UNIVERSITY OF LEEDS

An automated algorithmic approach for activity recognition and step detection in the presence of functional compromise

Valeria Filippou

Submitted in accordance with the requirements for the degree of Doctor of Philosophy (PhD)

The University of Leeds

School of Mechanical Engineering

February 2021

Intellectual Property and Publication Statements

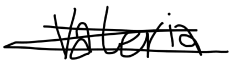
The candidate confirms that the work submitted is his/her/their own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

The work in Chapter 3 of the thesis has appeared in publication as follows: V. Filippou, A. C. Redmond, J. Bennion, M. R. Backhouse and D. Wong, "Capturing accelerometer outputs in healthy volunteers under normal and simulated-pathological conditions using ML classifiers*," 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 2020, pp. 4604-4607, doi: 10.1109/EMBC44109.2020.9176201.

V. Filippou led on all aspects of design, planning, data capture, data analysis, interpretation and writing the manuscript and was the lead presenter.

J. Bennion contributed to the data capture, interpretation and writing of the manuscript.

A. C. Redmond, M. R. Backhouse and D. Wong were the supervisors and contributed to the design, planning, interpretation and writing of the manuscript.

Signed 

Acknowledgements

Undertaking this PhD has been a truly life-changing experience for me and it would not have been possible to accomplish without the continuous support and guidance that I received from many people.

I would like to say a huge thank you to my supervisors Prof Anthony C Redmond, Dr Michael R Backhouse and Dr David Wong for all their help and advice throughout this PhD. Their knowledge and experience have encouraged me during the whole time of my academic research as I have always felt that I have all the support that I needed.

I want to thank Jacquline Bennion for her help during the initial stage of the pilot study undertaken for my PhD.

I would particularly like to thank Dr Claire Brockett, iMBE team and my fellow students for the support I had when I needed it.

Special thanks also to the Patient and Public Involvement group and the clinicians who took the time to discuss with me about my PhD. Their input was valuable in understanding a different perspective of my project.

I would also like to thank all the people who took part in the pilot study. Without their input, this PhD would not have been successful.

I gratefully acknowledge the funding received towards my PhD from the Engineering and Physical Sciences Research Council (EPSRC) Centre for Doctoral Training in Tissue Engineering and Regenerative Medicine.

Last but not least, I wish to acknowledge the support and great love of my family and friends; especially Christiana and Nick. They kept me going on and this work would not have been possible without their input.

Abstract

Wearable technology is a potential stepping stone towards personalised healthcare. It provides the opportunity to collect objective physical activity data from the users and could enable clinicians to make more informed decisions and hence provide better treatments. Current physical activity monitors generally work well in healthy populations but can be problematic when used in some patient groups with severely abnormal function.

We studied healthy volunteers to assess how different algorithms might perform for those with normal and simulated-pathological conditions. Participants (n=30) were recruited from the University of Leeds to perform nine predefined activities under normal and simulated-pathological conditions using two MOX accelerometers on wrist and ankle (Maastricht Instruments, NL). Condition classification was performed using a Support Vector Machine algorithm. Activity classification was performed with five different Machine Learning algorithms: Support Vector Machine, k-Nearest Neighbour, Random Forest, Multilayer Perceptron, and Naïve Bayes. A step count algorithm was developed based on pattern recognition approach, using two main techniques, Dynamic Time Warping and Dynamic Time Warping-Barycentre Averaging. Finally, synthetic acceleration signal was generated that represented walking activities since there was limited access to patient data and to refine synthetic data generation in this field. Three dynamic coupled equations were used to represent the morphology of the desired signal.

Wrist and ankle locations performed similarly and the wrist location was used for further analysis. Both condition and activity classification algorithms achieved good performance metrics i.e. that the volunteer has been correctly classified in the right condition, and the activities performed have been correctly recognised. Additionally, the novel step count algorithm achieved more accurate results for both conditions in comparison to existing algorithms from the literature. Finally, the signal generation approach seems promising since the normal condition

synthetic signals matched closely to their associated original signals.

Algorithms developed for a specific group or even person with functional pathology, using techniques such as Dynamic Time Warping-Barycentre Averaging produce better results than traditional algorithms trained on data from a different group.

Publications and conferences

Publications

V. Filippou, A. C. Redmond, J. Bennion, M. R. Backhouse and D. Wong, "Capturing accelerometer outputs in healthy volunteers under normal and simulated-pathological conditions using ML classifiers*," 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 2020, pp. 4604-4607, doi: 10.1109/EMBC44109.2020.9176201.

Conference Presentations

I gave oral and poster presentations at national meetings and student-organised conferences. I also presented my work at international conferences:

- 6th International Conference on Ambulatory Monitoring of Physical Activity and Movement (ICAMPAM) 2019, Maastricht: Poster presentation
- Leeds Institute for Data Analytics (LIDA) Seminar series 2019, Leeds: Oral presentation
- Healthcare Technologies and Early Career Awards 2019, London: Poster presentation
- 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) 2020, Mondreal: Oral presentation

Contents

List of Figures

List of Tables

List of Abbreviations

ADL(s) Activitie(s) of Daily Living

AI Artificial Intelligence

ANN Artificial Neural Network

ARC Activity Recognition Chain

C Regularisation parameter

CART Classification and regression trees

CIs Confidence Intervals

CNNs Convolutional Neural Networks

CO_2 Carbon Dioxide

CV Cross Validation

DBA Dynamic Time Warping Barycentre Averaging

DFT Discrete Fourier Transform

DLW Doubly Labelled Water

DTW Dynamic Time Warping

ECG Electrocardiogram

EE Energy Expenditure

FFT Fast Fourier Transform

FITT Frequency, Intensity, Time, Type

FN False Negative

FP False Positive

GANs Generative Adversarial Networks

GB Gaussian Naive Bayes

^{2}**H** Hydrogen isotope

HAR Human Activity Recognition

Hz Hertz

IMU Inertial Measurement Unit

IQR Interquartile Range

kNN k-Nearest Neighbour

LOA Limits of Agreement

ML Machine Learning

MLP Multilayer Perceptron

MMH Maximal Margin Hyperplane

MPSD Mean Power Spectral Density

NCDs Non-Communicable Diseases

$\mathbf{O_2}$ Oxygen

^{18}**O** Oxygen isotope

NN Neural Network

PA Physical Activity

PCA Principal Component Analysis

PPI Patient and Public Involvement

RA Rheumatoid Arthritis

RF Random Forest

RMS Root mean square

RMSE Root Mean Square Error

SI International System of Units

Std Standard deviation

SVM Support Vector Machine

TP True Positive

TN True Negative

UK United Kingdom

WHO World Health Organisation

1D One-dimensional

2D Two-dimensional

3D Three-dimensional

Chapter 1

Introduction

1.1 Background

People with musculoskeletal disease often walk with impaired gait. That reduced mobility can increase risk of secondary complications such as cardiovascular disease, so there is a drive to increase patient physical activity. In other words, increased physical activity prevents and delays onset of several chronic conditions (Phillips et al. 2018). Generally, physical activity is very important since it influences peoples' health and well-being (including mental health). However, measuring physical activity in people with impaired gait is challenging.

