different risk of status transition if we assume the model work sufficiently.

In addition, 11 subjects who have follow-up visits to the hospital and take the experiments several weeks after their initial visits. All these 6MWT data are selected as the actual test set to: 1) evaluate model performance; 2) analyze medical causality if the status change is confirmed.

CHAPTER 9

PREDICTION OF PULMONARY FUNCTION

9.1 Pilot Study I: Severity Can Be detected for Patient in the Same Demographic Cohort

The first pilot study of prediction pulmonary function test started from the collaboration with University of Illinois Hospital and Health System back to 2012.

The protocol of clinical experiment were still under debating back to the then and there were lots of problem while collecting patients data, experimentally or technically. Thus the data quality is very poor and we separate this dataset from our main study. However, the story of pilot study is still very intuitive and it actually helped us to fix the direction of the main study.

As shown in Table 9.1, there are 22 female subjects recruited from University of Illinois Hospital and Health System, whose walking data are used for this part of study. The support vector machine model using the gait parameters was used to predict FEV1% (and hence GOLD) for the 22 female COPD patients, whose demographics are shown in Table 9.1. The results of this model are shown in Table 9.2, where the unified prediction for all females provided 71.73% accuracy, whereas the cohort prediction of females between the ages of 50 and 64 years provided 89.22% accuracy. The gait model is the same in both cases, but the more specific training set provided higher accuracy. As noted, only female patients had enough samples to analyze, and the only cohort with at least 10 patients was the female middle-age group. Accuracy for particular status was also higher within this cohort; for example, the subcohort with mild COPD was 99% correctly predicted.

As noted, for sufficient data to predict the GOLD stages from gait parameters, I chose female gender as our cohort group, giving 22 patients who completed the 6MWT and measured FEV1%. GaitTrack implements pre-

	GOLD1	GOLD2	GOLD 3
Number of patients	6	9	7
Mean age (years)	57.3	62.3	59.5
Average height (cm)	154.3 (18.6)	163.5 (3.2)	171.3 (8.2)
Average weight (kg)	78.8 (17.1)	86.4 (30.8)	74.3 (19.9)
Average 6MWD (m)	306.3 (38.1)	208.7 (67.5)	232.8 (65.2)

Table 9.1: Demographic Information of Female Chronic Obstructive Pulmonary Disease Patients

processing that removes all periods for non-walking for gait analysis. It is expected that demographic information can be used to increase the accuracy of the model. here a smaller dataset contains demographic of one gender group (female) and two age groups (50-64 years and 65-80 years). I constructed a support vector machine model with linear kernel using the gait parameters with all 22 patients and another model using only the 12 patients between the ages of 50 and 64 years. The overall classification accuracy shows clustering patients by age increases the accuracy of the model. As Table 9.2 shows, the overall classification accuracy of the younger cohort group is 89.22% compared with 71.73% for the entire group. Thus in the following study, demographic information is involved in model training.

Model using the gait parameters closely predicts GOLD status, whereas the raw walk distance of the patients does not. Shown in Table 9.1, the average walk distance for the GOLD 2 patients was actually lower than the walk distance for the more severe GOLD 3 and even GOLD 4 patients. Therefore, gait parameters appear to more accurately track a patient’s GOLD health status than timed walktests.

	Unified Prediction			Cohort Prediction		
Actual Status	GOLD1	GOLD2	More Severe	GOLD1	GOLD2	More Severe
GOLD1	78.54%	19.51%	1.95%	99.24%	0	0.76%
GOLD2	16.96%	65.18%	17.85%	0	87.63%	12.37%
More Severe	2.96%	23.47%	73.57%	1.23%	11.6%	87.16%

Table 9.2: Support Vector Machine Classification Results with Demographic Cohorts

9.2 Pilot Study II: An Accurate Universal Model Fits All Demographic Cohorts

This part of pilot study is about distinguishing different GOLD levels in COPD group [77]. Our lab recruited twenty-seven pulmonary patients at NorthShore University Health System, under IRB approval start November 2014. All chronic patients going through pulmonary rehabilitation in Respiratory Therapy in Evanston Hospital are now offered the option of participating in our study. Such patients are given our provided smartphones for recording their motion during a standard six-minute walk test during each visit, carried in a fanny pack. The six-minute walk tests are performed on a thirty-meter straight walkway in hospital corridor, with cones at the terminals. The patient walks back and forth on the walkway under the supervision of nurses for six minutes and the distance is recorded. The patient is permitted to stop and rest anytime, although the clock keeps going. Our software does not include this stationary data. All six-minute walk tests follow the ATS guidelines in a completely standard way [78].

The demographic information for each group is shown in Table 9.3. There were two patients, Patients 12 and 24, who did not have pulmonary function tests in the record, so we eliminated these in our dataset. Patient 2 did not perform a six-minute walk test, which means there is no eligible walk data collected for this patient. This leaves twenty-four patients with walking data eligible for analysis with the models.

Table 9.3: Demographic information of each group of patients by GOLD levels. Age, height and weight are in average (minimum - maximum) format. These demographics plus gender are used to adjust the model.

	GOLD 1	GOLD 2	GOLD 3
Patients [Female]	2[1]	13[5]	9[3]
Age (year)	69[65-73]	80[67-95]	72[55-85]
Height (m)	1.68[1.55-1.80]	1.66[1.24-1.83]	1.69[1.55-1.83]
Weight (kg)	109.5[93.0-126.1]	80.3[54.4-112.0]	81.2[45.4-118.4]

9.2.1 Methodology

An SVM model is trained for GOLD stage classification. The model input is the raw three-dimensional acceleration collected by carried smartphones,

and the output is a class of GOLD stage from 1 to 4 for each patient.

Preprocessing

MoveSense collects and processes the raw sensor data to extract qualified walking data. In six-minute walk test, subjects are allowed to rest during the walk test, so the dataset must be cleaned to detect if the patient is walking or stationary mode. We introduce a classification algorithm based on standard deviation analysis [79].

I calculate the magnitude of the raw acceleration data to compress 3D signals into 1D signals, then split the signal of each session into ten-second length samples with five-second sliding window [46]. Rather than a static threshold of standard deviation, the classification algorithm dynamically assigns standard deviation threshold for each session by reading the whole signal thread beforehand in one-second intervals. The algorithm can detect walking or stationary mode for six-minute walk test, see Figure 9.1. When the standard deviation threshold is set, each one-second walking has a binary decision for walking/not-walking. Then for all ten-second samples, if the ratio of walking is larger than 0.7, we select this sample as an eligible sample based on training sets.

Feature Selection

After preprocessing, good walking data for each patient is obtained. I compute input features for training the model by feature selection approach (FSA). Feature selection approach (FSA) is based on empirical knowledge. A primary parameter to measure gait is the cadence, which is the number of strides within a unit of time. We calculated cadence by counting steps in each ten-second interval [47] and dividing by length of sample.

Related studies in motion tracking by wearable devices extract a series of features from raw acceleration data. Besides the cadence and spatio-temporal gait parameters, demographic information must also be considered in model training, just as a spirometer uses it to adjust the raw values from patient respiration. Here I keep the four basic demographic parameters: age and sex, height and weight as factors in the input feature vector. With such input feature vector, universal models were trained which ideally can be applied on

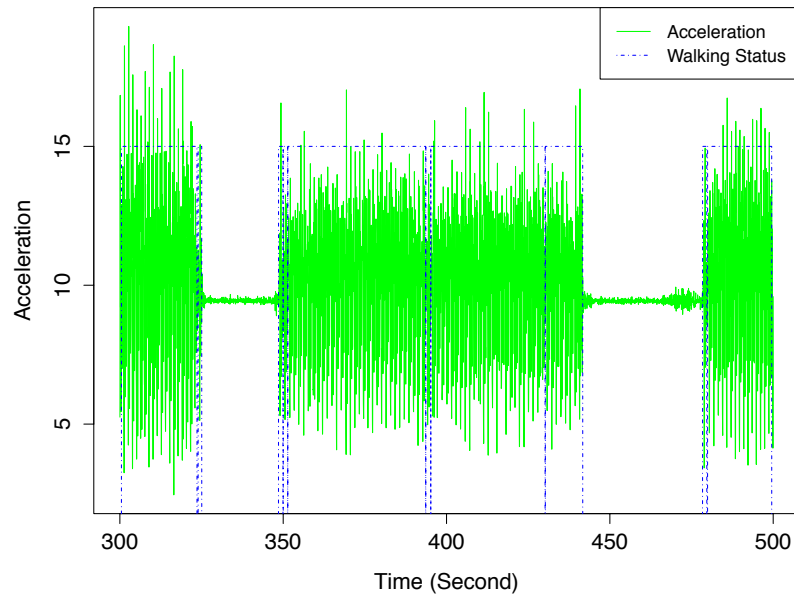


Figure 9.1: Sample result of walking detection algorithm. Green lines represent raw acceleration and blue dash lines represent walking status: higher level means this period is detected as walking and lower level means period is detected as stationary.

general population. Overall, for each ten-second sample, the input feature vector $\vec{x}$ contains thirteen independent features – covering stepping, moving, and demographics.

Model Training

Support vector machines (SVM) are a class of machine learning algorithms, widely used in learning classification models. We apply RBFSVM, the SVM with radial basis function kernel, to train the model for GOLD classification. SVM allows classification by remapping multi-dimensional vector $\vec{x}$ into a higher dimensional space and determining a hyperplane or a set of hyperplanes that separate different classes¹⁹. The hyperplanes optimize the separation between data points in the given space, yielding the classification Equation 9.1. SVM classification requires a training set to find the variables α_i to determine $f(x)$ with a given kernel function and given inputs^[80].

$$f(x) = (i = 1)^N \alpha_i * K(x_i, x) \quad (9.1)$$

From Equation 9.1 we know that different kernel functions affect the overall performance of the trained model. In this study I applied RBFSVM, using the radial basis function kernel to train the SVM model²¹, as Equation 9.2. I set the hyper-parameter $\gamma = 1/D$, based on the number of dimensions in training samples. $\vec{x}_i$ and $\vec{x}_j$ are two arbitrary sample vectors in the training set. In our dataset patients status are labeled as one of the three GOLD stages (GOLD 1, GOLD 2 and GOLD 3). One-against-all strategy is applied to train this three-class classification model^[81]. The tolerance of termination criterion is set as 0.001 and the insensitive-loss variable ϵ is set as 0.1.

$$K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2), \quad \gamma = \frac{1}{D} \quad (9.2)$$

Voting Status with Thresholding

A classification model obtains GOLD stages for each ten-second walking samples. However, each six-minute walk test contains multiple ten-second walking samples, up to 37 if no stopping during the test. For cardiopulmonary patients, the severity will not change within a single walk test. But a walk-

ing sample may be affected by other factors and lead to false classification. Majority voting is robust rule to obtain single decision from multiple samples. It is applied to obtain predicted GOLD stages for each patient. Based on empirical observations, a strict threshold of 85% is applied to accept the voting, which means only when the majority group gets more than 85% of the votes, the decision will be accepted. Otherwise, if none of the groups obtains higher than 85% votes, I mark the status as pending and clinically analyze the risk of this walking session.

9.3 Results

The dataset contains 1204 ten-second walking samples (39 six-minute walk tests) from 24 different patients, including 769 labeled samples from the first 6MWT session of each patient. I trained classification models only using all labeled samples. Ten-fold cross validation is applied to self-validate the model. We then utilize the trained model to predict all unlabeled samples.

9.3.1 Evaluating the Model

Besides the RBFSVM model, I train an SVM with linear kernel as baseline. The results are shown in Table 9.4. The overall accuracy of the RBFSVM model is 98.60% (100% for GOLD 1, 99.28% for GOLD 2 and 98.37% for GOLD 3). While the overall accuracy of the linear SVM model is 86.35% (100% for GOLD 1, 88.31% for GOLD 2 and 82.02% for GOLD 3). The RBFSVM model significantly improves classification accuracy for GOLD 2 and GOLD 3 samples from 82% to 98%.

Table 9.4: Self-validation results for RBFSVM model and linear SVM model. We compare the accuracy of classification, which is all correct classified samples divided by the total number of samples. This is by ten-second interval, not by patient.

	RBF-SVM	Linear SVM
Overall	0.986	0.864
GOLD 1	1.000	1.000
GOLD 2	0.993	0.883
GOLD 3	0.984	0.820

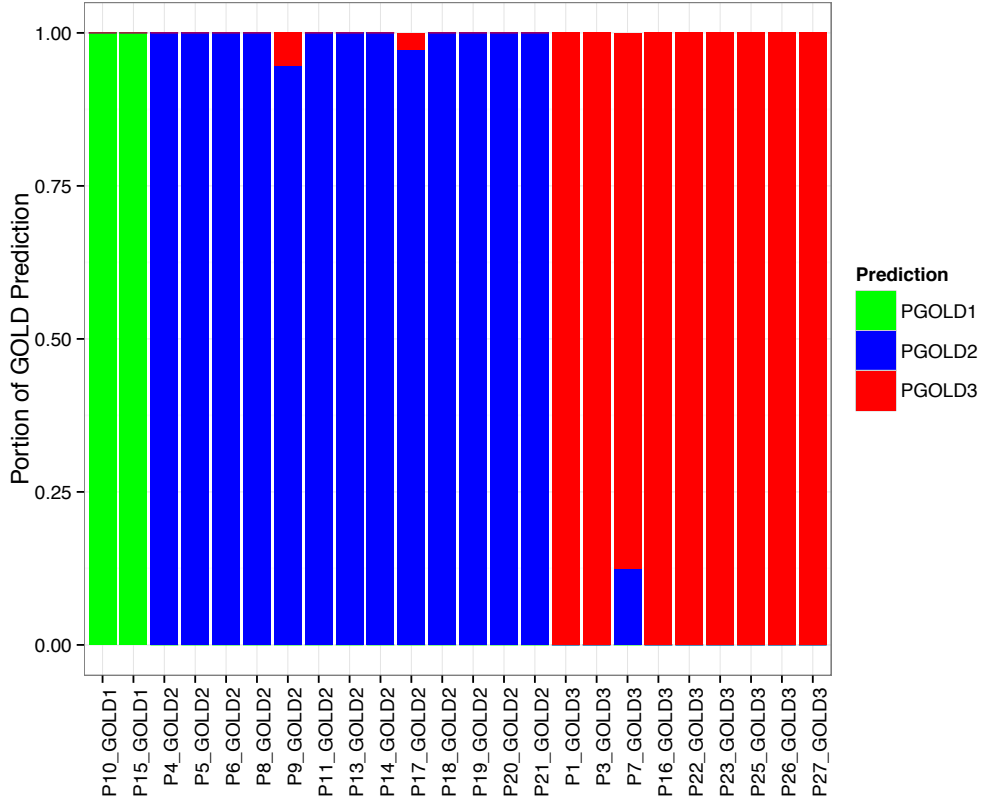


Figure 9.2: Evaluation of GOLD classification for labeled 6MWT sessions. The bar-plot shows the portion of predicted GOLD stages for all labeled walking samples of each patient. Except for Patient 7, 9 and 17, all other patients are 100% correctly predicted within every interval. After voting, all patients are predicted as their actual GOLD stages.

After obtaining the classification results for ten-second samples, majority voting is applied for categorizing each patient into a target GOLD stage. We perfectly categorize all patients into the corresponding GOLD stages summarized from their FEV1% predicted values, shown in Figure 9.2.

9.4 Main Study: Predicting Health Status for Healthy and COPD Subjects

In the main study, I combined the dataset collected from Carle Foundation Hospital and NorthShore University HealthSystem to obtain a full variation of status from healthy to severe COPD. It is to show that a classification

model can be trained to well predict a subject’s status from motion data. Totally sixty-six subjects participate in this part of study, including fifty-five pulmonary patients and eleven healthy subjects. All subjects performed at least a full session of standard six-minute walk test (6MWT), walking back and forth for six minutes on a straight walkway, under the supervision of nurses. Subjects are permitted to stop and rest anytime during the walk test, although the clock keeps going. The distance of 6MWT was recorded for each subject. All six-minute walk tests follow the ATS guidelines[78].

All COPD patients are labeled according to their GOLD levels. Moreover, all healthy subjects are labeled as GOLD 0, indicating that they do not have cardiopulmonary diseases. It is a simplified definition of “healthy”, just in order to see if the model can distinguish COPD from non-COPD subjects. Since the more severe (GOLD 4) patients have significant lack of mobility, measuring their health status is a difficult with motion sensors. So the only focus is on GOLD 0 to GOLD 3 in this study. There are five cardiopulmonary patient missing PFT values so they are eliminated. Two healthy subjects do not have height and weight information and are eliminated as well. Overall, there are fifty-nine subjects data: nine GOLD 0 subjects, twenty-three GOLD 1 subjects, eighteen GOLD 2 subjects and nine GOLD 3 subjects. The demographic information is shown in Table 9.5.