Health tracking has received increased attention recently due to technological improvements (Majumder et al. 2017). One way to measure daily physical activity is via commercially available activity monitors. Although many monitors have been developed, almost all have been designed for healthy people with a typical gait pattern, and who are able to perform moderate/vigorous-intensity activities. For people who walk abnormally or slowly, and for light-intensity activities, the devices are often inaccurate (Walker et al. 2016). Therefore, objective assessment of physical activity is not commonly used in clinical practice, and instead, self-reported activity of patients is used as a proxy.

Figure 1.1 demonstrates the life of a patient with chronic condition(s) throughout a year. Typically, patients with long term conditions visit the clinic once or twice a year. At this appointment, they are expected to recall details of their wellbeing, including their physical activity, since they were last seen, with many physical activity questionnaires focusing on the most

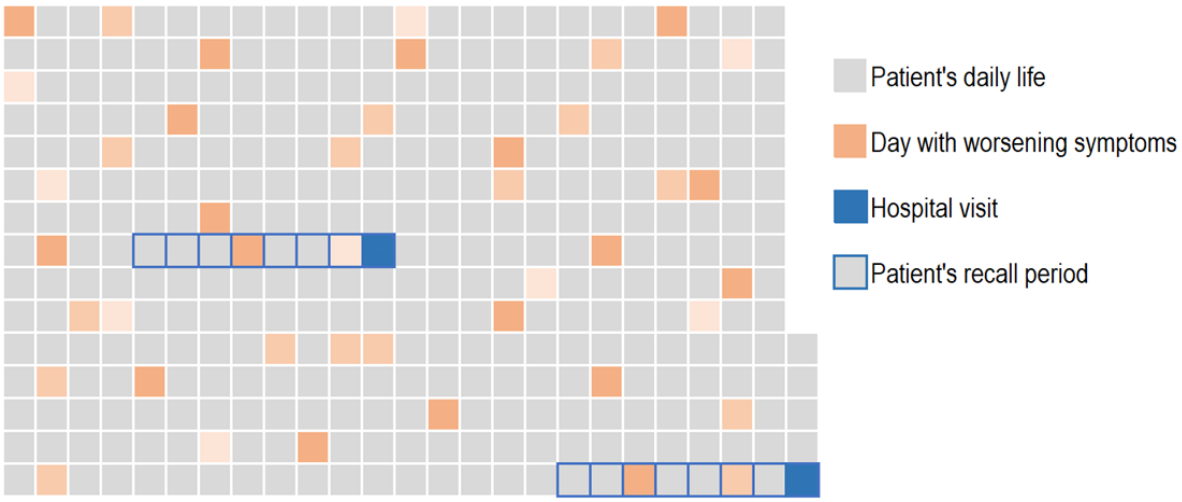


Figure 1.1: Schematic for patient's daily life for a whole year

recent week. This can be problematic for two reasons; 1) a seven day period is not always representative for the progression of the disease of the patient, and 2) the feedback from patients is subject to recall bias. Remote patient monitoring offers the opportunity to provide a more objective, long term overview of the progression of the disease.

Remote patient monitoring will benefit clinicians and patients. For example, clinicians could be able to continuously collect objective patient data to improve their healthcare decision-making (Baig et al. 2017). Patients could be provided with a patient-specific treatment with improved feedback and support (Baig et al. 2017). Additionally, patients would not need to recall any of the activities that they performed over a long period of time (Miller et al. 2013).

The potential clinical benefits of improved monitoring of physical activity mean that analysis of activity tracker data warrants further attention. There is a clear need to develop algorithms that can perform accurately in both healthy and pathological populations. Therefore, this thesis explores the use of signal processing and machine learning algorithms to identify the activity levels of users performing tasks of daily living. Additionally, a step count algorithm is described to calculate the number of steps of the user. Importantly, the step count algorithm targets users with atypical gait patterns.

Accelerometer sensors have been used extensively in wearables to measure physical activity as they can be used for both activity recognition and step count purposes. The work of this thesis provides the methodology required for, and implementations of, both activity monitoring and step count using an accelerometer.

1.2 Aim and objectives

The governing hypothesis being explored in this thesis is that tuneable activity monitor algorithms could produce better outputs than standard algorithms if specifically trained to recognize data relating to abnormal activities. Hence, the overall aim of this thesis is to explore the development of tuneable algorithms to more accurately measure physical activity in people with walking impairments.

Objectives:

1. To explore the effect of sensor mounting locations on condition classification and activity recognition (Chapter 3)
2. To explore the performance of a machine learning algorithm in identifying whether a patient is moving normally (Chapter 3)
3. To explore the performance of a set of machine learning algorithms in identifying different types of physical activity in healthy participants under normal and simulated-pathological conditions (Chapter 3)
4. To compare the ability of previously published step count algorithms to correctly count the number of steps taken by healthy participants performing normal and simulated-pathological gaits (Chapter 4)
5. To develop a novel step count algorithm to count the number of steps in healthy participants performing normal and simulated-pathological gaits (Chapter 5)
6. To generate synthetic acceleration data that represent normal and atypical walking patterns in order to test the relevant algorithms (Chapter 6)

1.3 Thesis overview

The rest of the thesis will be structured as follows:

Chapter 2 reviews the existing literature around physical activity and the current methods that are used to measure it. It outlines the advantages and disadvantages of each method before concluding that the best method for this work was accelerometry. Different aspects of the accelerometer sensor are then discussed in more detail, including how the sensor works and

what the output signal means. Additionally, the chapter also provides a complete overview of the Activity-Recognition-Chain process. The theory of popular machine learning and step count algorithms is also presented. Furthermore, an overview of the literature regarding the different methods used for activity recognition and step count is presented. Finally, relevant methods of generating synthetic walking signals are also discussed.

Chapter 3 outlines the design of a pilot study carried out with healthy volunteers. In this chapter the Activity-Recognition-Chain process, which was followed to analyse the acceleration data, is discussed step-by-step. The purpose of this chapter is to analyse physical activity data from healthy volunteers under normal and simulated-pathological conditions. The results show that there is a need to develop population-specific activity recognition algorithms.

Chapter 4 presents four potentially suitable step count algorithms from the literature; 1) peak detection, 2) thresholding (frequency-domain), 3) thresholding (time-domain), and 4) template-matching. The performance of each algorithm is examined in both normal and simulated-pathological gaits. Results from these analyses demonstrated suboptimal performance of the existing algorithms in simulated-pathological gait and confirmed the need to develop an improved algorithm.

Chapter 5 describes the development of a novel person-specific step count algorithm. It is based on the template-matching technique and uses Dynamic Time Warping for a similarity measure between the template and the acceleration signal. Finally, analysis within this chapter confirms that the new algorithm performs better in simulated-pathological gait, when compared to existing algorithms.

Chapter 6 describes a mathematical model that is used to generate synthetic data and was developed because of the challenges in conducting research on patients, especially during the COVID-19 pandemic. The model is inspired by McSharry and colleagues, who previously used this method to generate synthetic electrocardiogram signals (McSharry et al. 2003). The synthetic signals are validated using the data collected from the healthy volunteers. Therefore, this chapter provides an important methodological development to enable the synthesis of realistic accelerometer signals which will aid future advances in physical activity monitoring.