Table 9.5: Demographic information of each group by GOLD levels. Patient [Female] shows the number of patient [the number of female patient] in each group. Age, height and weight are in median [minimum - maximum] format.

	GOLD 0	GOLD 1	GOLD 2	GOLD 3	Overall
Patients [F]	9[6]	23[14]	18[8]	9[3]	59[31]
Age	28[20-60]	65[43-81]	76[61-95]	75[55-85]	68[20-95]
Height (m)	1.72[1.57-1.88]	1.68[1.47-1.88]	1.68[1.24-1.83]	1.68[1.55-1.83]	1.68[1.24-1.88]
Weight (kg)	75.8[51.3-104.3]	85.28[49.4-126.1]	78.93[48.08-117.03]	81.65[45.36-118.39]	80.51[45.36-126.1]

For the main PFT study, four machine learning models for GOLD stage prediction have been trained. The models accept raw three-dimensional acceleration, collected by MoveSense software, as input. After processing, the predictive models output a specific GOLD stage out of the defined four GOLD stages: GOLD 0, GOLD 1, GOLD 2 and GOLD 3.

9.4.1 Data Preparation

As described in previous studies, MoveSense process the raw acceleration and extract qualified walking data. This process is necessary because in standard 6MWT, subjects are allowed to rest anytime during the walk test and we must detect if the patient is steadily walking or stationary. We developed an algorithm for detecting walking based on standard deviation analysis[77].

I calculated the magnitude of the raw acceleration to compress three-dimensional signals into one-dimensional signals. Each walk test session was split into ten-second samples with a five-second sliding window. The walking detection algorithm sets dynamic threshold of standard deviation by reading whole thread of signals in the walk test session beforehand. For each ten-second sample, if the walking portion is larger than 0.7, we keep this sample as a good walking sample, otherwise, we dismiss it.

9.4.2 Feature Selection

The good walking algorithm has been cast on raw acceleration input. Good walking data for each subject is extracted after preprocessing. Feature selection approach (FSA), which is based on empirical knowledge of medicine and kinesiology, is applied to compute the input features[44].

9.4.3 State-of-arts Models

State-of-art learning models are selected to draw the baseline of PFT prediction. In this study I applied linear SVM, RBF-SVM and Decision Tree, which is discussed in Section 8.3.1.

Decision Tree is a typical tree-based learning structure. The “trees” are a set of decision making strategies that can be represented as a tree structure. A decision tree model can be established in a supervised behavior with labeled training set. The Gini index, shown as Equation 9.3, is usually applied for tree model optimization.

$$Gini(D) = 1 - \sum_{i=1}^m p_i^2 \quad (9.3)$$

Recursive partitioning (*rpart*) routines [82] is applied to obtain classifica-

tion tree for GOLD prediction. With default setting the *rpart* program will take the input data and label, and then build the tree model. It also helps plotting the general tree structure so that we can observe details how the tree model is making classification decisions.

9.4.4 Applying DPClass Model

In this study, DPClass[71], an effective but concise discriminative pattern-based classification framework, is applied to obtain a predictive model for GOLD, where a discriminative pattern is a conjunctive clause of several conditions, for example, (Age < 80.5) AND (Height $\geq$ 1.5). DPClass can automatically extract the top discriminative patterns in model training. The generated discriminative patterns, or discriminative rules reflect interactions of all input features. Detail of the technique is descussed in the Section 8.4.

9.4.5 Model Training

The two different gait feature generating approaches: FSA and CEA are combined with four different model training techniques. Totally eight models are trained as comparing the performance. For model interpretation, we stored the tree structure by decision tree and all selected top 30 discriminative patterns by DPClass for deeper analysis.

9.4.6 Results

The dataset contains 3,748 ten-second walking samples, including 485 GOLD 0, 1445 GOLD 1, 1191 GOLD 2 and 766 GOLD 3. Each sample is labeled as the GOLD stage of the corresponding subjects, from GOLD 0 to GOLD 3. We trained different supervised models for the same input with different learning methods we described above. We applied standard ten-fold cross validation for model validation to avoid over-fitting. Then, randomly split the training set into ten parts, and each time select one out of the ten folds as the test set and the other nine folds to train the models. After all ten folds being tested by models trained with other folds, we obtain the prediction

Table 9.6: Model Validation.

FSA	GOLD 0	GOLD 1	GOLD 2	GOLD 3	Overall
Linear SVM	0.9979	0.8042	0.5852	0.6619	0.7326
RBF-SVM	0.9958	0.9972	0.9916	0.9935	0.9938
Decision Tree	1.0000	0.9785	0.9454	0.9256	0.9594
DPClass	0.9977	0.9561	0.9488	0.9566	0.9632
CEA	GOLD 0	GOLD 1	GOLD 2	GOLD 3	Overall
Linear SVM	0.9897	0.8242	0.6784	0.4073	0.7175
RBF-SVM	0.9959	0.9945	0.9882	0.9778	0.9887
Decision Tree	0.9959	0.9356	0.9530	0.8982	0.9337
DPClass	1.0000	0.9605	0.9611	0.9249	0.9607

results of all samples in the manner that none of these samples contributes to model training.

9.4.7 Predictive Models Accurately Detect Different GOLD Stages

The validation results are shown in Table 9.6. The FSA and RBF-SVM combination achieves the highest accuracy, 0.9938. DPClass generally performs the second best either combine with FSA or CEA approach. All eight predictive models can distinguish healthy subjects and cardiopulmonary patients very well. However, distinguishing the slowest patients GOLD 2 and GOLD 3 is the most difficult task for all predictive models. The linear SVM model has very low classification power for distinguish GOLD 2 and GOLD 3.

Each six-minute walk test contains multiple ten-second walking samples. For cardiopulmonary patients, the severity will not change within a single walk test, while other factors can affect a single walking sample so it may lead to false prediction. Majority voting strategy is applied to summarize a single GOLD stage outcome from all walking samples in the same walk test session.

Only the linear SVM model has false prediction on subject level. Two subjects are misclassified: CF015, a GOLD 1 to be predicted as GOLD 2, and CF003, a GOLD 2 to be predicted as GOLD 1. All other predictive models have perfect prediction on subject level. That is to say, all healthy and cardiopulmonary patients can be categorized to the correct cohort group

they belong to.

Tree Model Returns Dominant Features for GOLD Prediction

An advantage of tree structure models is that it outlines the classification strategy based on input features. We trained a universal model with all samples and obtains the classification structure. The misclassification error rate is 0.0267. The number of terminal nodes are 25. The minimum depth of the tree is one while maximum depth of the tree is six. Three demographic features are used to construct the tree: age, height and weight. Five spatio-temporal gait parameters are used to construct the tree: MeanAcc, StdAcc, RMS, PF and Shannon Entropy. All young subjects (age < 49.5) are classified as healthy. For GOLD detection, both demographics and gait features affect the results. The most crucial gait features for classification is the mean of acceleration (MeanAcc), standard deviation of acceleration (StdAcc) and root mean square (RMS). A brief decision structure is shown in Figure 9.3

DPClass Model Returns Top Discriminative Patterns for GOLD Prediction

Mining discriminative patterns is the major novelty of DPClass[71]. Unlike decision trees, which directly make decisions on each node by a specific input feature, DPClass selects the top discriminative patterns and automatically generates decision rules from any combination input features. For this specific model, a sample decision rule could be “(Age<80.5) and (Weight>= 63.28) AND (MeanAcc >= 9.67) AND (StdAcc>= 1.09936) AND (RMS< 1.05517)”, which means if a sample satisfied this rule, the program will return a 1 in the generated feature space . Otherwise, the program will return 0. From the patients datasets, we generated the top 30 rules (discriminative patterns) for GOLD prediction, shown in Table 9.8. The self-evaluation accuracy of DPClass model is 0.9789.

Each discriminative pattern contains threshold for several input features. Unlike the decision tree model, where demographics functions as standalone nodes for decision making, in the DPClass model, demographics participated on every single rule. That is, there is at least one feature out of age, sex, height and weight threshold influencing the binary decision making for every of the top 30 rules. In more detail, age participates in 24/30 of the rules,

Table 9.7: Proportion of Rules Containing Each Feature. We count the number of appearance of a specific feature in all top 30 rules.

Feature	Portion of Rules
Age	24/30
Weight	18/30
Height	12/30
Sex	2/30
Cadence	2/30
MeanAcc	15/30
StdAcc	5/30
AC	3/30
CV	3/30
MCR	1/30
RMS	11/30
PF	4/30

sex participates in 2/30 of the rules, height participates in 12/30 of the rules and weight participates in 18/30 of the rules. Even though cadence is an important gait feature in kinesiology, it only affects two out of thirty rules in the DPClass model for motion correlating to function. Except for Shannon entropy, all spatio-temporal gait parameters contribute to the discriminative rules.

9.4.8 Discussion

Our lab have studied monitoring health status for cardiopulmonary patients with simple smartphones for years and demonstrated motion information extracted from phone sensors have predicting power for GOLD levels. In this part, I show that a universal model, trained with about sixty patients 6MWT data, has perfect accuracy at detecting all status from GOLD 0 all the way to GOLD 3. Meanwhile, tree-based models are applied so that we obtained more comprehensive information about the model mechanisms. This gives us a way to better interpret how the predictive model works and how the input features, containing demographics and gait features, interact with each other.

This work introduces a potential solution to monitor health status for cardiopulmonary patients, by establishing predictive models with human mo-

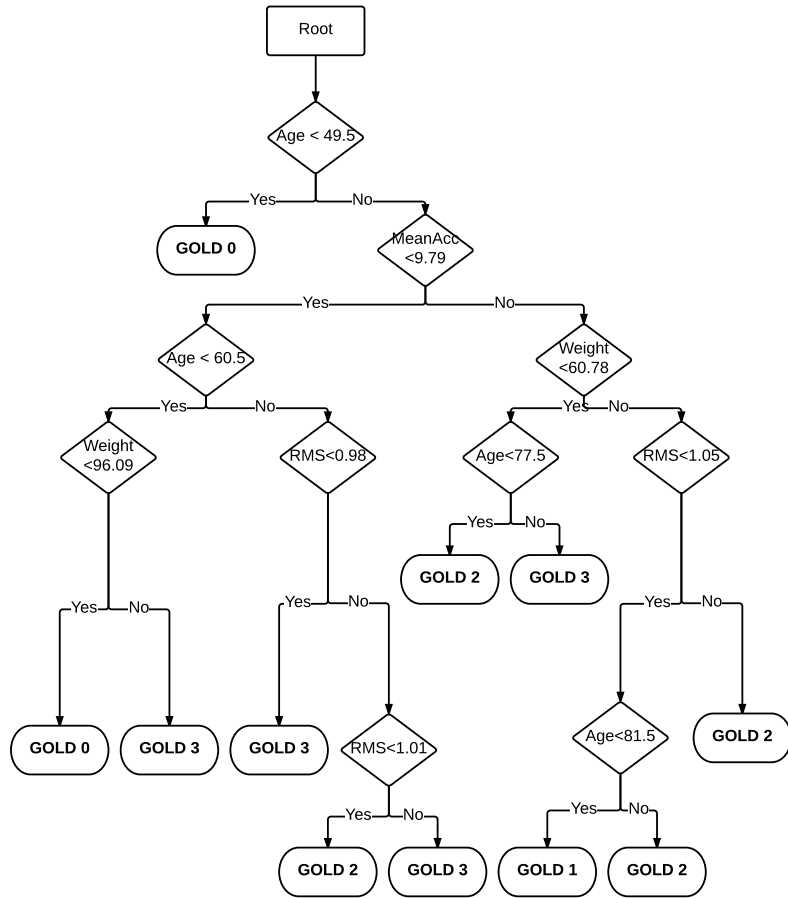


Figure 9.3: Brief Decision Tree Structure. Each Node represents a specific feature dominating decision making on this step. Decision trees making binary decision trees on each node.

Table 9.8: Discriminative patterns in DPClass model.

RANK	DISCRIMINATIVE RULES
1	(Age < 78.5) AND (Weight >= 63.28) AND (MeanAcc >= 9.87816) AND (MeanAcc < 10.1604)
2	(Age < 81.5) AND (Height >= 1.537) AND (Weight >= 59.875) AND (MeanAcc >= 9.76779) AND (MeanAcc >= 9.88939) AND (StdAcc < 2.79557)
3	(Age < 49.5) AND (MeanAcc < 9.86086)
4	(Weight >= 63.28) AND (MeanAcc >= 9.74457) AND (MeanAcc >= 9.86863) AND (RMS < 1.04414) AND (RMS >= 1.0138)
5	(Age < 80.5) AND (Weight >= 63.28) AND (MeanAcc >= 9.67007) AND (StdAcc >= 1.09936) AND (RMS < 1.05517)
6	(Weight >= 67.815) AND (RMS >= 1.01061)
7	(Age >= 51) AND (Age >= 66.5) AND (Weight < 117.71) AND (PF < 2.74542)
8	(Age >= 49.5) AND (Weight >= 61.01) AND (RMS >= 0.993072) AND (RMS >= 1.01945) AND (RMS < 1.05535)
9	(Age < 63.5) AND (Height >= 1.635) AND (Weight < 115.215) AND (MeanAcc < 9.76779)
10	(MeanAcc >= 9.78772) AND (RMS < 1.06968)
11	(Age >= 66.5) AND (Height < 1.7763) AND (AC >= 0.507291)
12	(Age >= 66.5) AND (Height < 1.775) AND (Weight >= 56.475) AND (MeanAcc < 9.88532)
13	(Age >= 66.5) AND (Age < 69.5) AND (Height < 1.7763) AND (AC >= 0.507291)
14	(Age >= 52.5) AND (Height < 1.7513) AND (cadence >= 0.675) AND (StdAcc >= 0.745093) AND (StdAcc < 1.2303) AND (RMS < 1.01391)
15	(Age < 49.5)
16	(Age >= 66.5) AND (Height >= 1.7763) AND (Weight < 89.815)
17	(Age >= 77.5) AND (Weight < 56.475) AND (CV >= 0.0941185)
18	(Age >= 60.5) AND (Weight >= 55.565) AND (Weight >= 78.925) AND (MeanAcc < 9.76984) AND (StdAcc < 1.32393) AND (PF < 1.84692)
19	(Height < 1.74) AND (Height >= 1.395) AND (cadence >= 0.725) AND (RMS < 1.0102) AND (CV < 0.114057) AND (CV >= 0.0844477)
20	(Height < 1.665) AND (Weight < 58.97) AND (AC >= 0.681601)
21	(Age < 64) AND (Weight >= 48.31) AND (MeanAcc < 9.71001)
22	(Age >= 49.5)
23	(Age >= 66.5) AND (Height >= 1.6637) AND (Weight >= 56.475) AND (Weight < 93.215) AND (MeanAcc >= 9.88532)
24	(Age >= 33) AND (Age < 66.5) AND (MeanAcc >= 9.7852) AND (StdAcc >= 1.11517) AND (StdAcc < 2.99304)
25	(Age >= 52.5) AND (Height >= 1.6505) AND (MeanAcc >= 9.72688) AND (MeanAcc < 9.88382)
26	(Age < 66.5) AND (Height >= 1.653) AND (Weight >= 56.475) AND (Weight < 108.865) AND (MeanAcc < 9.88532) AND (RMS < 1.00835)
27	(Weight >= 65.545) AND (Weight < 122.015) AND (Weight >= 94.57) AND (MeanAcc >= 9.98547) AND (RMS >= 1.01831) AND (CV >= 0.187085)
28	(Age < 56) AND (Weight < 82.775)
29	(Age >= 49.5) AND (Sex >= 1.5) AND (MCR >= 0.388519) AND (RMS >= 1.00043) AND (PF >= 1.64725)
30	(Age >= 49.5) AND (Sex >= 1.5) AND (RMS >= 1.00043) AND (PF >= 1.64725)

tion data. We collected 6MWT motion data for sixty-six subjects, including cardiopulmonary patients and healthy subjects. We integrated two tasks: 1) detecting healthy status or unhealthy status and 2) detecting different GOLD stages within unhealthy status; and showed that universal models can predict GOLD status with very high accuracy.