Chapter 7 draws together and discusses the main outcomes from each study as an integrated body of work. It considers the relevance of these findings in the context of existing literature, and

the general strengths and limitations of each study. In this context, it concludes by proposing key areas for future research.

Chapter 2

Literature review

2.1 Introduction

Chronic non-communicable diseases (NCDs) are one of the biggest health and development challenges in the 21st century and were responsible for the 71% of the world's deaths in 2016 (WHO 2018). NCDs are associated with common risk factors such as obesity, tobacco and alcohol use, unhealthy diets as well as physical inactivity, which are aspects of an unhealthy lifestyle (Maher et al. 2012). Among these modifiable risk factors, lack of physical activity (PA) is the most common (Warburton et al. 2006). Worldwide, one in four adults is insufficiently active (WHO 2017), and in the UK, 39% of adults do not meet physical activity (PA) recommendations (BHF 2017). Physical inactivity has been identified as one of the leading risk factors for premature mortality and worldwide deaths by the World Health Organization (WHO) (Dumith et al. 2011; Kohl et al. 2012; Lee et al. 2012; Taylor 2014; McPhail et al. 2014; Laarhoven et al. 2016; WHO 2017; Caron et al. 2017). Due to that, it is important to be able to measure PA accurately to ensure adequate activity levels are reached.

2.2 Physical activity and current technologies for measuring PA

2.2.1 Physical activity

2.2.1.1 Definition

PA has been defined by (Caspersen et al. 1985) as any "bodily movement produced by skeletal muscles which result in energy expenditure (EE)". In other words, PA is the behaviour that

increases EE above resting levels (Hills et al. 2014).

PA consists of four dimensions: frequency, intensity, time and type (FITT) (Butte et al. 2012; Strath et al. 2013). This information shows how often or how much (frequency), hard (intensity) and long (time), as well as what type of physical activity is performed (Hills et al. 2014). PA can be quantified based on these dimensions. Table 2.1 demonstrates a few examples for each category.

Table 2.1: Dimensions of physical activity and their parameters (Sliepen et al. 2018).

Dimension	Parameter
Frequency	Number of level steps
	Number of ascending steps
	Number of descending steps
Intensity	Walking cadence
Time	Time spent walking
	Time spent sitting
Type	Walking
	Sitting
	Standing

2.2.1.2 Guidelines

Health guidelines on levels of PA have been developed and recommended to the public by the WHO. The guidelines suggest that adults should participate in moderate PA for at least 150 minutes a week, or vigorous PA for at least 75 minutes a week, or an appropriate combination of moderate and vigorous PA (González et al. 2017). Whilst the WHO only include moderate and vigorous PA, recent studies have demonstrated that light intensity PA also provides health benefits (Wannamethee and Shaper 2001; Pate et al. 2008; Calabró et al. 2014). Most of the activities of daily living (ADL) fall in the category of light intensity activities (Calabró et al. 2014).

2.2.1.3 PA and health

Regular PA has been shown to improve health outcomes. For example, this includes: reducing blood pressure and systemic inflammation; improving body composition, coronary blood flow and cholesterol levels (Warburton et al. 2006). Additionally, regular PA reduces the likelihood of 35 chronic conditions, such as cancer, musculoskeletal disorders, and cardiovascular diseases (Booth et al. 2012; Pedersen and Saltin 2015). The relationship between PA and chronic diseases

is a vicious circle: decreasing amounts of PA increases the likelihood of a person developing a chronic disease. The chronic disease further decreases the amount of PA a person can perform, thereby exacerbating their condition. Consequently, it is widely recognised that people with chronic disease(s) tend also to undertake lower levels of PA (Durstine et al. 2013).

Physical inactivity is the fourth leading risk factor for mortality as identified by the WHO (Veldhuijzen van Zanten et al. 2015). Physical inactivity is defined as any activity level insufficient to meet the current PA guidelines (Lee et al. 2012). Therefore, it is essential to know how physically active or inactive an individual is, since PA is a modifiable risk factor (Schrack et al. 2016). This means that with the appropriate guidance, individuals can improve their PA behaviour and adapt to a healthier lifestyle (Strath and Rowley 2018).

2.2.1.4 Assessment

The importance of accurately monitoring and increasing PA has been well documented across a range of long-term conditions. There is, therefore, a need to measure PA accurately and precisely to enable us to develop and evaluate programmes that aim to increase PA and to monitor patient's wellbeing.

Various methods have been developed to measure PA (Vanhees et al. 2005). Traditionally, PA has been assessed using subjective methods such as questionnaires. However, due to some of the well recognised limitations of subjective measures, technological advancements, and demand for better quality of life, objective methods are becoming increasingly more widely used (Taraldsen et al. 2012; Broderick et al. 2014). This section discusses in detail various methods for assessing PA by identifying their advantages and disadvantages.

Self-reports Questionnaires, such as the International PA questionnaire, and the Global PA questionnaire, as well as activity diaries are examples of subjective methods used to assess PA (Yang and Hsu 2010). They are simple to use, inexpensive and they offer the ability to assess sedentary behaviour, such as recreational, transport-related activity and occupational (Broderick et al. 2014). Another advantage of such methods is the ease of administration in large groups (Mynarski et al. 2012).

However, there are several limitations to these subjective measures which limit their use in clinical populations. They are reported to provide inconsistent assessment results since they depend

on subjective interpretation and individual observation (Mynarski et al. 2012; Broderick et al. 2014), and they are prone to recall bias since people tend to report socially desirable outcomes rather than true outcomes. Questionnaires, in particular, may also be cultural- and age-specific which means that they cannot be used in different countries and populations (Taraldsen et al. 2012; Miller et al. 2013). There are also specific issues with these methods amongst older adults, which is particularly important given the prevalence of multiple NCDs amongst this population. Older adults might fail to remember exactly which activities they carried out throughout their day (Taraldsen et al. 2012; Miller et al. 2013; Sylvia et al. 2014). Another key limitation is that they are not very good at capturing low intensity activities which is where people with chronic conditions get most of their PA from (Miller et al. 2013). This is because most questionnaires include questions about moderate and vigorous physical activity tasks (Sylvia et al. 2014).

Direct calorimetry Human energy metabolism involves the transformation of energy from the combustion of fuel in the form of carbohydrate, protein, fat, or alcohol. In this process, oxygen (O_2) is consumed and carbon dioxide (CO_2) is produced via an exothermic reaction. The measurement of energy expenditure (EE) involves the measurement of heat loss for each subject using a calorimeter, which is referred to as direct calorimetry (Ndahimana and Kim 2017). This technique is used to quantify metabolic rate and it is often used to validate other objective and subjective methods. However, since this method requires direct observation, complete assessment is time consuming and requires a lab setting using expensive equipment, so it prohibits real world assessment and is not suitable for large-scale studies (Vanhees et al. 2005).

Indirect calorimetry The indirect calorimetry technique measures respiratory gas volume, such as O_2 and CO_2. This can be done using several methods, for example face mask, canopy and Douglas bag (Ndahimana and Kim 2017). Like direct calorimetry, this method is expensive and requires technical expertise. Additionally, the assessment must again be performed in a lab setting since the user should be connected to the machine. It is non-invasive and produces accurate results (Ndahimana and Kim 2017) but not well suited to large or real-world studies.