The predictive models are more robust at subject level, after majority voting on sample level predictions. Even the linear SVM model, performing the worst with only 71% accuracy, only have two false predictions at subject level. Moreover, the false predictions are for a GOLD 1 patient and a GOLD 2 patient. As the most confusing categories for a motion based detector, GOLD 1 and GOLD 2 patients have similar level of mobility, which can be reflected by their 6MWDs. A GOLD 1 patient can perform poor in a 6MWT with bad behavior and a GOLD 2 patient can achieve much higher 6MWD after rehab training.

The result shows that RBF-SVM model leads to much better performance than a linear SVM model within the range of GOLD 1 to GOLD 3. In this study, after involving healthy subjects data, RBF-SVM model still obtains close-to-perfect predicting results, even at the sample level.

A major novelty in this part of work is that the tree-based pattern mining technique, DPClass, is applied to provide more insights about model mechanisms. The decision tree model initially demonstrated that demographic information determines how models will select the gait feature thresholds. Additionally, DPClass is designed to not only build an accurate model but also generate more explainable features using discriminative patterns. The model complexity of DPClass is much smaller than RBF-SVM. For example, DPClass only uses 30 dimensions in its final model, while RBF-SVM will use N dimensions (N is the number of patients). It is not surprising that RBF-SVM has a slightly better accuracy due to its much higher model complexity. However, RBF-SVM is more like a blackbox without explanations. With the explicit patterns, DPClass model shows how demographic information interacts with gait features.

Age is the dominant demographic feature in both decision tree model and DPClass model. Young subjects have less probability to have cardiopulmonary diseases, while also moving more steadily. For cardiopulmonary patients, different age ranges indicates different thresholds on gait features for classification. Besides age, both height and weight plays significant roles in

model prediction. This also matches why BMIs are medically important. Height has been adjusted in pulmonary function tests by demographics built into spirometers, but weight has not. So weight should participate more in demographic calibration than height. We observed that both the decision tree model and DPClass model shows weight is more important than height, which matches the medical assumption above. Unlike age, height and weight, sex turns out to be less important, which makes sense because older males and females do not perform differently in steady-pace walking, as reflected on their 6MWD records. The tree structure and discriminative rules provides potential answers to several questions we found in our previous studies, which we will discuss in the next section.

9.4.9 Why Model Interpretation

Our lab have been studying using smartphone sensors to monitor human health since 2011. The results were progressively better, but the models always had some errors that we tried to explain with clinical variation. By increasing the number of subjects and refining the techniques, the universal predictive model is now able to detect status from GOLD 0 to GOLD 3 within the cardiopulmonary status. However, several questions related to the mechanisms of model prediction still remain, causing inaccuracies in the models.

I discuss below how the significant features in the discriminative rules explain these questions. As Table9.7 shows, there are 5 features (out of 13 total) which occur in at least 10 rules. These are the most significant features, which are automatically identified by the pattern model, the next most significant appears in only 5 rules. These features are discussed, showing that in fact each of the questions raised by our previous studies depend on these significant features.

Age dominates healthy and unhealthy status classification. We showed the initial results for detecting healthy and unhealthy status with six subjects, three healthy and three unhealthy[47]. The result shows that healthy and unhealthy subjects are clearly distinguished by their gait speed. However, such correlation is not strong enough to show the causality. Age bias occurs in our dataset and most healthy subjects are younger than pa-

tients. We know that young people generally walk faster than old people. Thus, we always question on whether age is the dominant factor for distinguishing the two groups, not health status.

The tree structure in the decision tree model and the discriminative rules in the DPClass model both demonstrate that age matters. The decision tree will directly classify all subjects into healthy category (GOLD 0) if age is younger than 49.5 years. However, this does not mean that the predictive models cannot detect healthy subjects above age 49.5. Actually, we have one healthy subject in the same age range (>60) as cardiopulmonary patients, and all his ten-second samples are correctly classified into GOLD 0. DPClass also gives more detailed age thresholds for each discriminative rule. These rules are not only for detecting healthy versus unhealthy status but different severity levels. Thus, age is not a factor that causes bias in our dataset but a demographic factor that essentially affects model prediction.

BMI significantly distinguishes demographic cohorts. Our lab had published a pilot study that discusses demographic effects on model prediction for different GOLD stages[48]. With an empirically picked demographic cohort of sex (female) and age (50-64), the predicting accuracy significantly increases. However, this is not the full story for demographic effects. BMI, which is a combined effect of height and weight, is shown to be significant for model predicting, after we merged the demographic information and gait features into the model input, leading to a close-to-perfect prediction. However, then we did not know how BMI affects model prediction. In the current study, we observed that weight dominates decision making for several nodes in the decision tree. In fact, height (12/30) and weight (18/30) both appear frequently in the discriminative rules for the DPClass model. All these evidences show that BMI plays significant role in GOLD prediction.

However sex does not make significant difference for model prediction from the result of decision tree and DPClass. Sex does not dominate any tree nodes and only contributes in two out of thirty discriminative rules. There are two potential reasons: 1) Sex difference does not affect 6MWDs significantly ($p>0.05$). Unlike fatigue tests which test the upper bound of mobility, natural walking of males and females are less distinguishable because the gait indicators do not vary much. 2) Pulmonary function test has already adjusted the predicted FEV1% values by demographics including age, sex, height and race. Unlike age and height, sex or gender is a simple binary indicator, so

easy to adjust the predicted FEV1% against sex across the whole population. Also, since weight is not in the adjustment of PFTs but height is, raw weight value has stronger predictive power, so weight plays a more important role in our model.

MeanAcc more important than Cadence. We carefully studied on natural walking and showed that simple step counting or walk distance is not good enough for cardiopulmonary status prediction. Neither cadence nor 6MWD can be directly used to distinguish GOLD stages [46]. Since cadence, defined the number of strides in a minute, is an essential parameter in gait analysis for general purpose, we still assume that cadence is a distinguishing factor[15]. However, both the decision tree structure and discriminative rules show that cadence is not comparatively significant as the mean acceleration, which is a feature continuously measuring 3D spatial motion of human body. We observed that cadence does not dominate tree nodes but mean acceleration (MeanAcc) does, and cadence participates in only two out of thirty discriminative rules while mean acceleration participates in fifteen out of thirty discriminative rules. Thus, a predictive model for cardiopulmonary status requires more information about motion than simple steps or speed. Note this implies that fitness devices, which only compute steps, cannot predict function, unlike smart phones with appropriate software, which can detect the slowdown and speedup patterns during natural walking..

RMS is more important than StdAcc. Future prediction is critical to prescribing treatments. This involves predicting “risk”, which is the change in status. We have made efforts for risk assessment, which is one step beyond status detection. In particular, we developed a model for predicting transitions in-between two consecutive GOLD stages[43]. In general, instability reflects risks. It means that if the status is more unstable, the higher the risk that this patient will have status change in the future. Both the standard deviation of acceleration (StdAcc) and root mean square of acceleration (RMS) reflect the stability of walking. RMS measures the overall acceleration while StdAcc measures the offside around mean acceleration. Nishiguchi *at el.* claimed that RMS is a better measure for stability for human walking[31]. Here in this study we attested that RMS is more important than StdAcc for model prediction. RMS dominates two tree nodes in decision tree model while StdAcc does not dominate any tree nodes. Meanwhile, we also observed that RMS contributes to eleven of thirty discriminative rules

and StdAcc only contributes to five of them. Thus, for a potential risk assessment for clinical instability, the root mean square of acceleration could be a better indicator than the standard deviation of acceleration.

CHAPTER 10

PREDICTION OF OXYGEN SATURATION

In this chapter, I investigated the feasibility of using gait to predict oxygen saturation, which can be used as an biometric indicator for COPD status. In the previous section, I defined the application on oxygen saturation, after looking at various vital signs to see which vital signs are most correlated with FEV1%. Oxygen saturation becomes the most promising vital sign to predict FEV1% status since it is stable, and convey enough information. This part of study includes building machine learning models to predict oxygen saturation directly from raw gait parameters, and, analysis about common patterns of oxygen desaturation in COPD patients during 6MWT. Strong correlation has been discovered in the oxygen saturation pattern and COPD status. More precisely, it brings new thought about assessing risk by measuring oxygen saturation using smartphones. Ideally, smartphones that are simply carried during daily activities by chronic patients could substitute for pulse oximeters with medical accuracy. Then the most important composite vital sign of oxygen saturation would be immediately and widely available.

10.1 Pilot: Predicting Raw Pulse Oximetry Values

The very first step is to predict raw SpO_2 values, as what is tried for gait speed prediction. The assumption is iff the model have a good chance to match the raw vital sign, it can be potentially used to predict status.

Thus, our lab designed models to predict oxygen saturation values from spatio-temporal gait parameters in the way similar to gait speed[47][48]. As well, MoveSense platform [29][46] is applied to extract the gait parameters from smartphones and build machine learning models to correlate oxygen saturation during both walk tests and free walking.

To understand the values, oxygen saturation is commonly in the range from

100 to 80. A young healthy person with mild exertion will stay steady around 100, while an old unhealthy patient will have organ failure if steadily drops below 80. During a walk test with pulmonary patients[37], older persons will usually start around 97, then slowly drop to their status level of 80, for GOLD I, II and III. As their shortness of breath forces them to slow down in walk speed due to inability of their lungs to keep up with oxygen demand, their saturation will recover a bit, but they will never recover to their initial levels, instead stabilizing at values corresponding to their pulmonary function levels.

Universal models are trained for validation with all the subject's samples merged and the prediction based on training with the whole population. There are two strategies for the validation: 1) standard 10-fold cross validation and 2) cross validation over subjects. The first strategy is to randomly parse the sample set into 10 partitions. Each time I select one sample as the test set with the remaining forming the training set. The second strategy is to select all samples from a specific subject as the test set with all remaining samples as the training. Since the subjects testing the model each time do not contribute to model training, the validation result of the second strategy could be more practically meaningful. I define the absolute error rate as the mean of absolute values of residuals divided by the observed values, as Equation (10.1). Since the variation of oxygen saturation is from lower 80 to 100, the background absolute error rate is larger than 0.05.

$$r = \frac{1}{N} * \sum_{i=1}^N ||pred_i - obs_i|| / obs_i \quad (10.1)$$

FSA Validation. There are two feature selection approaches validated in this study. The first approach includes 20 features generated by the magnitude of acceleration and the second approach includes 50 features generated by the three dimensional values of acceleration. I applies both Gaussian process regression and support vector regression in validation of feature selection approaches, the result is shown in Table 10.1.

The result shows that models trained by support vector regression obtains a lower average absolute error rate ($r=0.0301$) than models trained by Gaussian process regression ($r=0.0329$). In 10-fold cross validation, training with the fifty features generated from the raw values of three-dimensional acceleration

Table 10.1: Validation result of feature selection approaches in absolute error rate. The average r shows that Support vector regression obtains a lower average r than Gaussian process regression. In 10-fold cross validation, selecting 50 features obtains a lower average r than selecting 20 features. In cross validation over subjects, selecting 50 features obtains a higher average r than selecting 20 features.

	FSA ₂₀		FSA ₅₀		Average
	10-Fold	Subject	10-Fold	Subject	
SVR	0.0209	0.0393	0.0173	0.0430	0.0301
GPR	0.0201	0.0379	0.0181	0.0553	0.0329
Average	0.0205	0.0386	0.0177	0.0492	

results in a lower absolute error rate ($r=0.0177$) than training with the twenty features generated from the magnitude of the three-dimensional acceleration ($r=0.0205$). However, in cross validation over subjects, training with the fifty features generated from the raw values of three-dimensional acceleration results in higher absolute error rate ($r=0.0492$) than training with the twenty features generated from the magnitude of the three-dimensional acceleration ($r=0.0386$).

CEA Validation. I validated the component extraction approaches on two signal sets: 1) the magnitude of acceleration and 2) the three-dimensional raw acceleration. The results are shown in Table 10.2.

	CEA _{Mag}		CEA _{3D}		Average
	10-Fold	Subject	10-Fold	Subject	
SVR	0.0221	0.0483	0.0236	0.0456	0.0349
GPR	0.0229	0.0446	0.0242	0.0430	0.0337
Average	0.0225	0.0465	0.0239	0.0443	

Table 10.2: Validation result of component extraction approaches in absolute error rate. The result shows that Gaussian process regression obtains lower average r than support vector regression. In 10-fold cross validation, components extracted from magnitude values obtains lower average r than components extracted from three-dimensional values. In cross validation over subjects, components extracted from three-dimensional values obtains lower average r than components extracted from magnitude values.

The results show that models trained using GPR($r=0.0337$) obtain lower average absolute error rates than models trained by SVR ($r=0.0349$). For 10-fold cross validation, components extracted from the magnitude values ($r=0.0225$) is more accurate than components extracted from the three-

dimensional values ($r=0.0239$). For cross validation over subjects, models trained with the components extracted from the three-dimensional values ($r=0.0443$) is more accurate than models trained with the components extracted from the magnitude values ($r=0.0465$).

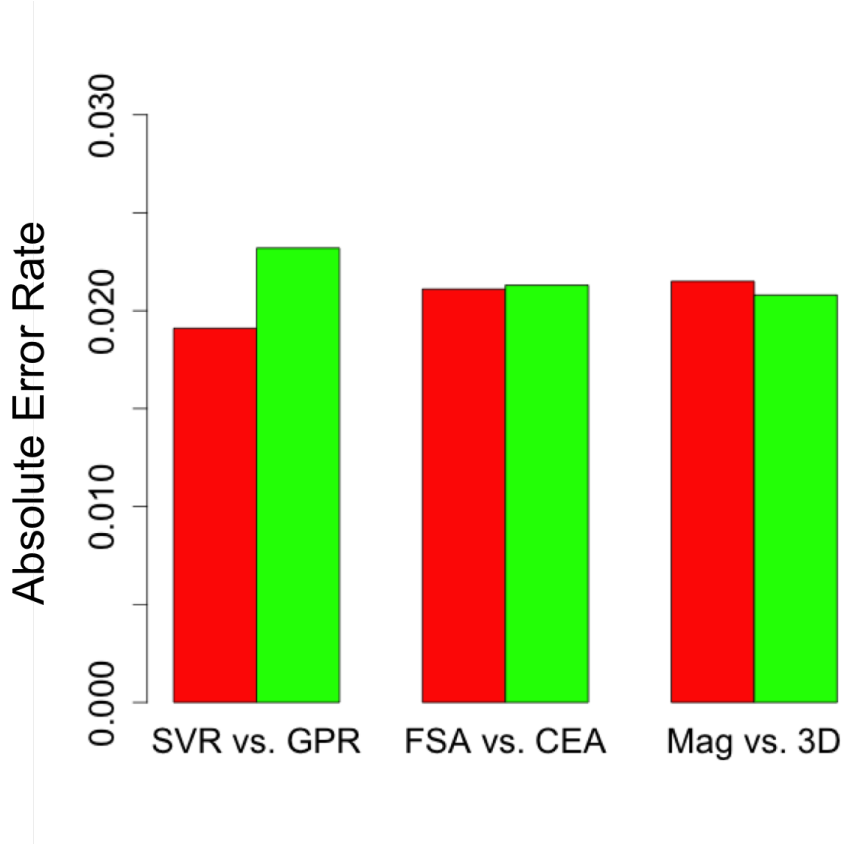


Figure 10.1: Comparison of absolute error rate (r) obtained by 10-fold cross validation. Support vector regression (SVR) versus Gaussian process regression, feature selection approach (FSA) versus component extraction approach (CEA) and magnitude of acceleration versus raw values of acceleration.