Doubly labelled water Doubly labelled water (DLW) is the "gold standard" technique that is used to assess total EE (Ndahimana and Kim 2017). This method uses the stable isotopes of water to assess EE, water flux, and body composition. It follows the exponential disappearance

of the oxygen (^{18}O) and hydrogen (^{2}H) stable isotopes in body water after initial labelling of the body water pool. ^{18}O is lost from the human body as both CO_2 and H_2O, in contrast to ^{2}H that is lost in the form of water. Therefore, the loss difference between them is the production of CO_2 during that period (Westerterp 2009; Buchowski 2014). Even though this method is currently the gold standard, it has several disadvantages. First, it is an expensive technique and unsuitable for large-scale studies. Using this method, the EE is only estimated and therefore a distinction between the EE of PA and basal metabolic rate, as well as diet-induced EE is not possible (Butler et al. 2004; Vanhees et al. 2005). In addition, it does not provide information about PA over timescales of days or weeks (Hills et al. 2014).

Pedometry Pedometers are electromechanical or electronic devices that are commonly used to count the number of steps taken throughout a day. An up and down movement occurs when a human is walking. Typically, within a pedometer, a lever arm is stimulated to oscillate by the person's steps. Each oscillation occurs when a step is taken and there is corresponding movement at the hip; therefore the pedometer counts the steps taken during human walking (Wise and Hongu 2014). In the clinical setting, pedometers can be used to improve or support the daily activities of patients (Yang and Hsu 2010; Broderick et al. 2014). To achieve good results, the step count needs to be accurate so that any targeted interventions can be calibrated to an individual. Despite this, some important parameters of PA cannot be measured from a pedometer because of the way it works. For example, activity intensity, duration of individual bouts of PA and sedentary time cannot be measured. The pedometer provides an output in steps, which means that EE estimates will be inaccurate (Yang and Hsu 2010; Broderick et al. 2014; Trost and O'Neil 2014). This might be because small steps expend less energy than big steps, but both are considered equal by a pedometer. Furthermore, at slower speeds, the pedometer's accuracy is reduced (Broderick et al. 2014) because most of the (pedometer) algorithms are based on thresholds that were developed in a healthy population that walk normally, and at a normal pace. For example, the amplitude of vertical acceleration is reduced at slower speeds, hence the threshold may not be representative to count accurately each step (Ehrler et al. 2016).

Accelerometry In terms of human movement, acceleration data can reflect the frequency and intensity of motion, since acceleration is directly proportional to external force when the measured mass is constant. Additionally, several parameters such as vibration frequency, rotation,

and tilt can be derived from an acceleration signal. Tilt sensing can be used to identify different body postures of the wearer. These sensors have become practical and useful in wearable devices to assess PA since they have smaller size with lower power consumption, and therefore they have been used widely (Yang and Hsu 2010). They can be used for human activity recognition (HAR), for estimating EE (Hills et al. 2014; Sazonov et al. 2014; Santos-Lozano et al. 2017) and for detecting falls (Castillo et al. 2014). Due to the features offered by accelerometry, this approach has become one of the most popular choices to use for measuring PA objectively. One of the most important features of the accelerometer approach is the ability to provide information about the four dimensions of PA; frequency, intensity, time and type (FITT).

Raw accelerometry - limitation Even though accelerometers are currently one of the best options to measure PA in real-life, they still have limitations (Atallah et al. 2011; Gjoreski et al. 2016). The sensor location affects the accuracy of the recorded signal. The most common sensor mounting locations are wrist, ankle, chest, waist and hip. They can be directly attached to the skin, clipped to clothing, or they can be carried in pockets. Theoretically, the sensor might be placed slightly differently each time it is attached by the wearer, hence producing slightly different outcomes. The main drawback of accelerometers is that they might not provide sufficiently detailed information or accurately classify different intensity activities on their own. Other sensors, such as gyroscopes, microphones and electrocardiogram (ECG) sensors, along with accelerometers can be integrated to provide more informed results (Atallah et al. 2010).

Accelerometers embedded in wearables - limitations Commercially available wearables often use proprietary algorithms to measure PA, which creates some limitations. The first is that the underlying algorithms are usually not available to either researchers or consumers (only the resulting PA). This means that results estimated by devices from different manufacturers cannot be easily compared or validated (Mancuso et al. 2014). The second problem is that even when open source algorithms are used, many algorithms are targeted towards a specific sub population, and do not work well in other groups (Backhouse et al. 2013; Mancuso et al. 2014). Therefore, the algorithms might be less accurate for someone with a different walking pattern and slower pace than the usual. It is important to be able to measure these activities as well, since most of the primary ADLs are of low intensity (Pate et al. 2008; Calabró et al. 2014). Another issue of accelerometers is that they only provide EE or activity counts based on intensity count thresholds. These counts are used to represent a sum of acceleration, which was

counted as an activity having exceeded a threshold. The sum of acceleration values represent an acceleration into an epoch, which is a fixed recording interval. Thus the data typically describes general PA levels, rather than specific types of activities (Lipperts et al. 2017). Table 2.2 summarises the advantages and disadvantages for each method.

Table 2.2: Advantages and disadvantages of physical activity assessment methods.

Method	Advantages	Disadvantages	References
Doubly labelled water	Non-invasive Precise EE results Accurate EE results	Expensive Time-consuming No contextual information Does not quantify FITT Unsuitable for large-scale studies Needs technical expertise No indication of specific activites	(Vanhees et al. 2005; Yang and Hsu 2010; Strath et al. 2013; Sylvia et al. 2014; Hills et al. 2014; Ndahimana and Kim 2017)
Indirect calorimetry	Accurate EE results Precise EE results Non-invasice	Needs technical expertise Expensive Short time assessment Limited to lab-setting	(Vanhees et al. 2005) (Yang and Hsu 2010; Strath et al. 2013; Ndahimana and Kim 2017)
Direct calorimetry	Accurate metabolic equivalent results	Expensive Subject confindement for 24+ hours	(Ndahimana and Kim 2017)
Pedometry	Easy to use Low cost Accurate for running and moderate walking Low burden to participant	Does not quantify FITT Influenced by placement Walking or running specific, no upper movements Inaccurate EE results	(Vanhees et al. 2005; Yang and Hsu 2010)

Continued on next page

Table 2.2 – continued from previous page

Method	Advantages	Disadvantages	References
Pedometry	Small size Portable Non-invasive Lightweight Large-scale application Motivation tool Suitable in free-living	Proprietary algorithms	(Butte et al. 2012; Strath et al. 2013; Sylvia et al. 2014; Hills et al. 2014; Trost and O'Neil 2014; Ndahimana and Kim 2017)
Accelerometry	FITT information Concurrent measure of movement Captures large amount of data EE results Activity counts (intensity) Cost-effective	Needs technical expertise Inter- and intra-monitor variability Thresholds influence measurements of PA intensity Proprietary algorithms Influenced by location Influenced by attachment method	(Culhane et al. 2005; Vanhees et al. 2005; Westerterp 2009; Yang and Hsu 2010; Butte et al. 2012; Strath et al. 2013; Sylvia et al. 2014; Hills et al. 2014)
	Long-term monitoring Suitable in free-living Step count Small size Wireless Non-invasive Lightweight Portable Low burden to participant	Influenced by external vibrational artefact No contextual information	(Trost and O'Neil 2014; Schrack et al. 2016; Ndahimana and Kim 2017)
Self-report	Low cost Accurate for intense activities	Recall Inaccurate for light/moderate intensity activities	(Vanhees et al. 2005) (Westerterp 2009)