In this section I discussed several approaches to predict oxygen saturation from smart phone spatio-temporal motion. Due to the small number of participants (fifteen patients), it is appropriate to assume that all participants' data will contribute to the model training. With a large population, patients in the same cohort will contribute to model training. Thus, the result from standard 10-fold cross validation will be more persuasive in this study. The feature selection approach obtains an average error rate 0.0191 and the component extraction approach obtains an average error rate 0.0232;

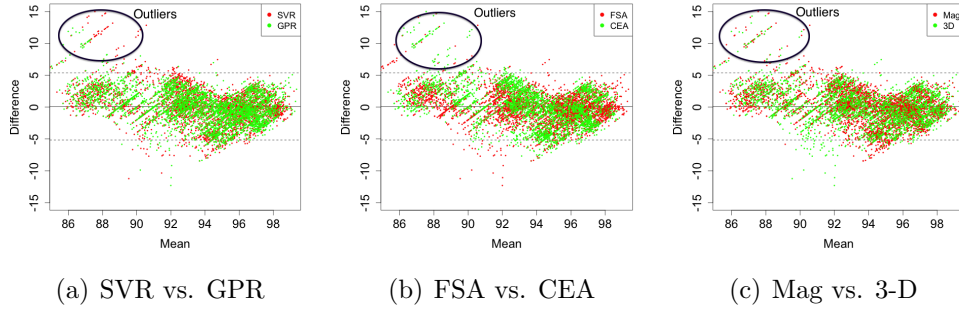


Figure 10.2: Bland-Altman plots for oxygen saturation prediction results. Mean= $(pred + obs)/2$; Difference= $pred - obs$. Confidence interval (CI) is $p < 0.05$. (a) Comparison between two regression models: support vector regression (SVR) and Gaussian process regression (GPR). (b) Comparison between two parameter generating approaches: feature selection approach (FSA) and component extraction approach (CEA). (c) Comparison between two acceleration processing strategies: magnitude of acceleration (Mag) and raw value of three-dimensional acceleration (3D)

Support vector regression obtains an average error rate 0.0211 and Gaussian vector regression obtains an average error rate 0.0213; processing on magnitude of three-dimensional acceleration obtains an average error rate 0.0215 and processing on raw values of three-dimensional acceleration obtains an average error rate 0.0208, shown in Figure 10.1. The result indicates that the alternative approaches on each step of model training limitedly affect the performance. However, there is still an optimal path of training the models: the input should be gait features selected by raw three-dimensional acceleration, and the model should be trained with support vector regression. Figure 10.2 shows the Bland-Altman plots for outlier analysis with our oxygen saturation prediction with the major outliers at the top-left of graphs. Most of these outliers are from the validation results of GOLD III. This is because there is only one subject in this stage. The universal models are biased to GOLD I and GOLD II for this sample set.

Cohort Study. A cohort is a group of individuals within a population with common characteristics[4]. GOLD III was eliminated for containing only one subject. Eight subjects were categorized into cohort GOLD I and six subjects were in cohort GOLD II. Intuitively, more healthy mild COPD patients have less variation while walking than less healthy moderate COPD patients, who must slowdown more often.

	SVR			
	FSA ₂₀	FSA ₅₀	CEA _{Mag}	CEA _{3D}
GOLD 1	0.0079	0.0084	0.0072	0.0077
GOLD 2	0.0149	0.0132	0.0151	0.0161
Average	0.114	0.108	0.0112	0.0119
Universal	0.0209	0.0173	0.0221	0.0236
	GPR			
	FSA ₂₀	FSA ₅₀	CEA _{Mag}	CEA _{3D}
GOLD 1	0.0077	0.0087	0.0072	0.0077
GOLD 2	0.0147	0.0132	0.0152	0.0162
Average	0.0112	0.0110	0.0112	0.0120
Universal	0.0201	0.0181	0.0229	0.0242

Table 10.3: Validation of regression models on cohorts. The result indicates that training on cohorts obtains a lower absolute error rate than training universally.

I trained regression models separately for each cohort using both support vector regression (SVR) and Gaussian process regression (GPR). I also compared different gait parameter computation approaches: 1) computing gait parameters by magnitude values and by three-dimensional values of acceleration, 2) computing gait parameters from feature selection approaches, selecting twenty features (FSA₂₀) and fifty features (FSA₅₀). The models were evaluated by 10-fold cross validation. The validation results are shown in Table 10.3. For all methods, training on cohorts obtains a lower absolute error rate ($r=0.0113$) than training universally ($r=0.0212$). It is also interesting that models on Cohort GOLD I obtains lower absolute error rates ($\bar{r} = 0.0078$) than models on Cohort GOLD II ($\bar{r} = 0.0148$). As discussed later, higher status patients have more constant saturation.

Free Walking SpO₂ Prediction. I evaluate the oxygen saturation prediction model with free-walking data. All free walking samples are processed into the same format as the walk test samples. Fifty gait features were generated for each sample using the feature selection approach (FSA₅₀). The models were trained using the walk test samples and tested using the free-walk samples. I trained the prediction models in two ways: universal training and cohort training. The absolute error rate (r) is 0.0265 for universal trained model. The average absolute error rate ($\bar{r}$) is 0.0201 for cohort trained models, in which the absolute error rate (r) is 0.0195 for Cohort I and the absolute error rate (r) is 0.0204 for Cohort II. The free walking evaluation also demonstrated that model trained on cohorts performs better than on

universal samples, Table 10.4.

	Cohort I	Cohort II
Absolute Error Rate r	0.195	0.0204
Average	0.0201	
Universal	0.0265	

Table 10.4: Free-walking evaluation of oxygen saturation prediction models. The average error rate of cohort models is lower than the universal model.

The Bland-Altman plot for evaluation result of free-walking tests is shown in Figure 10.3. The plot shows that for free walking, most of the outliers are generated from the universal prediction model (CI: $p < 0.05$). The cohort prediction models are effective in oxygen saturation prediction. The prediction model of cohort I does not generate any outliers. The plot indicates that for free walking oxygen saturation prediction, the cohort model performs better than the universal model. Since the models were trained on the walktests, the free-living walking errors are naturally higher, as a harder task with more variation.

10.2 Predicting Transitions in Oxygen Saturation

After the saturation categories being established from the pulse-oximeter, I want to demonstrate that these categories can be predicted from a gait model using phone sensors as the inputs. Our hypothesis is that during 6MWT, walking pattern and SpO_2 patterns are correlated. We previously developed the gait model and used it to predict pulmonary function and walk distance.

As I did for GOLD prediction, two feature generation approaches by four different machine learning techniques are used to form the technical foundation. the classification models are trained to predict three categories (GOLD1, GOLD1T and GOLD2) from walking patterns, applying the feature selection approach and SVM. Then we collect all prediction results for each subject and perform majority voting to obtain the status of this subject. For example, considering all samples for a subject, if most windows of walking are predicted as GOLD1, the status of this subject is predicted as GOLD1. This medical assumption is cardiopulmonary patients are usually walking at the pace related to their status.

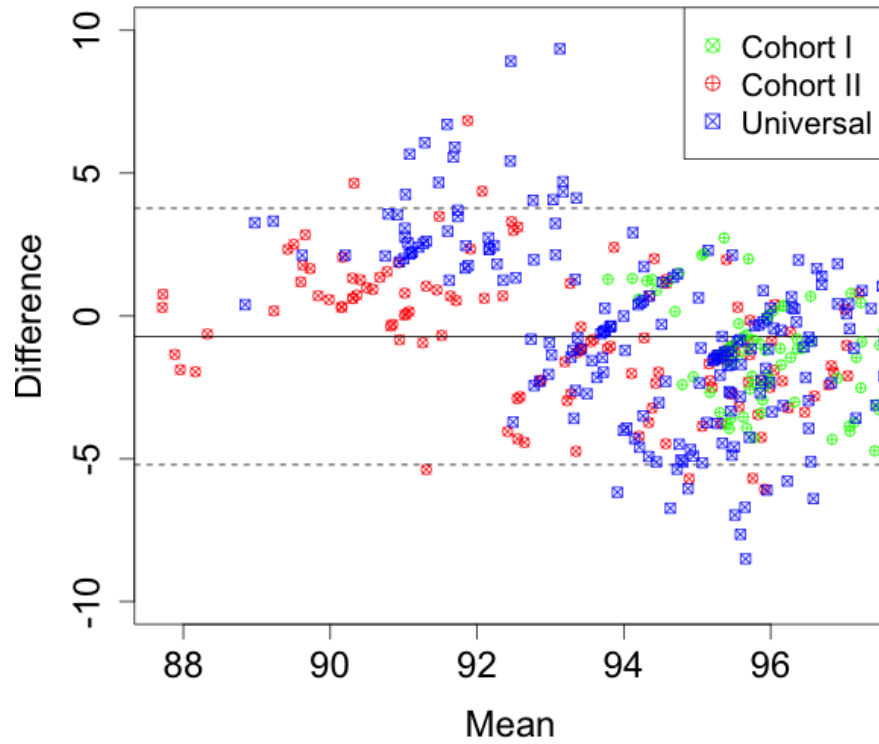


Figure 10.3: Bland-Altman plot for free walking evaluation. $\text{Mean} = (\text{pred} + \text{obs})/2$; $\text{Difference} = \text{pred} - \text{obs}$. Confidence interval (CI) is $p < 0.05$. Most of the outliers are generated by the universal prediction model. Prediction model of Cohort I does not generate any outliers.

10.2.1 Subjects

I conducted experimental measurements on twenty-eight senior patients in Pulmonary Rehabilitation Center at Carle Foundation Hospital in Urbana Illinois. In this study, each subject is required to wear a Nonin Onyx II Model 9560 pulse oximeter during the entire test. Their oxygen saturation values are recorded through Bluetooth connection between the pulse oximeter and smart phone. Thus there are continuous recordings, enabling analysis of the entire curve for oxygen saturation during active walking. This is the first such analysis of the resultant patterns for saturation categories that we are aware of, as previous studies in medical literature considered only desaturation occurrences with ambulatory patients having pulmonary disease.

10.2.2 Sample Selection

Pulmonary function tests (PFTs) are performed with a spirometer in clinical conditions to measure severity level of pulmonary diseases. In obstructive diseases, such as COPD, the ratio of Forced Expiratory Volume in 1 second (FEV1) and Forced Vital Capacity is used as an effective indicator. The FEV1% predicted is a normalized version of FEV1%, which is defined as the ratio of the patients FEV1% divided by the average FEV1% in the population within the same demographic range. The Global Initiative for Chronic Obstructive Lung Disease (GOLD) defines a standard for COPD severity levels based on the predicted FEV1% values. There are four stages representing levels: mild, moderate, severe, very severe.

In this analysis, I focus on ambulatory patients who are the most mobile, as the best candidates for health monitors in daily living. This means the study will focus on GOLD 1 (FEV1% ≥ 80) and GOLD 2 (FEV1% 50-79). So I eliminated the samples with FEV1% less than 50, with two subjects in this GOLD 3 (FEV1% 30-49) category. I also eliminate two subjects who did not complete the walk test and two more where the pulse oximeter did not record. Finally, I eliminated two subjects whose medical conditions caused the pulmonary function for validation to have changed significantly, so that the PFT no longer reflected the function being measured by the 6MWT, due to time delay between tests. The first had a diagnosis of bronchiectasis, a lung condition where function changes within short periods. The second had

change of medication between PFT and 6MWT, with our model correctly showing a marked increase in function. All subjects were permitted to use their inhalers before testing, but she changed from albuterol (for asthma) to tiotropium (for COPD).

Starting with 28 patients tested, this left 20 patients to analyze. For health status, there were 14 GOLD 1 and 6 GOLD 2 status, consisting of 11 female and 9 male. The demographic information used for adjusting the raw values is: age and sex, height and weight, as shown in Table 10.5. I calculated p-values for each demographic feature by Student's t-test between GOLD 1 and GOLD 2 samples. The t-test result shows all p-values are larger than 0.05, which indicates within the 95% confidence interval, so there is no obvious correlation between demographic information and GOLD stages in this sample set. For this specific population subset, demographics does not affect GOLD stages significantly.

	GOLD 1	GOLD 2I
Total Number (Female)	14 (8)	6 (3)
Age (year)	64.5 (43 - 81)	70.5 (54 - 80)
Height (cm)	169.09 (147.32 - 187.96)	170.60 (157.48 - 182.88)
Weight (kg)	83.27 (49.44 - 119.75)	86.25 (54.43 - 124.28)

Table 10.5: Demographic Information of Tested Subjects. The age, height and weight values are in the format of mean (min - max).

10.2.3 Identify oxygen saturation patterns in GOLD status

I have defined the three different oxygen saturation patterns: 1) the SpO_2 value of the subject starts above 93 and keeps stable; 2) the SpO_2 value of the subject starts above 93 and then varies below 93; 3) the SpO_2 value of the subject starts below 93 and keeps stable. Patterns 1) and 3) obviously correspond to GOLD 1 and GOLD 2, whereas 2) is a transition status in-between GOLD 1 and GOLD 2, which we name it GOLD 1T. If we had enough patterns tested with GOLD 3, it is expected that there would have been GOLD 2T and GOLD 3, just as the GOLD 1T and GOLD 2.

By simply observing the dot plot of SpO_2 values over the 6MWT, variation of the transition group could be detected, show in Figure 10.4. To demonstrate validity, t-test is run on oxygen saturation between GOLD 1

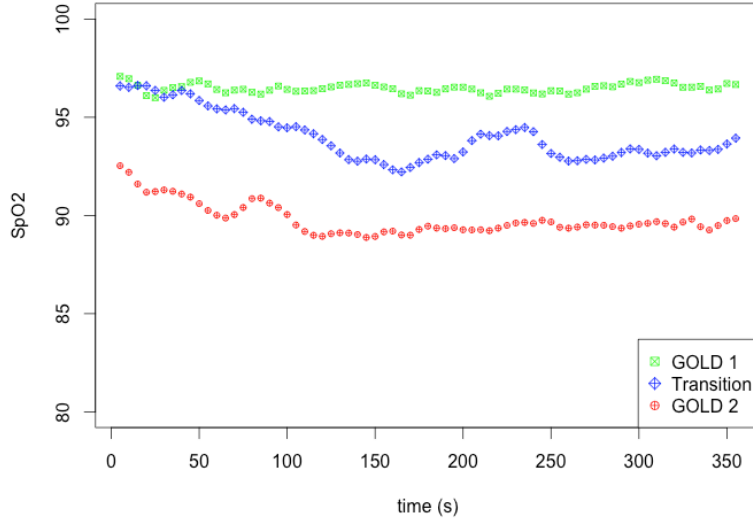


Figure 10.4: Mean oxygen saturation (SpO_2) for GOLD1, GOLD1T and GOLD2 in the 6MWT

and GOLD 1T (p-value < 0.001), plus GOLD 1T to GOLD 2 (p-value < 0.001). We compare the SpO_2 patterns of the three groups by considering the mean SpO_2 of each group over 10-second windows.

I created a novel heuristic model to represent the categorization of the three different groups: GOLD1, GOLD2 and transition group GOLD1T, in between the first two groups. For the GOLD1 group, the SpO_2 will be below the overall mean, with the SpO_2 value of the transition group varying in between. Our heuristic model captures these differences between each group and thus can predict health status, including transitions.

With the heuristic model, I calculated the mean and standard deviation (SD). The result is shown in Figure 10.5, with the subjects displayed in mean order for each category. Mean and SD are independent, as subjects at the boundaries between GOLD and transition demonstrate in the chart. The mean of SpO_2 corresponds to patients' GOLD status, and the SD corresponds to their clinical stability. The 20 subjects naturally partition, corresponding to clinically stable ($\text{SD} < 1.5$) or clinically unstable ($\text{SD} > 1.5$).

The heuristic model was carefully reviewed and proved by pulmonary

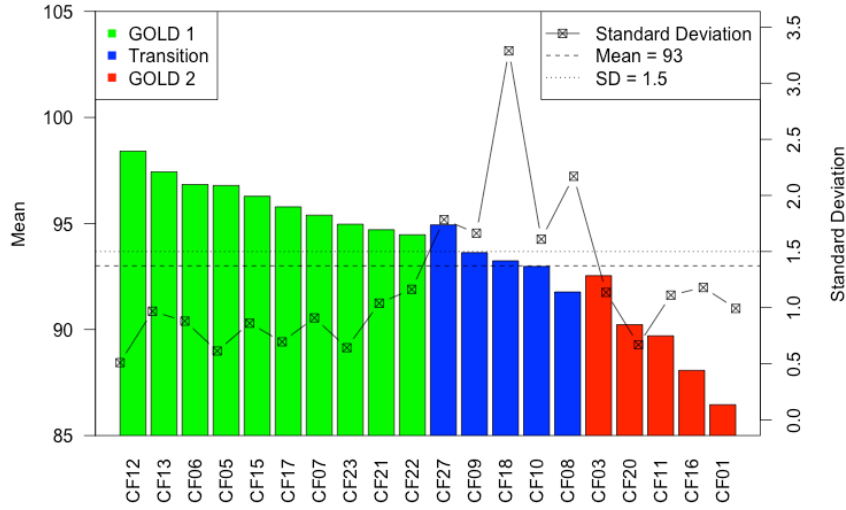


Figure 10.5: Mean and standard deviation (SD) of oxygen saturation for each subject.

physicians, which can be used to label each patient. Then as the six minute walk distance and pulmonary function tests, we yield a regression model to predict actual SpO2 values, as well as a classification to predict oxygen saturation patterns. However, raw SpO2 values are not standard measure for COPD diagnosis unless desaturation occurs, and it will not be recorded in the patient EMR so raw value prediction has restricted meaning. Oxygen saturation patterns are not medically proved yet since no other studies tried to record and analyze all SpO2 values during 6MWT before. Thus qualified physicians are necessary to verify the categorization of oxygen saturation patterns and label each patient. In the preliminary work I trained both regression models for raw SpO2 values and classification models for oxygen saturation patterns and both of the models have potential predicting power.