Continued on next page

Table 2.2 – continued from previous page

Method	Advantages	Disadvantages	References
Self-report	Ease of administration	High burden on participants (diary)	(Yang and Hsu 2010; Strath et al. 2013; Sylvia et al. 2014; Hills et al. 2014; Bassett et al. 2015; Ndahimana and Kim 2017)
	Determining discrete categories of activity level	Subjective interpretation	
	Simple to use	Low reliability and validity	
	Capture contextual information	Need to be population and culture specific	
	Large-scale application	Less robust in assessing EE	

2.2.1.5 Summary

Increasing PA offers many benefits for people with chronic conditions. To help develop strategies to improve PA in this population and monitor progress, it is important to be able to accurately measure the activity of patients over time and in the real world. This would help to manage the diseases more effectively, to monitor progress of chronic diseases and to promote healthy behaviours in populations of people at increased risk of negative effects associated with physical inactivity (Tao et al. 2012; Lipstein et al. 2016).

Subjective measures have been traditionally used to assess PA since they are inexpensive and can be easily administered in large groups (Mynarski et al. 2012; Broderick et al. 2014). Most common subjective methods depend on the memory of the participant, and on subjective interpretation (Taraldsen et al. 2012; Mynarski et al. 2012; Miller et al. 2013; Broderick et al. 2014).

Objective techniques provide more accurate measurements. Of these techniques, accelerometers are most commonly used to assess PA. The reason for that is their ability to automatically, continuously and for long term period to measure PA of participants in both free-living and lab-based environments. All four PA dimensions can be measured using accelerometers (Yang and Hsu 2010). Therefore, the next section will explore accelerometers in more detail. For example: a) the different components of the acceleration; b) the underlying mechanism of the accelerometer; c) the morphology of the acceleration signal while in motion and rest; and d) the variability of the signal while walking.

2.3 Accelerometry

2.3.1 Definition

The definition of acceleration is the rate of change in velocity over a given time. An accelerometer is a motion sensor that measures accelerations of a body along a sensitive axis (Hills et al. 2014). Acceleration is measured in either gravitational units (g) or SI units (m/s^2) (Hills et al. 2014). It can measure acceleration in up to three orthogonal axes (Mathie et al. 2004), that are sometimes called vertical, anterior-posterior, and medio-lateral (Godfrey et al. 2008; Tao et al. 2012).

The output signal from an accelerometer represents the overall acceleration and/or deceleration of the body on which the sensor is attached (Strath et al. 2013). The signal is composed of three main components; gravity, movement and noise (Hees et al. 2013). The first component is associated with gravitational acceleration, the second component is associated with the movement of the subject and the third component is associated with external vibrations acted on the body and from the movement of soft tissue (Mathie et al. 2004).

When the accelerometer is static, the resultant signal is entirely due to gravity. This means that the total acceleration measured is 1g. The gravity component can be used to identify the orientation of the accelerometer. When the accelerometer is moving, a change in the output acceleration signal is detected. The changes in the signal are influenced by the intensity of the motion of the accelerometer. When the movement is more intense, the resulting accelerations are greater and a bigger change in the signal is detected (Gjoreski and Gams 2011). This can be seen in Figures 2.1 and 2.2 which represent the activities under both normal and simulated-pathological conditions respectively.

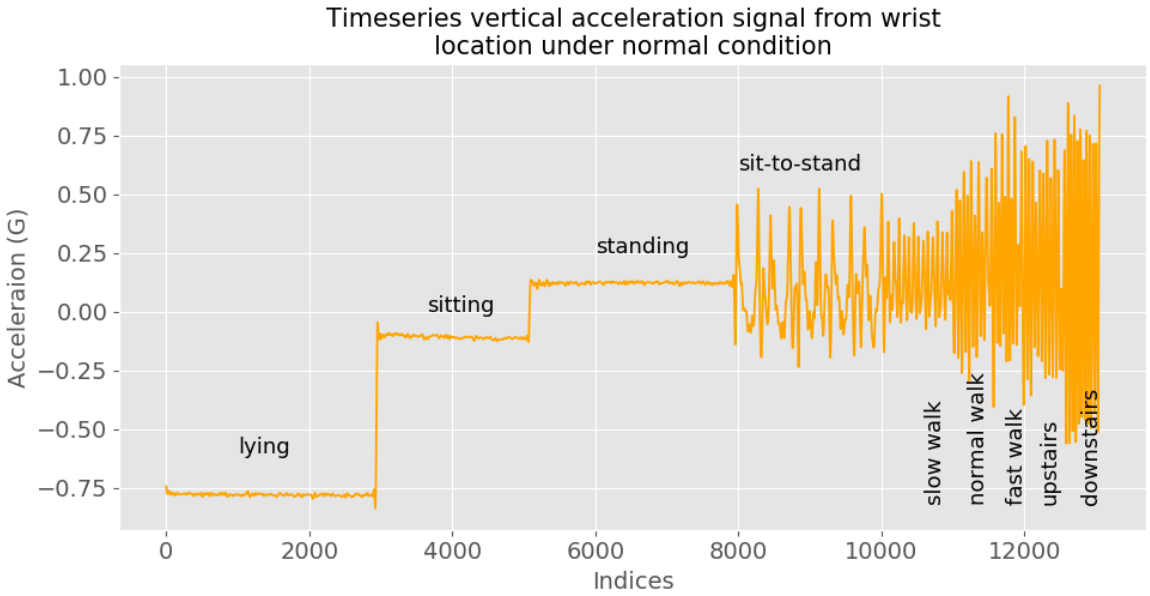


Figure 2.1: Vertical acceleration signal collected from the wrist representing the following daily activities under normal condition: lying, sitting, standing, stand-to-sit, slow walk, normal walk, fast walk, ascending stairs and descending stairs