The categorization can then be used to label the subjects, with the left category corresponding to GOLD1 (10 subjects), the middle category corresponding to the transition group (5 patients) and the right category corresponding to GOLD2 (5 subjects). The subjects in the transition group vary in actual status from GOLD1 on the left to GOLD2 on the right; the first

4 subjects have $FEV1\% > 80$, whereas the last (CF08) has $FEV1\% < 80$. The mechanism of this model for each patients is to check the SD of their SpO_2 values to decide if they are in stable groups ($SD < 1.5$) or transition groups ($SD > 1.5$). If the patient is categorized into the stable group, then I look at the mean of the SpO_2 values to decide if they are in stable groups GOLD1 (Mean > 93) or stable GOLD2 (Mean < 93). This is why the boundary patients have higher mean values than the lowest of the previous category (CF22 and CF08), as their SD places them into GOLD1T and GOLD2, for the boundary patients CF27 and CF03.

10.2.4 Computed Walking Patterns Predict Oxygen Saturation Patterns

Now that the saturation categories are established from the pulse oximeter, I demonstrated that these categories can be predicted from a gait model using phone sensors data as inputs. The hypothesis is that during 6MWT, walking pattern and oxygen saturation pattern are correlated. Here I apply the similar gait model technique I have been used for GOLD prediction and gait speed prediction.

The eight spatio-temporal parameters are selected, together with the demographic information of each subject. I apply the two feature generating approaches: FSA and CEA, plus for different learning techniques: linear SVM, RBF SVM, decision tree and DPClass. The result is shown in Table 10.6.

Oxygen saturation is perhaps the most important vital sign in clinical use. It is closely correlated with clinical outcomes for health status, thus is diagnostic as warning sign for many conditions. While motion sensors are not directly measuring heart rate or lung function, motion is correlated to these for cardiopulmonary patients. Predicting saturation patterns from carried phones is a breakthrough, showing that gait speed (a proposed 6th vital sign) can predict oxygen saturation (a proposed 5th vital sign).

The results reveal a new aspect of oxygen saturation, from analyzing curves of continuous measures. The saturation for the 20 patients clusters naturally into 3 categories, based on the mean when clinically stable and standard deviation when clinically unstable, as shown in Figure 10.5. The saturation

Table 10.6: Oxygen Saturation Pattern Model Validation. I validate eight different predictive models by 10-fold cross validation. The first four rows represents the FSA results and the last four rows represent the CEA results. Each column represents the prediction accuracy for samples in the corresponding category. The accuracy is calculated by the portion of number of correctly predicted samples in the total number of samples in this category.

FSA	GOLD 1	GOLD 1T	GOLD 2	Overall
Linear SVM	0.9240	0.7063	0.9672	0.8787
RBF-SVM	0.9951	0.9901	0.9964	0.9941
Decision Tree	0.9951	1	1	0.9975
DPClass	1	1	1	1
CEA	GOLD 1	GOLD 1T	GOLD 2	Overall
Linear SVM	0.9547	0.8845	0.9380	0.9331
RBF-SVM	0.9984	0.9868	1	0.9958
Decision Tree	1	0.9868	0.9891	0.9941
DPClass	0.9982	0.9963	0.9960	0.9974

categories corresponded to a higher status (GOLD 1) with steady values above 93 and a lower status (GOLD 2) with steady values below 93, plus a Transition category status with varying values around 93, between these two steady levels. We label these categories accordingly as 1-1T-2, corresponding to GOLD levels with Transition.

There were another 2 patients whose values would fit into categories 2T and 3, but I eliminated these for modeling, due to small sample size with GOLD 3 status. More subjects would enable 5-way categories, as with athletes running 10 minutes on treadmills. When there are more categories, the thresholds for mean and standard deviation may be different.

The results also demonstrate that our gait model can accurately predict oxygen saturation. That is, motion can predict saturation for broad categories. This is a novel computational model from a novel clinical observation. The gait model had perfect prediction of saturation patterns, after voting was applied to classify by majority of sampled windows. The voting is necessary since the patients change patterns while walking, for example, going slower when they become short of breath before returning to their usual pace at their current status.

The perfect prediction of saturation category leads to a broader impact for adult medicine. If transition patterns can be separated cleanly from status

patterns, then change in status can be automatically detected. Thus, passive monitors from carried phones can continuously detect when a patient is within a transition category between status levels. Such transition detection can provide an accurate monitor, alerting that status has changed so that treatments may need to change.

Mobile phones can measure health status at stable levels and during transition stages. Passive monitors from phone sensors can predict clinical stability from oxygen saturation. Predicting transitions in saturation categories may become a major advance in clinical monitoring for population health.

CHAPTER 11

GAIT SPEED PREDICTION

Predicting gait speed is necessary. Whenever a new technique is released, the first step is to prove to the customer to believe what it does is correct. In medical literature, gait speed is proved to be important for diagnosis on cardio-pulmonary patients. Physicians will believe the predictive models only in the situation that we could confirm that it can predict speed correctly.

The raw motion obtained from the phone system must next be converted into gait speed [83]. In order to restrict error rate, the phone system must identify and correct for the position on the body, determine if the subject is walking, identifying steps, and ascertain speed. This work provides initial solution to identify and correct acceleration for body position, conduct activity recognition to identify good walk, and attain accurate step count and is not further improved in this work [47]. After proper pre-processing the analysis focuses on improving the accuracy of the gait speed estimation. The distance values are then scaled by age, sex, height and weight producing a measure found to closely correlate with status in pulmonary patients. We adapted the system to conduct 6MWTs. The program records the subject's heart rate and blood oxygen level through a Bluetooth connected pulse oximeter. Motion sensor data is recorded at 60Hz. After recording, all data are securely sent to a remote server for further analysis.

Table 11.1: Subject Demographics

	Healthy		Pulmonary	
	M	F	M	F
18-39	2	5	-	-
40-49	-	2	-	-
50-59	-	-	2	4
60-69	1	-	5	7
70-79	-	-	5	3

11.1 Predicting Speed for Six Minute Walk Test

Data are selected from thirty-eight subjects at the Calre Foundation Hospital Pulmonary Rehabilitation Center in Urbana, Illinois including twenty-eight patients diagnosed with a pulmonary disease requiring a pulmonary function test and ten subjects without a diagnosed pulmonary disease as seen in Table 11.1.

For each subject, the 6MWT data is divided into individual laps by analyzing the rotational matrix taken from the magnetometer and identifying when we see a change of π radians in direction. I verify the laps against manual timing taken with a stop watch during the 6MWT. To ensure correctness of labeling, we manually checked all turning labels so that the labeling error won't affect speed estimation.

The proposed technical approaches has been applied in speed prediction. Due to the fact that the predictive model must return raw speed values, the regression models are selected. Thus the technical foundation in speed prediction is slightly different from it from the other two medical applications, which are both classification problems.

For gait feature selection step, FSA and CEA are applied, while for machine learning step, Gaussian process regression (GPR), support vector regression (SVR) and regression tree are applied. For SVR, two different kernels are selected: the linear kernel and the radial basis kernel, as shown in section 8.3.2.

11.1.1 Speed Prediction Accuracy

Since sample size is small, demographic cohorts are not split out, but merged with raw gait parameters as input features. Predicting speed on natural walking is more difficult than speed prediction at a fixed speed [19]. When the patient's speed is fixed, prediction models are simply trained with all input vectors for the set speed. In a natural walk, the subject's speed constantly varies. Thus, instantaneous speed is constantly in therefore train our models using supervised learning with the input vector being derived from ten second windows and the target being the average lap time.

The results of predicting speed per lap are given in Table 11.2. Overall, the prediction error is much lower when trained using FSA than CEA. The

best performing approach is the combination of FSA and RBF-SVR.

Table 11.2: Error Rates for Speed Prediction for both FSA and CEA input features

	FSA			
	GPR	Linear SVR	RBF SVR	RT
Error	5.61%	3.23%	2.58%	18.60%
RMS	0.072	0.046	0.036	0.224
	CEA			
	GPR	Linear SVR	RBF SVR	RT
Error	9.20%	8.68%	5.94%	11.28%
RMS	0.125	0.124	0.084	0.138

11.1.2 6MWT Distance Prediction Accuracy

Two strategies can be applied to predict total walking distance for the 6MWT. The general approach is to split the walking data into 10 second sample windows and use our speed prediction model to predict window speed. Then multiply the window time by the speed to get total distance. We call this the free walking method as the method is not constrained by laps. The second method is to identify the turns during the 6MWT and multiply the time of each lap by the average speed predicted by the windows in the lap. We call this the lap based method. We estimate the total distance based on both the free walking and lap methods described above. The results are shown in Figure 11.1. As expected, the lap method produces better results than the free method since the training was based on the lap speed. For the full model, the lap method error was 3.78% for pulmonary patients and 1.82% for healthy while the free walking method produced errors of 15.22% for pulmonary and 17.24% for healthy subjects.

11.1.3 Pulmonary Patient Free Walk

The twenty-eight pulmonary patients were asked if they were willing to walk around the rehabilitation clinic after conducting the 6MWT. Twenty-four patients were willing to freely walk while carrying the pack around a 120 foot (36.6 meter) oval. The number of laps walked depended on ability with a range of one lap to five laps. We then used our universal free walk distance

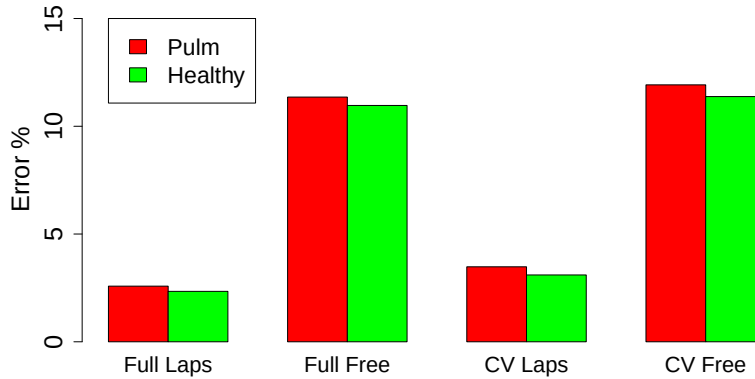


Figure 11.1: Error in Distance Estimation for 6MWT for the Full and Cross Validation Models in Healthy and Pulmonary Patients.

model to predict the distance and compared to the actual distance completed by counting the number down from twenty-four to twenty patients. With these patients, who free walked from 2 to 5 oval laps, the error was 11.2%. Thus the free walk prediction is less than the target error, even though the model is trained on controlled walk tests but used on uncontrolled free walks. The Bland-Altman plot shows that our measurement on average estimates low by roughly 7.7 meters. Thus we underestimate the walked distance when patients walk freely on an oval course due to lack of turns.

11.2 Healthy Controls with Fitness Devices

To better evaluate the performance of MoveSense in prediction mobility, we compare the distance estimation from MoveSense against common fitness and health devices. To directly compare, we conduct testing on five healthy subjects with one female and four males from the ages of 21 to 60 years old. Due to the cumbersome nature of carrying six devices, the IRB would not allow us to test on pulmonary patients. We have our subjects carry a Fitbit Flex and Jawbone Up24 on their wrist, an Omron HJ-720ITC pedometer on their belt and an iPhone 5s running Moves (a commercial app from ProtoGeo now a FaceBook subsidiary), a Motorola Droid Mini running MoveSense and an LG Optimus Zone running MoveSense in a pack at the L3 position.

Subjects walked between two minutes simulating eating a meal at a table. Finally, the subjects walked through the 65 meters of hallways two more times and ascended and descended a set of stairs. Each fitness device was calibrated with the subject’s personal stride length and demographic information. We recorded the distance reported from each device and the distance from our previously trained universal free distance model for both phones running MoveSense. These were then compared to the known measured distances.

The results of the experiment are shown in Figures 11.2. We find that the Fitbit Flex gives average readings 400% error. The JawBone Up24, Omron pedometer, and Moves app all have errors on the order of 50%. Our models running on the Mini and Zone phones attain average errors of 8.3% and 7.6% respectively. We note that lower distances attain a smaller error with the result being skewed to the walk test with the longest length. Overall the MoveSense system estimates walk distance at nearly an order of magnitude better than popular fitness devices, shown in Figure 11.3.

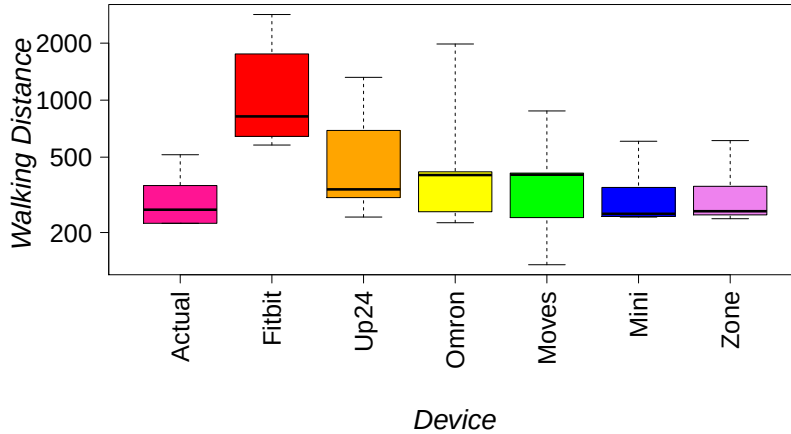


Figure 11.2: Box Plot of Actual versus Predicted Distance for All Devices.

11.3 Rethink the Correlation between Speed and Status

Gait speed, or the six-minute walk distance (6MWD), has been used to indicate COPD severity. It is statistically approved with large sample size [34].

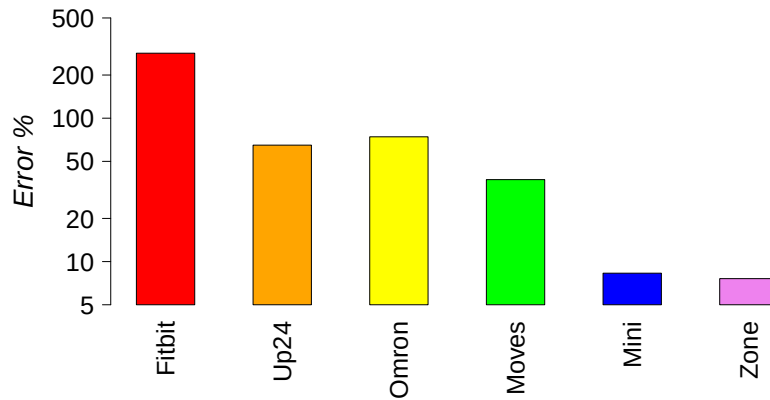


Figure 11.3: Comparison of mobility prediction between MoveSense and popular fitness devices.

But when a physician wants to apply the result for individual status diagnosis, he needs to rethink if the 6MWD actually reflects the status for this specific patient.

My observation on the smaller set of COPD patients indicates that 6MWD may not be that significant for individual diagnosis due to the distance cannot significantly distinguish GOLD 2 and GOLD 3. We can clearly see that in the Bland-altman plot results, the 6MWD of GOLD 2 and GOLD 3 subjects are mixed up, in Figure 11.4.

That is to say, for individual diagnosis, theoretical conclusion can only be used after attested by observations. Instead, both pulmonary doctors and pulmonary nurses confirmed that the gait patterns are different among different severity groups from their expertise by watching hundreds of pulmonary patients walk. Their information input is much more than a single number of 6MWD but the observed gait pattern. A predictive model should consider as much as the patterns and simulate the actual diagnosis process of physicians to achieve the goal of perfect prediction. This is why I need to predict status directly, rather than just predict speed.