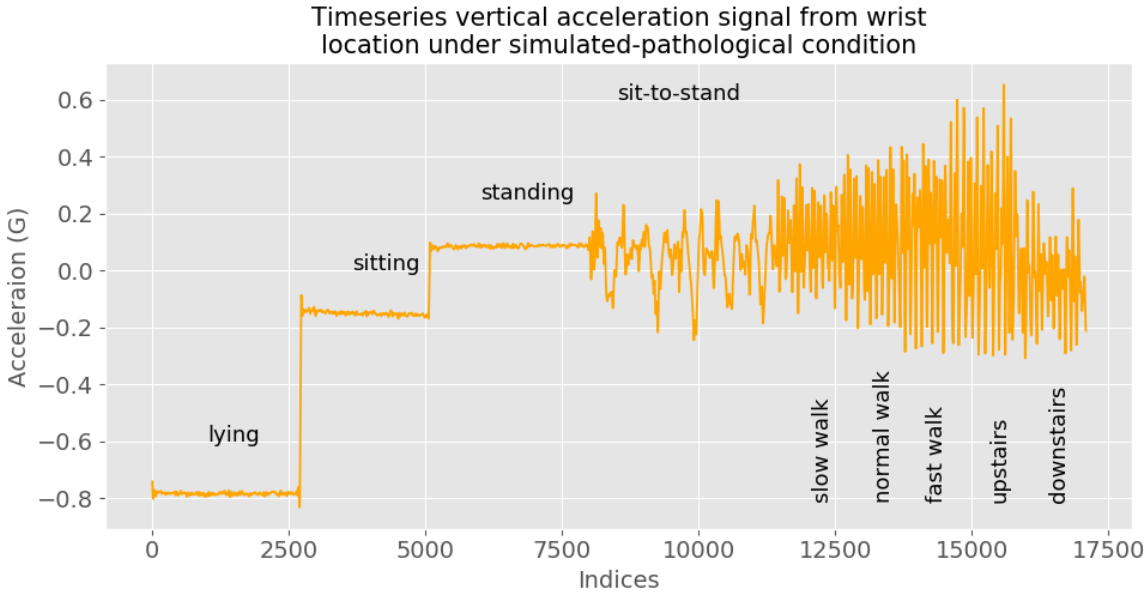


Figure 2.2: Vertical acceleration signal collected from the wrist representing the following daily activities under simulated-pathological condition: lying, sitting, standing, stand-to-sit, slow walk, normal walk, fast walk, ascending stairs and descending stairs

2.3.2 Accelerometer system function

There are three main types of accelerometers; piezoelectric, piezo-resistive and capacitive (Shany et al. 2012). Although the accelerometers are built based on different manufacturing techniques and designs, they all use the same underlying basis, which is a mass-spring system (Mathie

et al. 2004). In other words, accelerometers work based on Newton's 2nd law and Hooke's law (Kavanagh and Menz 2008).

Newtons 2nd law:

$$F = m \times \frac{v_1 - v_0}{t_1 - t_0} = m \times \alpha \tag{2.1}$$

Hooke's law:

$$F = k \times x \tag{2.2}$$

where m is mass, v is velocity, t is time, α is acceleration, k is the spring's stiffness and x is distance.

The mass-spring system reacts when a compressive force is applied. This reaction produces a proportional force to the initial force acted on the system. Based on that principle, acceleration can be calculated according to the system's displacement when mass and spring constant are known (Gomes 2014).

One common type of accelerometer is the capacitive, which is composed of microstructures that are built into a polysilicon surface (Lingesan and Rajesh 2018). The sensor contains differential capacitors that comprised with fixed independent plates, and plates attached on the moving mass. When the sensor moves, the mass generates a reaction force that is applied to the springs. As shown in Figure 2.3, the springs are attached at the anchor of the system. The accelerometer sensor obeys the mass-spring system, described previously. Hence, it has been validated that the deformation of the spring is linear with acceleration (Jarchi et al. 2018). Due to the reaction force while the sensor is moving, an electrical output signal is produced by the differential capacitors. The signal is proportional to the magnitude of the acceleration acting on the sensor (Kavanagh and Menz 2008).

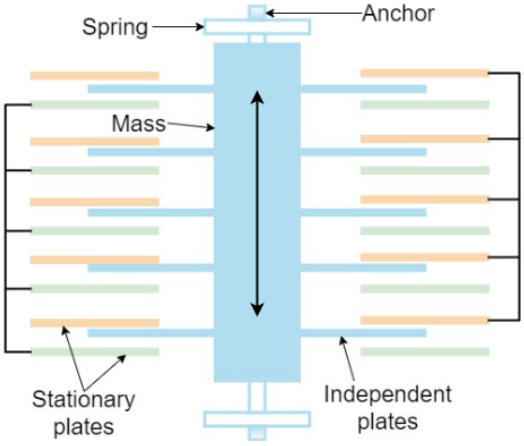


Figure 2.3: Representation of a general capacitive accelerometer

2.3.3 Acceleration signal morphology

Tri-axial accelerometers are used in wearables. This means that they can be placed near or on several body locations, such as wrist, ankle, thigh, waist, pockets, etc. These particular locations, that are associated with wearables, are used because they have common characteristics in both women and men: a large continuous surface, and low flexibility and movement (Gemperle et al. 1998; Yang and Hsu 2010; Mancuso et al. 2014). The location is often chosen depending on the patients' condition and their requirements (Mancuso et al. 2014). For example, wearables are often worn on the wrist to maximise patient convenience. However, accelerometers worn on the upper part of the human body tend to slightly underestimate PA in comparison to the wearables that are worn on the lower part of the human body, ankle, and shoe (Mancuso et al. 2014; Walker et al. 2016). In addition, patients might have some physical impairments as a result of the illness, surgeries, or therapies that restrict their movement in the limb that they have to wear the monitor (Walker et al. 2016). Hence, the accelerometer output is influenced by its position and orientation, the posture of the wearer and the activity performed (Godfrey et al. 2008).

2.3.3.1 Static activities

As mentioned above, when the accelerometer is static, the acceleration value should be zero. However, the accelerometer is affected by gravity. Hence, if the accelerometer is placed so that its vertical dimension is perpendicular to the ground, the measured acceleration value will be 1g. Due to noise, the acceleration will not be exactly 1g (Figure 2.4). Figure 2.4 demonstrates the three-dimensional (3D) acceleration while standing, where x is the vertical dimension, y

is medio-lateral dimension (back-forth) and z is antero-posterior dimension (right-left). The vertical dimension (up-down) is affected by gravity, therefore it is the furthest from zero. The medio-lateral dimension is almost zero because there is no motion in this plane. And the antero-posterior dimension is slightly off from zero because it might not be aligned with ground.

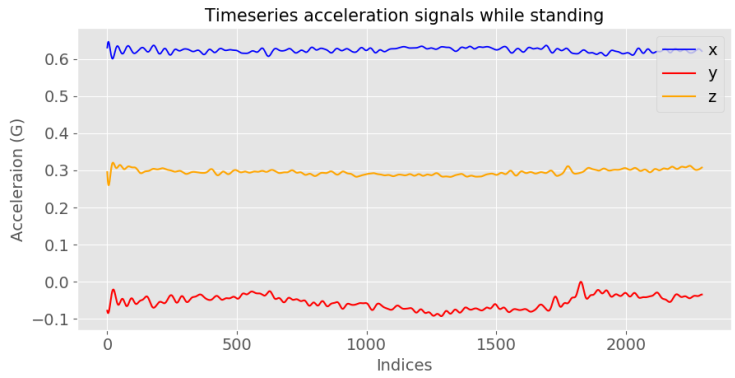


Figure 2.4: Vertical (x), medio-lateral (y) and antero-posterior (z) acceleration signals representing standing activity

2.3.3.2 Dynamic activities

Gait is characterised as a series of alternating rhythmical movements of legs, arms and trunk, where a forward body movement is created (Anwary et al. 2018). Human walking can be represented as a gait cycle, which is composed of two consecutive steps (Cola et al. 2017). Walking is a repetitive process of multiple gait cycles (Tao et al. 2012), and can be derived from anterior-posterior and vertical accelerations. These two directions are responsible for most of the total power of the signal (Butte et al. 2012). In terms of acceleration signal, each gait cycle can be observed as a series of deflections away from the baseline (gravitational) value. These deflections are represented by maximum peak values (Anwary et al. 2018) and reflect the arm and leg movement during walking. The peaks are created because of the foot contact on the ground at each step (Cola et al. 2017).

As the foot contacts the ground, an impact is created and transmitted to the sensor using the human body as a medium. The acceleration signal has greater magnitude and intensity when the accelerometer is attached on the foot in comparison to the signal created when the sensor is attached on the wrist. The reason for this is because the accelerometer is farther from the foot (Cola et al. 2016). Additionally, the magnitude of the deflections of the acceleration signal, representing two consecutive steps, is influenced by the side of the body where the sensor is

attached. Higher acceleration values are recorded for the steps that were made with the leg closer to the accelerometer. In terms of the wrist accelerometer, the acceleration magnitude is also influenced from the arm swing, which might reduce the amplitude of the acceleration (Cola et al. 2017).

A gait cycle is composed of eight phases; heel strike, loading response, mid-stance, terminal stance, pre-swing, toe-off, mid-swing and terminal swing as shown in Figure 2.5. This cycle represents the behaviour of the legs during walking. In terms of the arms movement, they swing out of phase with the legs. Hence, while walking, the right side of the pelvis, the right leg, the left arm and the left side of the shoulder girdle move forward at the same time (Whittle 2007).

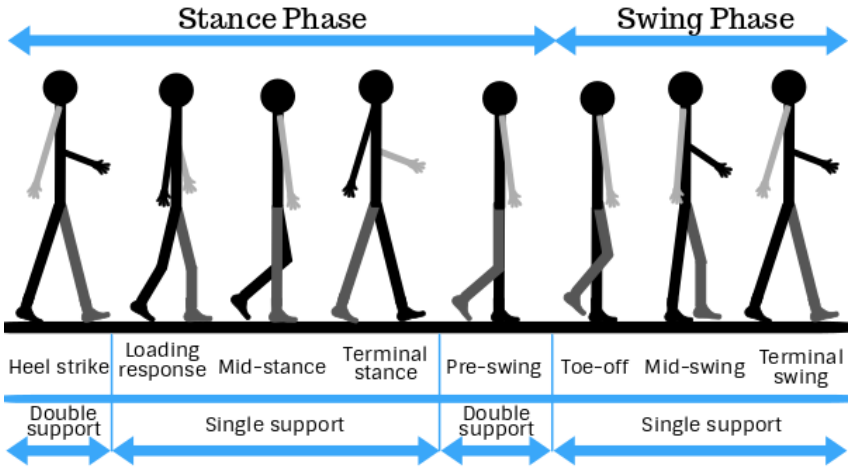


Figure 2.5: Phases of gait cycle

The walking acceleration signal may be divided into different sections. When a step is made using the leg that is on the same site with the sensor, the acceleration recorded is higher. The following are the sections of the acceleration signal derived from the wrist location, as shown in Figure 2.6:

- Forward valley: a deflection away from the baseline caused by the arm's direction, which is perpendicular to the ground and tends to move forward
- Forward peak: a deflection away from the baseline caused by the arm's direction, which is beyond the user's hip
- Stance-point: a deflection away from the baseline caused by the arm's direction, which is perpendicular to the ground and tends to move forward

- Backward peak: a deflection away from the baseline caused by the arm's direction, which is behind the user's hip

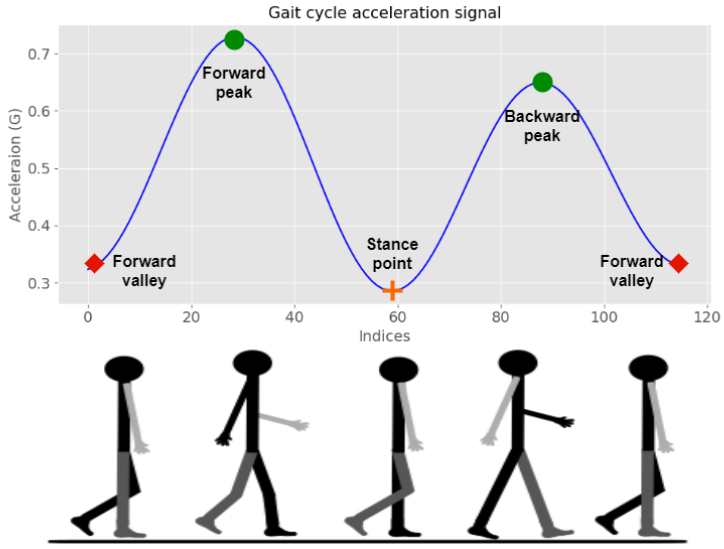


Figure 2.6: Acceleration signal of a single gait cycle with important points

2.3.4 Walking variability

There is intra- and inter-variability of gait between humans while walking. This means that each walking step varies in length, time, and width between steps from the same individual (intra-) and between individuals (inter-) (Collins and Kuo 2013). In terms of the gait cycle, this translates in variations of amplitude, period and pattern in the acceleration signal.

Several factors might be responsible for walking variability. For example, the nervous system, the structure of the physical body and the floor surface which is often not completely flat. The conscious choice of walking speed also leads to variability of walking. Humans can do several things while walking, for example carrying objects or having their hands in pockets, therefore their speed is influenced by additional activities they perform (Collins and Kuo 2013).

Walking variability exists in both "normal" and "pathological" gaits. Regarding the former, a more consistent walking pattern is visible. Regarding the latter, the walking pattern might become less visible and with greater amount of noise (Kirtley 2006).

2.4 Human activity recognition

Human activity recognition (HAR) is an area of research that is used to "read" motions and gestures of the human body. For instance, standing, walking, running, texting, etc. Sensors, like accelerometers, heart-rate monitors, and global positioning systems (GPS) are used on and around the subject's body to identify any activity performed (Ann and Theng 2014). HAR provides useful information about the behaviour of the user, such as if the user is active or not. Devices are then used to assist the user proactively while carrying out their tasks (Bulling et al. 2014). Generally, the area of activity recognition has become more popular since wearables and smartphones can be used to recognise activities. HAR is used in healthcare research for two main reasons: 1) to assist patients with chronic diseases and 2) to support diagnosis of patients by detecting anomalous behaviours or tracking health conditions (Bulling et al. 2014; Banos et al. 2014).

2.4.1 HAR challenges

Though HAR is commonly used by researchers, several challenges remain (Bulling et al. 2014). The first challenge is to ensure robust placement of sensors to maximise effective activity recognition and step counting. Some sensors are sensitive to orientation and position. This might be an issue, since it is not possible to always place the sensor at the exact same position (Avci et al. 2010; Atallah et al. 2011; Ann and Theng 2014). However, a device that is located on the wrist can be expected to be consistently worn with the same orientation with respect to the arm of the user. Therefore, accelerations from the local coordinate system of the 3D accelerometer can be used for gait analysis (Cola et al. 2016). Another challenge is human variation (Avci et al. 2010). Like walking, other activities may be performed differently by each individual, or even within the same individual from time to time. This can be classified as intra-class variability (Bulling et al. 2014). A similar problem is inter-class similarity, in which different activities may have similar characteristics that make them difficult to differentiate (Bulling et al. 2014).