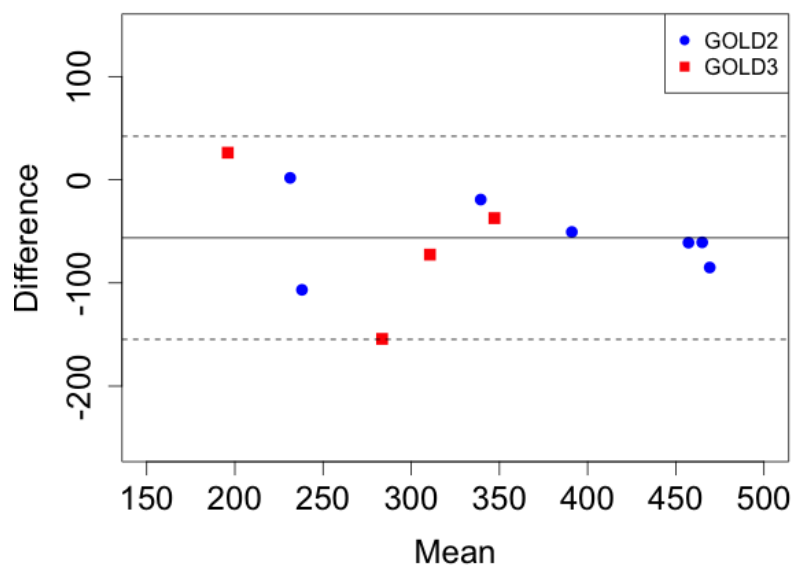


Figure 11.4: Figure. Bland-Altman plot of 6MWD comparison. Blue dots represent 6MWDs GOLD 2 patient sessions. Red squares represent 6MWDs of GOLD 3 patient sessions. GOLD severity levels are not well separated by 6MWD.

CHAPTER 12

LIMITATIONS

All previous studies provide a possible solution for using smart phones to monitor Free-Living Health, specifically for pulmonary patients. Generally to complete the monitoring system and process raw sensor data, there are four consecutive steps: Good walking detection, feature extraction, model training and medical application. Within each step, the previous work only covers a subpart of potential techniques, which means, there are ways to improve the overall performance by properly introducing alternative techniques in each step.

12.1 Device and Sampling

This study applies one single smartphone to monitor health, which means we need to face to the nature weakness and limitation of the hardware. Most medical acceleration pick 100Hz (Actigraph, Dynaport, etc.) as their sampling frequency. Yang et al. [53] briefly compares the efficiency of different sampling rates for monitoring motion. MoveSense picks 60Hz as a compromise between reliable phone sensor sampling and rich enough information. The 60Hz is in the middle of Yang's outcome of efficient sampling frequency, 10Hz to 100Hz, depending on computation of different features and it is sufficient for all our applications since old patients move very slowly and their motions vary significantly among different COPD status. However, applications in further studies may require higher sampling frequency, and it is expected the cheap devices could easily hit higher steady sampling frequency rate in the next years. Also note that currently a Fitbit brace-band maintain a 1Hz sampling rate, compromising with the storage and battery issue, which is expected to be enhanced quickly. Thus, we can expect that even cheap braceband will give us sampling rate equal or above 10Hz, which makes the

device sustainable for monitoring health status.

12.2 Good Walking Detection

Good walking detection requires activity recognition from 3-dimensional acceleration data. Lockhart et al. discuss on several limitations with activity recognition methodology and data sets [Lockhart2014]. Even though there has been lots of study about activity recognition and it become mature as a field, most of the current AR studies are not extensible to a ubiquitous level. That is to say, most of the current work has sufficient limitations on their experimental methodology and often make unrealistic assumptions. Most activity recognition strongly biased on their own training set so the trained models become inaccurate if they are applied on another dataset directly.

In our study, to ensure higher recognition rate we applied activity recognition by training the model with our own dataset and we manually checked the result even after the automation to make sure the processed data are clean enough for following analysis. This procedure may not work when a large dataset streams in.

Besides, good walking detection is similar to, but not the same as activity recognition. It is a binary classification of walking and non-walking status, instead of multi-way classification of all considered activates. Such character simplifies the problem, which lead to a higher chance to obtain high quality recognition and extract good walking sets from daily free walking. In this specific situation, we expect low false positive rate, which means the detected good walking must be real good walking. And we are lenient on missing detection of good walking. Thus, the good walking detection model should obtain an comparatively high precision and the recall can be sacrificed to achieve higher precision at some level.

Another important issue is the adjustment of phone positions. Sensor position recognition has been discussed in related works [33]. However, similarly as activity recognition, the technique is hard to be applied to ubiquitous situation due to restrictions of methodologies and experimental settings. As for old patients, our investigation shows not many of them have pockets on their pants or around the lower back (L3) position. An ideal solution is after detecting common phone carrying positions, such as coat pockets, backpack,

purse, etc., there is a way to transform accelerations from these positions to acceleration when the phone is carried at the L3 position.

12.3 Feature Extraction

Currently, one phone can only collect data from one position at a time. However, most Kenisiology study, such as sports analytics, fall detection require tracking motions of multiple body position. There are not many options to deal with one dimensional acceleration signal. Previously I spent lots of time to explore good algorithms to count steps accurately. After that, I was focusing on computing the stride length directly from the raw signal. The theoretical ways are not suitable for the practical usage since the phone is bouncing while people walking. Thus, signal processing techniques are applied to retrieve enough features to train the model. However, since it is still not a typical signal processing application, there are few literature talking about the efficiency of each potential features. The only ones appeared in publications that relevant to human motion were selected and compared.

In previous studies we tried several strategies to extract features from raw spatio-temporal acceleration collected by phone sensors, including heuristic feature selection process and automatic component extraction process. For feature selection we basically calculate gait features which are attested to be effective by the literatures in relevant researches [52] [50] [31]. However, there are more gait features described other than the ones we applied, which means after involving more features, we can potentially increase the quality of the model. For component extraction, we tried principal component analysis and optimized the number of principal components we need to extract. However, the overall performance of component extraction is not better than feature selection, which indicates PCA is not an optimal solution for extracting features.

Thus, there could be better set of features or better solution if both the Kenisiology and signal processing domains promote techniques to monitor human motion. It is expected to happen soon since more and more data are collected and stored with personal wearable devices and the dataset will be attractive enough for expertise to focus more on.

12.4 Model Training

We trained predictive models for the applications in our previous studies, with a series of state-of-art techniques, including support vector machine, Gaussian process regression, artificial neural networks. We compared different techniques and select the one with better performance for a further study. However, for completing a Ph.D. dissertation, a direct comparison of all potential techniques should be included in every target applications. As the size of dataset increases, the performance of different techniques could vary. It is necessary to complete the comparison to capture the variation and select optimized approach for each application.

As previously mentioned, our models are based on sensor data from constrained walking rather than unconstrained walking. However, in real word application, most walking are unconstrained, which could be different from a constrained walking data from standard walk test.

12.5 Medical Applications

The whole project requires medical validation when we are willing to leverage the usage of applications. Since the topic is unique back to the time we started in 2011, we have to validate our results with acceptable medical measurements, such as 6MWT and the gait speed. However, gait speed, or the walk distance does not represents high prediction power of health status, just as heart rate, which varies a lot in both population view and chronic view. That is to say, for different subject in population, gait speed varies a lot and for a specific subject, his gait speed could also vary a lot chronically. Thus, our gait speed model can easily be validated but hard to be applied to indicate risk of health status variation. The pulmonary function and oxygen saturation related prediction have higher potential power to indicate health status variation. But the validation of its effectiveness requires more interaction with physician diagnosis and other medical indicators. We need real doctors to review health status of every patient in our experiment so the label status and risk of this subject is reliable for model training and model validation.

CHAPTER 13

POTENTIAL EXTENSIONS

The object of the project is to establish a real-time system to monitor human health. Besides using phone sensor data predicting health status, the real-time system also requires solutions to other issues. One prime issue is to collect good walking data in free-living situation with smartphones. The MoveSense App. for Android smartphones is created to collect phone sensor data and transmit the data back to the server. To complete the functionality of MoveSense and make it a real health monitoring phone App., It is necessary to integrate all predictive models into the software, which accepts and analyzes real-time data stream.

Also, the general philosophy of this study could be applied in related health monitoring problems. MoveSense applies cascading filters to divide the complex problem into several simpler problems. Currently, it applies filters to guarantee the reliability of outcomes for each intermediate step, as well as involves assumptions for each step to support the model functionality. These assumptions should also be attested to prove the feasibility of the system scientifically. Each of the issues could be potential extension topics.

13.1 Data Quality Improvement

In the limitation chapter I discussed about the limitation of using a simple smartphone to track health. As the market of new wearable devices growing, there must be more data from these devices rather than phones. And when technique improves, the computational power, battery and storage of these new devices will be capable to handle medical quality daily monitoring. New devices bring new thoughts to this type of research.

It could be easier to simulate Kinesiological motion study in patient's daily life with multiple-sensor wearable devices. Multiple sensors will generate

multi-dimensional data, which contains much richer information than one phone. Also the wearable devices will be attached tightly on human body so less noise will be generated, such as bouncing, rotating, etc. In the phone study, low pass filters are applied to reduce sampling noise, forcing us to lower the eligible sampling frequency, 60 Hz to 20 Hz. There are potential information loss in this data cleaning step. So if we can keep the raw sampling frequency in a cleaned dataset, it will help detecting more narrowed variation of different status. When it comes to other related topics, like fall study, these hardware level improvement could be very significant. To control sampling quality, this study applies multiple filters in a cascading manner. The desired good walking data should be steady walking for a comparatively long time (> 2 minutes) of old patients.

On the other hand, multi-dimensional sensor data increases the input complexity. How to extract sufficient features from raw signals will be another challenge. In this dissertation I proposed several methods to reduce feature dimension and select sufficient features for model training. The typical signal processing approaches include Fourier's transform, PCA, etc. Besides, DP-Class, as a novel pattern mining approach, has also been introduced to select sufficient feature patterns. Currently around ten features are generated for the input and most of them are significant for model training. When the input feature space increases to hundreds or even thousands, these feature reduction approaches will be more important for achieving better predictive models.

13.2 Length of Good Walking for Health Status Prediction

In the cascading filters, each step needs scientific proof of the optimization. In general, there are two ways to determine whether the parameters we pick are optimal or not: theory-driven or data-driven. It is feasible to trust knowledge from medical literature. However, for issues that are barely supported by any medical literature, data-driven approaches are sufficient alternatives to pass optimization process for intermediate steps.

One challenge issue of the cascading filters is to explore that as a sample, how long of the good walking is enough? In medical literature, there are

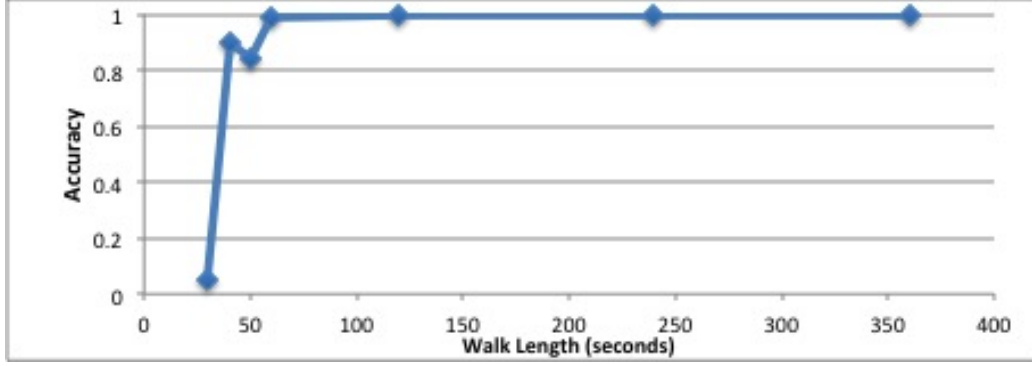


Figure 13.1: Classification Accuracy with Different Walk Length in Training. The accuracy in ten-fold cross validation reaches 100% after two-minute (120-second) length.

two popular types of walktests: 6-minute walk test and 4-meter walk test. But neither of them tells us about how long is the optimal walk length for monitoring health status. Thus, we need to design data-driven experiments to find an optimal solution, at least for the current dataset.

To simulate such experiments, we need to chop the 6MWT dataset into different length and assume that we only obtain limited length of good walk for each patient. An initiation of walking length analysis is done with the NorthShore patients. The result shows that when the length is 2 minutes obtains similar accuracy as when the length is 6 minutes, which is a full 6MWT, shown in Figure 13.1. We will operate a similar analysis on the full dataset to capture a minimum length of good walking. It is necessary for the real-time system do decide when to collect sufficient good walking data from the massive daily activity sensor data.

13.3 Good Walking Detection

Detecting good walking is a significant issue for the real-time monitoring system[24]. Activity recognition (AR)[84] and position recognition (PR) has been studied for years and currently there are commercial products being released. The Wireless Sensor Data Mining Group at Fordham University published a phone application, called Actitracker. They also published their activity dataset for helping research in this area.

With well-labeled training set, the models obtain accuracy over 70% for a

series of pre-selected activities[85]. As discussed in limitation section, such activity recognizer cannot be applied to patient directly. Thus, making use of the public dataset in this area to build a recognition model specific for old patients is another important issue.

13.3.1 Activity Recognition

Currently activity recognition research is dedicated to classify each type of activities in a pool, which mostly are all known activities in the dataset. Thus, with well-labeled data, a supervised learning model will usually achieve high predicting accuracy[33],[58]. However, we should always keep in mind that in real-life situation there must be activities that are unknown to the model. Also the population variation is another important factor. This dissertation study is aiming to a population of old and sick patients, but some existing models are trained by data from young and healthy students. The result shows that the models can be much less accurate even if we train the model with our own dataset. So we should be cautious if we want to apply any of the existing models for activity recognition. They may not work as well as they claimed.

Another important concern is the window size. Different studies choose different walk length for train and test their models. Most of the studies are just pick up the most convenient window size to chop the samples. A systematic study is necessary to compare the model performance under different window size. In real-life situation, we will compromise on a lot of other issues, such as computational power, memory and battery. An optimal window size for detecting activities is a prerequisite for applying the model.

13.3.2 Position Recognition and Adjustment

Park *et al.* did one more step than just recognizing different gestures or activities, recognizing the device location. The methodology is pretty much simple, labeling different position and training supervised SVM model[33]. The study yields to a high accuracy for position recognition but it is far away from sufficient enough for real-life situation. The acceleration patterns of different positions are also varied by the type of activity, demographic and

other factors. A young male always carry their phone in the pants pocket during normal life, but that is hardly happened for a old female since they don't usually wear pants with pockets. It is hard to collect all variations to train a supervised model. Also we still need to face with the window size issue and other issues as for the activity recognition task.

13.3.3 Pilot Study: Prototype Model

For this specific application, I would like to simplify the problem at my best capability to ensure the feasibility. The main issues for this specific application are extracting “good walking” activity when the phone or device is placed at the right position. (now at L3 but it could be better to adjust to a regular carrying position like pants pockets.) Also, since chronic diseases are not expected to have status change in shorter temporal period, like within a week or even a month, the model is not required to have high recall which captures all “good walking” pieces. On the other hand, the model must guarantee high precision to feed true positive samples for the status detection. This means we can manually tune the model to have close-to-perfect precision by sacrificing more on recall.

I have done pilot studies to apply both activity recognition and position recognition for detecting good walking with smartphones. “Good walking” detection is not equivalent to traditional activity recognition, which requires extremely high precision but allows comparatively low recall while classifying walking with other activities. Besides, good walking detection is more like a binary classification rather than a multiple-way classification so it does not require the model to be sensitive to the difference between other activities. It is more likely to be a one-class classification problem in the original prototype, Figure 13.2. Keep in mind that a 2 minute steady walking is a special activity against all other potential daily activities for old patients. Walking is probably the only comparatively long periodic activity in their life. This simple evidence will help us to eliminate most unqualified samples.

This framework takes as input accelerometer data in (x, y, z) coordinate. The data is divided into 10 second data stream, where these segments are units of classification. These data then go through two phases of recognition process, activity and position recognition.

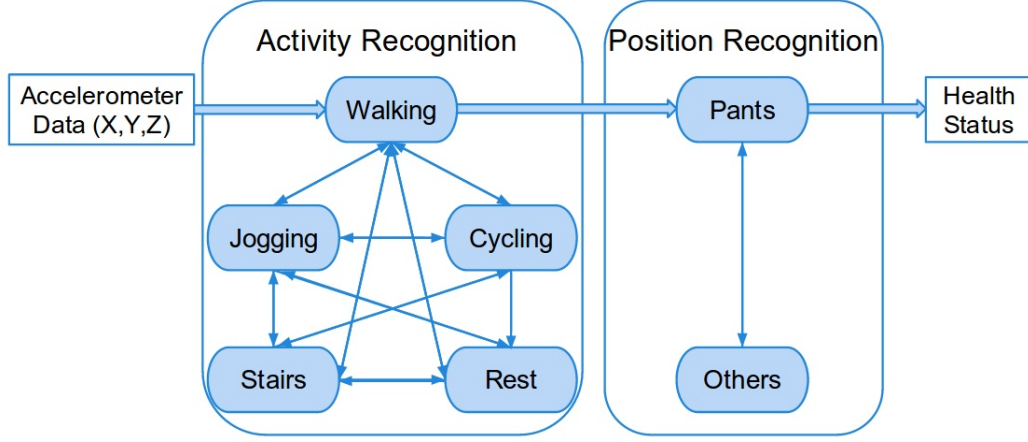


Figure 13.2: Overall Framework of GaitTrack

Activity (walking, jogging, cycling, stationary and going up the stairs) is first detected using accelerometer data stream via

$$y_a(t)* = \operatorname{argmax}_a(Pr(y_a(t)|x(t))) \quad (13.1)$$

where the accelerometer data at time segment t is $x(t)$, and the classified activity state is y_a* . Using the classification results, position of the phone is then detected. It is important to note the position takes into account the activity classification results. More formally

$$y_p(t)* = \operatorname{argmax}_p(Pr(y_p|x(t), y_a(t))) \quad (13.2)$$

The ultimate goal of detecting the health status, while not covered in this report, is then detected by

$$y_h(t)* = \operatorname{argmax}_h(Pr(y_h|x(t), y_a(t) = walking, y_p(t) = pants)) \quad (13.3)$$

Notice that only classify the health status if the activity is walking and is in the person's pants or equivalent positions.

It should be noted that for each of the two phases, the data goes through preprocessing stage. From the preprocessed dataset, it is then fed into classification algorithm.

An active learning strategy can be applied so that new feeds of activity samples can contribute to enhancing the model performance. We have a “other” class for all unknown activities and positions and other cases we

didn't consider in prior. Then we need to eliminate all samples that are not certain to be "good walk" at "right position". This strategy is not suitable for traditional activity recognition tasks but for retrieving "good walk" for health status prediction.

13.3.4 The Desired Position

In my pilot study, I reproduced the activity recognition model and built a position recognition model based on it with walking data from phones. Be aware that the key features here are replaceable, such as the "desired device position". Here I use "pants pocket" just because I assumed that usually a adult male will wear pants with pockets and the pockets are the closest place to L3 when you normally carry your phone. That means, based on different assumptions, the desired position could be differed, especially when we choose other wearable devices. The assumption I use here is there are several equivalent positions to L3, such as fanny pack, pants pockets, coat pockets, tote bags, etc. Other positions such as hand bags, purses, braceband, armband, necklace, which are further away from L3 position, are not equivalent positions.

What if the device is not placed at the desired position or the equivalent positions? The short answer is we can simply eliminate the data if detected as other positions. In daily monitoring context, we can assume the subject will constantly generate good walking data above day-to-day scale. That is to say, at least, we can obtain one qualified sample of good walking at desired position per day. As we know chronic disease are comparatively stable and the status won't change minute to minute, but week to week or even month to month. Thus, eliminating unqualified pieces of data has limited effect on the over all performance if the subjects follow the medical advice and wear the device at the right position most of the time.

Of course it will be better if there are ways to convert or adjust data collected from other positions to a standard position, as a normalization for different positions. However, it requires large additional effort and this topic itself could be a dissertation study. Similarly, adjusting data from different devices, different sampling rate or different time frame are also in this category. If resource permits, there is no harm to do deeper study on

these topics.

13.4 Passive Monitoring of Human Health From Daily Activities

Measuring motion is a potential solution for passive monitoring, which has compliance advantages over active phone applications. That is, the patient simply uses their personal phone as usual during daily activities, with no special actions necessary. The home trials are more complex than the hospital trials since patients do many different activities throughout their day, while periods when they are carrying their phones recorded by a continuous monitor. The passive monitor phone app in the home is a re-implementation of the active monitor in the clinic, which only records when the phone is in steady motion and only archives when the patient is good walking.

13.4.1 Feasibility of Passive Monitoring

The dissertation study demonstrated that the predictive models have high accuracy for walk test data. As we discussed in the above section, activity recognition and position detection strategies should be applied prior to feed the model. Both activity recognition and position detection can be simplified in the passive monitor. The recognition models only need to detect good walking at desired position, for a sufficient length of time. Since the amount of walking needed for accurate computation is very small, 1 or 2 minutes per session, the filter can be tuned up, for high precision only during periods when there is definite measurement of good walking.

13.4.2 Pilot Results

Fortunately the IRB protocol supports for initial home study and two sets of walking data were streamed back for two COPD patients at Northshore. Both of the patients took the phone back home with passive monitor turned on for a whole week. In addition, a prospective patient, who is a GOLD1 smoker with COPD, was involved too. Each of these patients carried a Galaxy S5 mini

loaded with our passive monitor for one week. The good walking filters on the phone and the server extracted the periods when qualified sessions were taking place. These home sessions were utilized as input to the universal model trained by the clinic experiments for status prediction.

One patient has 7 eligible good walking samples in record and another has 75. Both of the patients passive monitor results are identical to their GOLD status according to their pulmonary function test records. The smoker has been detected mostly GOLD 1 even he didn't take any pulmonary examinations.

13.4.3 Potential Applications

The pilot results are simple but supportive to two major medical application for passive monitoring: care routing and case finding. For care routing, the passive monitor could have continuous status report for each COPD patients, and the results could be helpful for tracking status changes. For case finding, the smoker example shows that the passive monitor could be used to detect if a subject has COPD or not, simply by carrying a phone for several weeks. Then the ones with high risk to be COPD could schedule physician diagnosis to confirm the status. The general comparison between the proposed passive monitor approach and traditional diagnosis approach is shown in Figure 13.3.

13.4.4 Manipulation for Similar Medical Applications

The philosophy of this dissertation study could be applied to similar medical applications. Such as health monitor for other diseases. I mentioned congestive heart failure (CHF) in previous discussion, which are also medically proved to be correlated with motion. Similarly, fall detection and prediction requires better analysis of body motion previous to the time point of subject's falling. All these applications require rigorous design from data collection, data cleaning, all the way to model training and evaluation. There are a large set of potential techniques for each step. How to better select the most feasible techniques and design the system will be a crucial factor to success.

This dissertation provides strategic level ideas for all these challenging applications. Monitoring health is, and will be a very significant application

More Helped, Fewer Seen By PCP

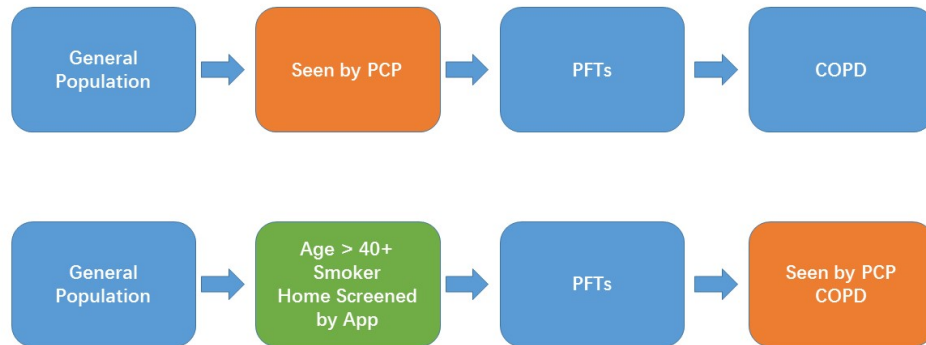


Figure 13.3: Comparison of traditional COPD diagnosis approach and proposed approach. Passive monitor will enhance case finding capacity.

domain for information retrieval and data mining techniques. I believe the improvement of technology will absolutely change the way how we treat our health issues, and make human life better.

CHAPTER 14

LESSONS LEARNED

The dissertation contains the major achievements and outcome of my Ph.D. study. However, I have learned much more during the five years of research experiences. I failed, succeeded, gained and lost. In this chapter I listed several general ideas I have learned beyond the results. Both input and output end are important for model selection. Here two aspects I am focusing on each side, to obtain sustainable model training, Figure 14.1.

14.1 Resolution

For monitoring health in general with machine learning models, whether to apply a classification technique, or a regression technique, should always be the fundamental question we ask to ourselves. In this dissertation, for different medical applications, I end up using different strategies: for gait speed prediction, I applied regression; for pulmonary functions, I applied classification and for oxygen saturation, I applied both. So why did I make these choices?

To deal with such questions, I always assess from both the input end and output end. From the input end I need to evaluate if we have enough samples. We all know that regression model works well if the training samples are equally distributed across the whole range of domain.

For gait speed, we collected a series of walking speed, from very tired sick patients to very energetic healthy people. Each lap their speed varies so it provides multiple sample points. From the output end, the output of gait speed must be numerical, an precise number for the current speed, which will then be converted to distance to match the 6-minute walk distance records. Thus, a regression model could be the only option[29].

For pulmonary function, on the other end, only classification model can

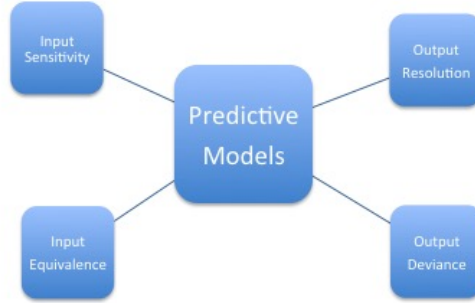


Figure 14.1: Main factors affect model quality.

be selected under my circumstance. First, from the input view, the sample points are rare. Each of our patient only have one set of PFT, and we only have about thirty patients. The range of FEV1% is from 0-100, covering GOLD levels from 1 to 4. A regression model could never work well in such situation and the over-fitting risk is very high. From the output end, it could be extremely doubtful if I claim we can predict PFT values from body motion to physicians, because these are totally separate systems. However, all physicians will agree with that both PFT and motion indicate health status, which is GOLD levels, in this context. Thus, predicting the GOLD levels for each patient is more valuable and persuasive and less risky, and a classification model is a perfect match for this task[77] [86].

For oxygen saturation, which is in the middle of the two above extreme cases. Simple oxygen saturation indicates big difference medically. De-saturation, which represents the oxygen saturation value drops sharply in a short time, is a significant health issue. At the same time, during our study, we found there are typical patterns of oxygen saturation values during 6MWT, strongly indicating different COPD severity levels and transitions.

Thus, from the output view both regression model and classification model have its value. Then we tried both with the input data, which is restricted. The classification model works much better and more reliable than the regression model. And this is a major reason why we prefer classification models in the following studies for oxygen saturation[43].

14.2 Sensitivity

The study contains a clinical trial to collect patient's data, which will be extremely expensive, both for money and time. The IRB also requires everything is well designed: how many patients, what kind of exercises each patient need to perform, etc. Such requirement is opposite to the nature of data science: any conclusions should be based on data. We were always struggling in estimating how many patients we need to train a good model before we get any data. The design of the experiment proposal will also need to be modified all the time along with data collection.

The number of initial recruitment depends on a general sense of persuasion. We had conversation with different physicians, investigated relevant literature to see what is a proper number in general for such kind of research. A comparatively small number of patient is acceptable to propose a novel idea about applying mobile phones to track health, which is 20 to 30, of three or four different categories. However, when it comes to assessing the feasibility of your proposed work medically, ten times more subjects should better be involved to cover major demographic cohorts, status variations and so forth. When it is for medical quality a commercial product, a releasable prototype should include data from more than a thousand, covering minor variations within each category. However, the basic philosophy is the more, the better.

Another important issue for sensitivity is over-fitting. More training data yields to better accuracy in model validation, which also increases the risk of over-fitting. Statistically we can adjust the portion of training and testing set in model validation to see when to hit a over-fitting point. Also we need to avoid bias in data collection and guarantee the labeling process is reliable. Also keep in mind that machine learning is similar to human learning. It does the best on handling tasks it has seen in the training set, and it will match any known tasks to the closest known task, which, in general, will

definitely lead to a worse performance.

The first and foremost step to avoid overfitting is proposing a correct assumption for training the model. Correlation studies must be prior to the machine learning step. Domain knowledge is always significant for artificial intelligence. In general, there is no ultimate artificial intelligent solution for everything. Even human beings need to be taught and trained to obtain skills, and machines are much dumber than human so it should be carefully told what it needs to do and how it can do it better.

14.3 Equivalence

Equivalence is another important topic before we estimating how to handle the proposed task. There are similar studies: similar topic, similar methodology, similar techniques in the literature.

At sensing level, the major task is based on sampling when the phone is held on the L3 position with a fanny pack, to match Kinesiology experiment results. However, L3 would not be a normal position when people carrying their phones. We can assume that the equivalence of L3 position to be the closest normal positions where a phone could be carried, like pants pockets, coat pockets, on the belt or so forth. All data collected from these positions are supposed to be feasible for health monitoring. But other positions such as hand bags, armbands, will not be regard as an equivalent position.

At model level, classification model will turn to be a regression model if the target classes are continuously uniformly distributed in space. That is, if it is necessary to build model to classify infinite classes, the classification technique and regression technique could switch. In medical context, how many classes are needed depends on the number of treatments we can provided. Theoretically each individual could be a unique class, and that is the ultimate goal for personalized medicine as well.

14.4 Deviance

Predicting current status is just the first step of monitoring health, because knowing the current status from a device will not be able to help medical

treatment for chronic diseases. Instead, detecting status deviance is a more important issue for applying motion techniques for monitoring health. The change point of status represents the best time to see your physicians and update your treatment before the disease getting worse.

For three different medical applications, I proposed three different ways to detect deviance, that is, to find the transition status. For gait speed and raw walking acceleration, autocorrelation coefficient (AC) is a major factor representing pattern changes, which calculates the similarity between each cycles. In general, detecting changes of periodic data sequence could apply this technique. For oxygen saturation or, similarly the heart rate R-R variability, a temporal data but not periodic in a 6-minute window frame, a combination of mean and standard deviation is enough to describe the different patterns and the transition in between. For pulmonary functions, actually each patient has one single sample point on the chronic sequence. Thus, directly compare the portion of predicted GOLD levels will indicates the internal variation of status. The health status is supposed to be stable during the whole experiment but if it shows instability, the patient could be under sort of changing situation.

REFERENCES

- [1] M. ANDERSON, “Technology device ownership: 2015,” *Pew Research Internet Project*, 2015.
- [2] S. Fox and M. Duggan, “Tracking for health,” *Pew Research Internet Project*, Jan. 2013. [Online]. Available: <http://www.pewinternet.org/2013/01/28/tracking-for-health/>
- [3] G. Rose and L. Colwell, “Randomised controlled trial of anti-smoking advice: final (20 year) results.” *Journal of epidemiology and community health*, vol. 46, no. 1, pp. 75–77, 1992.
- [4] B. R. Schatz and R. B. Berlin, *Healthcare Infrastructure: Health Systems for Individuals and Populations*. Series in Health Informatics. Springer London Limited, 2011.
- [5] V. C. Broaddus, R. C. Mason, J. D. Ernst, T. E. King Jr, S. C. Lazarus, J. F. Murray, J. A. Nadel, A. Slutsky, and M. Gotway, *Murray & Nadel’s Textbook of respiratory medicine*. Elsevier Health Sciences, 2015.
- [6] S. Studenski, S. Perera, K. Patel, C. Rosano, K. Faulkner, M. Inzitari, J. Brach, and et al., “Gait speed and survival in older adults,” *J American Medical Association JAMA*, vol. 305, pp. 50–58, 2010.
- [7] S. Studenski, “Bradypedia: is gait speed ready for clinical use?” *The Journal of Nutrition, Health & Aging*, vol. 13, no. 10, pp. 878–880, 2009.
- [8] N. S. Schenkel, L. Burdet, B. De Muralt, and J. Fitting, “Oxygen saturation during daily activities in chronic obstructive pulmonary disease,” *European Respiratory Journal*, vol. 9, no. 12, pp. 2584–2589, 1996.
- [9] B. Celli, W. MacNee, A. Agusti, A. Anzueto, B. Berg, A. S. Buist, P. Calverley, N. Chavannes, T. Dillard, B. Fahy et al., “Standards for the diagnosis and treatment of patients with copd: a summary of the ats/ers position paper,” *European Respiratory Journal*, vol. 23, no. 6, pp. 932–946, 2004.

- [10] S. of Left Ventricular Dysfunction Investigators et al., “Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure,” *N Engl j Med*, vol. 325, no. 5, pp. 293–302, 1991.
- [11] K. Dharmarajan, A. F. Hsieh, Z. Lin, H. Bueno, J. S. Ross, L. I. Horwitz, J. A. Barreto-Filho, N. Kim, L. G. Suter, S. M. Bernheim et al., “Hospital readmission performance and patterns of readmission: retrospective cohort study of medicare admissions,” *BMJ*, 2013.
- [12] M. R. Lemke, T. Wendorff, B. Mieth, K. Buhl, and M. Linnemann, “Spatiotemporal gait patterns during over ground locomotion in major depression compared with healthy controls,” *Journal of psychiatric research*, vol. 34, no. 4, pp. 277–283, 2000.
- [13] C. A. Camillo, F. Pitta, H. V. Possani, M. V. Barbosa, D. S. Marques, V. Cavalheri, V. S. Probst, and A. F. Brunetto, “Heart rate variability and disease characteristics in patients with copd,” *Lung*, vol. 186, no. 6, pp. 393–401, 2008.
- [14] M. Yamada, T. Aoyama, S. Mori, S. Nishiguchi, K. Okamoto, T. Ito, S. Muto, T. Ishihara, H. Yoshitomi, and H. Ito, “Objective assessment of abnormal gait in patients with rheumatoid arthritis using a smart-phone,” *Rheumatology international*, vol. 32, no. 12, pp. 3869–3874, 2012.
- [15] M. W. Whittle, *Gait analysis: an introduction*. Butterworth-Heinemann, 2014.
- [16] D. Longo, A. Paonessa, N. Specchio, L. N. Delfino, D. Claps, L. Fusco, F. Randisi, E. Genovese, F. Vigeveno, and G. Fariello, “Parry-romberg syndrome and rasmussen encephalitis: Possible association. clinical and neuroimaging features,” *Journal of Neuroimaging*, vol. 21, no. 2, pp. 188–193, 2011.
- [17] J. C. Perry, J. Andureu, F. I. Cavallaro, J. Veneman, S. Carmien, and T. Keller, “Effective game use in neurorehabilitation: user-centered perspectives,” *Handbook of Research on Improving Learning and Motivation through Educational Games*, IGI Global, 2010.
- [18] J. Yang, “Toward physical activity diary: motion recognition using simple acceleration features with mobile phones,” in *Interactive Multimedia for Consumer Electronics*. ACM, 2009, pp. 1–10.

- [19] H. Vathsangam, A. Emken, D. Spruijt-Metz, and G. S. Sukhatme, "Toward free-living walking speed estimation using Gaussian process-based regression with on-body accelerometers and gyroscopes," in *Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. IEEE, 2010, pp. 1–8.
- [20] W. Zijlstra and A. L. Hof, "Assessment of spatio-temporal gait parameters from trunk accelerations during human walking," *Gait & Posture*, vol. 18, no. 2, pp. 1–10, 2003.
- [21] N. N. Dissanayaka, A. Sellbach, S. Matheson, J. D. O'Sullivan, P. A. Silburn, G. J. Byrne, R. Marsh, and G. D. Mellick, "Anxiety disorders in parkinson's disease: prevalence and risk factors," *Movement Disorders*, vol. 25, no. 7, pp. 838–845, 2010.
- [22] R. Moe-Nilssen and J. L. Helbostad, "Estimation of gait cycle characteristics by trunk accelerometry," *Journal of biomechanics*, vol. 37, no. 1, pp. 121–126, 2004.
- [23] R. A. Rabinovich, Z. Louvaris, Y. Raste, D. Langer, H. Van Remoortel, S. Giavedoni, C. Burtin, E. M. Regueiro, I. Vogiatzis, N. S. Hopkinson et al., "Validity of physical activity monitors during daily life in patients with copd," *European Respiratory Journal*, vol. 42, no. 5, pp. 1205–1215, 2013.
- [24] H. Van Remoortel, Y. Raste, Z. Louvaris, S. Giavedoni, C. Burtin, D. Langer, F. Wilson, R. Rabinovich, I. Vogiatzis, N. S. Hopkinson et al., "Validity of six activity monitors in chronic obstructive pulmonary disease: a comparison with indirect calorimetry," *PloS one*, vol. 7, no. 6, p. e39198, 2012.
- [25] M. Jehn, A. Schmidt-Trucksäess, T. Schuster, H. Hanssen, M. Weis, M. Halle, and F. Koehler, "Accelerometer-based quantification of 6-minute walk test performance in patients with chronic heart failure: applicability in telemedicine," *Journal of Cardiac Failure*, vol. 15, no. 4, pp. 334–340, 2009.
- [26] M. L. Moy, N. A. Weston, E. J. Wilson, M. L. Hess, and C. R. Richardson, "A pilot study of an internet walking program and pedometer in copd," *Respiratory Medicine*, 2012.
- [27] M. L. Moy, M. Teylan, N. A. Weston, D. R. Gagnon, V. A. Danilack, and E. Garshick, "Daily step count is associated with plasma c-reactive protein and il-6 in a us cohort with COPD," *CHEST Journal*, vol. 145, no. 3, pp. 542–550, 2014. [Online]. Available: +

- [28] A. Albarbar, A. Badri, J. K. Sinha, and A. Starr, "Performance evaluation of mems accelerometers," *Measurement*, vol. 42, no. 5, pp. 790–795, 2009.
- [29] J. Juen, Q. Cheng, and B. R. Schatz, "Towards a natural walking monitor for pulmonary patients using simple smart phones," in *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, 2014.
- [30] Y. Jiang, Q. V. Liao, Q. Cheng, R. B. Berlin, and B. R. Schatz, "Designing and evaluating a clustering system for organizing and integrating patient drug outcomes in personal health messages," in *AMIA Annual Symposium Proceedings*, vol. 2012. American Medical Informatics Association, 2012, p. 417.
- [31] S. Nishiguchi, M. Yamada, K. Nagai, S. Mori, Y. Kajiwara, T. Sonoda, K. Yoshimura, H. Yoshitomi, H. Ito, K. Okamoto, and et al., "Reliability and validity of gait analysis by android-based smartphone," *Telemedicine and e-Health*, vol. 18, no. 4, pp. 292–296, 2012.
- [32] S. Nishiguchi, H. Ito, M. Yamada, H. Yoshitomi, M. Furu, T. Ito, A. Shinohara, T. Ura, K. Okamoto, and T. Aoyama, "Self-assessment tool of disease activity of rheumatoid arthritis by using a smartphone application," *Telemedicine and e-Health*, 2014.
- [33] J. Park, A. Patel, D. Curtis, S. Teller, and J. Ledlie, "Online pose classification and walking speed estimation using handheld devices," in *Ubiquitous Computing*. ACM, 2012, pp. 113–122.
- [34] V. Pinto-Plata, C. Cote, H. Cabral, J. Taylor, and B. Celli, "The 6-min walk distance: change over time and value as a predictor of survival in severe copd," *European Respiratory Journal*, vol. 23, no. 1, pp. 28–33, 2003. [Online]. Available: <http://erj.ersjournals.com/content/23/1/28>
- [35] C. Karpman, Z. S. DePew, N. K. LeBrasseur, P. J. Novotny, and R. P. Benzo, "Determinants of gait speed in copd," *Chest*, vol. 146, no. 1, pp. 104–110, 2014.
- [36] K. F. Rabe, S. Hurd, A. Anzueto, P. J. Barnes, S. A. Buist, P. Calverley, Y. Fukuchi, C. Jenkins, R. Rodriguez-Roisin, C. van Weel et al., "Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary," *American journal of respiratory and critical care medicine*, vol. 176, no. 6, pp. 532–555, 2007.

- [37] C. Casanova, C. Cote, J. M. Marin, V. Pinto-Plata, J. P. de Torres, A. Aguirre-Jaíme, C. Vassaux, and B. R. Celli, “Distance and oxygen desaturation during the 6-min walk test as predictors of long-term mortality in patients with copd,” *CHEST Journal*, vol. 134, no. 4, pp. 746–752, 2008.
- [38] C. Casanova, M. C. Hernández, A. Sánchez, I. García-Talavera, J. P. de Torres, J. Abreu, J. M. Valencia, A. Aguirre-Jaime, and B. R. Celli, “Twenty-four-hour ambulatory oximetry monitoring in copd patients with moderate hypoxemia,” *Respiratory Care*, vol. 51, no. 12, pp. 1416–1423, 2006.
- [39] C. Keller, J. Brimacombe, F. Agro, and J. Margreiter, “A pilot study of pharyngeal pulse oximetry with the laryngeal mask airway: a comparison with finger oximetry and arterial saturation measurements in healthy anesthetized patients,” *Anesthesia & Analgesia*, vol. 90, no. 2, pp. 440–444, 2000.
- [40] K. K. Tremper, “Pulse oximetry,” *CHEST Journal*, vol. 95, no. 4, pp. 713–715, 1989.
- [41] M. Loggan, G. R. Kerby, and S. K. Pingleton, “Is routine assessment of arterial oxygen saturation in pulmonary outpatients indicated?” *CHEST Journal*, vol. 94, no. 2, pp. 242–244, 1988. [Online]. Available: +
- [42] S. Minami, S. Yamamoto, Y. Ogata, T. Nakatani, Y. Takeuchi, M. Hamaguchi, T. Koba, and K. Komuta, “Ambulatory pulse oximetry monitoring in Japanese copd outpatients not receiving oxygen therapy,” *Multidisciplinary respiratory medicine*, vol. 9, no. 1, p. 24, 2014.
- [43] Q. Cheng, J. Juen, J. Hsu-Lumetta, and B. Schatz, “Predicting transitions in oxygen saturation using phone sensors,” *Telemedicine and e-Health*, 2015.
- [44] Q. Cheng, J. Juen, and B. R. Schatz, “Using mobile phones to simulate pulse oximeters: gait analysis predicts oxygen saturation,” in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2014, pp. 331–340.
- [45] A. C. on Proficiency Standards for Clinical Pulmonary Function Laboratories et al., “Aats statement: guidelines for the six-minute walk test.” *American journal of respiratory and critical care medicine*, vol. 166, no. 1, p. 111, 2002.
- [46] J. Juen, Q. Cheng, and B. Schatz, “A natural walking monitor for pulmonary patients using mobile phones,” *Biomedical and Health Informatics, IEEE Journal of*, vol. 19, no. 4, pp. 1399–1405, 2015.

- [47] Q. Cheng, J. Juen, Y. Li, V. Prieto-Centurion, J. A. Krishnan, and B. R. Schatz, "GaitTrack: Health monitoring of body motion from spatio-temporal parameters of simple smart phones," in *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*. ACM, 2013, p. 897.
- [48] J. Juen, Q. Cheng, V. Prieto-Centurion, J. A. Krishnan, and B. Schatz, "Health monitors for chronic disease by gait analysis with mobile phones," *Telemedicine and e-Health*, 2014.
- [49] R. Lemoyne, C. Coroian, T. Mastroianni, and W. Grundfest, "Accelerometers for quantification of gait and movement disorders: a perspective review," *Journal of Mechanics in Medicine and Biology*, vol. 8, no. 02, pp. 137–152, 2008.
- [50] M. Yang, H. Zheng, H. Wang, S. McClean, and N. Harris, "Assessing the utility of smart mobile phones in gait pattern analysis," *Health and Technology*, vol. 2, no. 1, pp. 81–88, 2012.
- [51] U. Grenander, *Probability and Statistics: The Harald Cramer Volume*, ser. Wiley Publications in Statistics. Alqvist & Wiksell, 1959. [Online]. Available: <http://books.google.com/books?id=UPc0AAAAMAAJ>
- [52] H. Lu, J. Yang, Z. Liu, N. D. Lane, T. Choudhury, and A. T. Campbell, "The Jigsaw continuous sensing engine for mobile phone applications," in *Embedded Networked Sensor Systems*. ACM, 2010, pp. 71–84.
- [53] M. Yang, H. Zheng, H. Wang, S. McClean, J. Hall, and N. Harris, "Assessing accelerometer based gait features to support gait analysis for people with complex regional pain syndrome," in *Proceedings of the 3rd International Conference on Pervasive Technologies Related to Assistive Environments*, ser. PETRA '10. New York, NY, USA: ACM, 2010, pp. 48:1–48:7.
- [54] J. Pang, Y. Shi, Y. Cao, D. Chen, X. Gu, B. Chen, Y. Wu, H. Zhan, and Y. Shi, "Study on distribution features of tender points in patients with knee osteoarthritis by cluster analysis," *Zhongguo gu shang= China journal of orthopaedics and traumatology*, vol. 27, no. 1, pp. 21–24, 2014.
- [55] C. E. Shannon, "A mathematical theory of communication," *ACM SIG-MOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.
- [56] J. A. Swets, *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Lawrence Erlbaum Associates, Inc, 1996.
- [57] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. John Wiley & Sons, 2004, vol. 46.

- [58] T. Bernecker, F. Graf, H.-P. Kriegel, C. Moennig, D. Dill, and C. Tuermer, “Activity recognition on 3d accelerometer data,” (*Technical Report*), 2012.
- [59] S. Frühwirth-Schnatter, “Data augmentation and dynamic linear models,” *Journal of time series analysis*, vol. 15, no. 2, pp. 183–202, 1994.
- [60] J. Durbin and S. J. Koopman, *Time series analysis by state space methods*. Oxford University Press, 2012, no. 38.
- [61] D. A. Jackson, “Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches,” *Ecology*, pp. 2204–2214, 1993.
- [62] V. N. Vapnik, “An overview of statistical learning theory,” *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.
- [63] J. MacQueen et al., “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 281–297. California, USA, 1967, p. 14.
- [64] H. Madsen and P. Thyregod, *Introduction to general and generalized linear models*. CRC Press, 2010.
- [65] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [66] S. S. Keerthi and C.-J. Lin, “Asymptotic behaviors of support vector machines with gaussian kernel,” *Neural computation*, vol. 15, no. 7, pp. 1667–1689, 2003.
- [67] M. A. Friedl and C. E. Brodley, “Decision tree classification of land cover from remotely sensed data,” *Remote sensing of environment*, vol. 61, no. 3, pp. 399–409, 1997.
- [68] A. Liaw and M. Wiener, “Classification and regression by randomforest,” *R news*, vol. 2, no. 3, pp. 18–22, 2002.
- [69] C. E. Rasmussen, “Gaussian processes in machine learning,” *Advanced lectures on machine learning*, pp. 63–71, 2004.
- [70] A. J. Smola and B. Schölkopf, “A tutorial on support vector regression,” *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.
- [71] J. Shang, W. Tong, J. Peng, and J. Han, “Dpclass: An effective but concise discriminative patterns-based classification framework,” in *SIAM International Conference on Data Mining (SDM 2016)*, 2016.

- [72] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced data," *University of California, Berkeley*, 2004.
- [73] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.
- [74] J. Friedman, T. Hastie, and R. Tibshirani, "glmnet 1.1: lasso and elastic-net regularized generalized linear models. r package," 2009.
- [75] R. Kohavi et al., "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2, 1995, pp. 1137–1145.
- [76] K. O. May, "A set of independent necessary and sufficient conditions for simple majority decision," *Econometrica: Journal of the Econometric Society*, pp. 680–684, 1952.
- [77] Q. Cheng, J. Juen, S. Bellam, D. Close, S. J., and B. Schatz, "Classification model for pulmonary function using motion analysis from phone sensors," *AMIA Annual Symposium Proceedings*, 2016.
- [78] ATS, "ATS statement: Guidelines for the six-minute walk test," *American Journal of Respiratory and Critical Care medicine*, vol. 166, no. 9, pp. 111–117, 2002.
- [79] C. S. Wallace and D. M. Boulton, "An information measure for classification," *The Computer Journal*, vol. 11, no. 2, pp. 185–194, 1968.
- [80] C. Chang and C. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.
- [81] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.
- [82] T. M. Therneau, B. Atkinson, and B. Ripley, "rpart: Recursive partitioning. r package version 3.1-46," *Computer software program retrieved from <http://CRAN.R-project.org/package=rpart>*, 2010.
- [83] Y. Song, S. Shin, S. Kim, D. Lee, and K. H. Lee, "Speed estimation from a tri-axial accelerometer using neural networks," in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*. IEEE, 2007, pp. 3224–3227.
- [84] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74–82, 2011.

- [85] B. A. Emken, M. Li, G. Thatte, S. Lee, M. Annavaram, U. Mitra, S. Narayanan, and D. Spruijt-Metz, "Recognition of physical activities in overweight hispanic youth using knowme networks," *Journal of physical activity & health*, vol. 9, no. 3, p. 432, 2012.
- [86] Q. Cheng, J. Juen, S. Bellam, N. Fulara, D. Close, J. C. Silverstein, and B. Schatz, "Predicting pulmonary function from phone sensors," *Telemedicine and e-Health*, 2017.