2.4.2 Activity recognition chain

The activity recognition chain (ARC) is a framework that is used to design and evaluate HAR systems. Specifically, it is a sequence of pattern recognition, signal processing and ML techniques. It consists of six elements: data acquisition, data pre-processing, data segmentation,

feature extraction and selection, training and classification, and performance evaluation (Bulling et al. 2014).

Figure 2.7 demonstrates the ARC process, which was adopted for the analysis of this thesis. From left to right, a set of N sources, in our case accelerometer sensors, delivers raw acceleration signals. The pre-processing part consists of several sub-parts, such as labelling the data, filtering to reduce noise and calculation of other measures derived from the collected acceleration data. The next step is to segment the signals into windows of given length to capture the dynamics of the signals. Subsequently, several features are calculated for each window using a feature-extraction process. In terms of features, time- and frequency-domain functions are often used. When all features are calculated, a feature vector is formed and it is used as an input to the classifier (Banos et al. 2013).

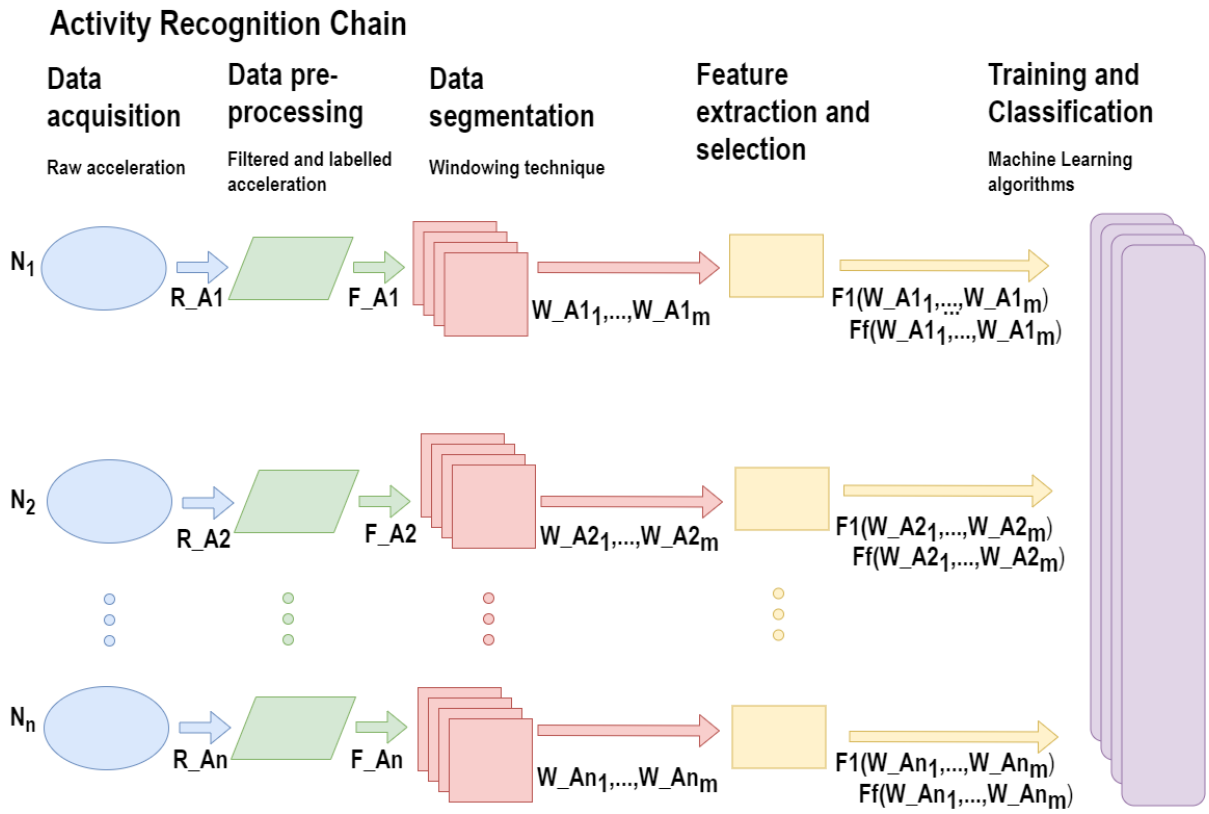


Figure 2.7: Activity recognition chain process (Banos et al. 2014)

2.4.2.1 Data acquisition

The first step to create a robust activity recognition system is to design and develop a process to collect comprehensive data about the target group. These data can take several forms, attaching sensors on different body locations of the participants and/or positioning sensors in

the environment (lab or home) to measure participants' movements. In addition to sensor data, demographic characteristics and sensor device specifications ought to be acquired (Saez et al. 2016).

Accelerometers might have up to three axes, e.g. 1D, 2D and 3D. This means that a sensor with three axes provides data from three different directions. For example, an accelerometer can record 3D acceleration in x, y and z direction. In general, the output of the sensors can be described as:

$$N_i = (d_{i1}, d_{i2}, d_{i3}, \ldots, d_{it}), for\ i = 1, \ldots, n \tag{2.3}$$

where n denotes the number of sensors, and d_{ij} the multiple values at a time t. The sampled data is recorded at regular intervals by the sensors. The result is to provide a multivariate time series. Sometimes the sampling rates of different sensor types are different, therefore synchronising across multimodal sensor data is an important step before performing any other action.

2.4.2.2 Data pre-processing

After successfully collecting data, the data is pre-processed to extract useful information. Often, the raw signal recorded by the sensors contains artefacts from aquisition, noise, missing samples or invalid data. Therefore, this stage is required to remove any artefacts, to synchronise and to prepare the captured signals for the next stage. Signal pre-processing is critical since it is essential to preserve signal characteristics in terms of the activity data. The processed data should retain the important information of the data, but with no artefacts.

Depending on the type of data, different pre-processing methods are used. The reduction of noise and artefacts is often carried out using different filters. Filters that might be used for this purpose are low-pass Butterworth and median filter (Hassan et al. 2018). The low pass filter blocks high frequencies, and passes low frequencies up to a specified cut-off frequency threshold. A median filter is often used to smooth the data between different windows (Wang et al. 2011).

Normalisation is a common method used to pre-process acceleration signals (Bulling et al. 2014). It is used to scale heterogeneous data into "new" comparable data. At this stage any missing values are estimated, which is a hard problem to resolve. For very short pieces of missing data

a 3-point median filter might be used to average the previous and the next values of the missing value (Saez et al. 2016). However, there are other more complex ways to fill in missing data when bigger chunks are missing, such as imputation with k-nearest neighbour or C4.5 decision tree (Batista and Monard 2003). Additionally, labelling the data is an important pre-processing step since the labels are used for the ML algorithms and also for removing unwanted data.

Finally, in some cases, the data recorded by the sensors, such as acceleration, can be used to calculate other types of signal, for instance velocity and jerk signals, using mathematical equations. Usually, this is done since there is no access to sensors that can directly measure these attributes.

2.4.2.3 Data segmentation

The sensor signals are divided into smaller data sections, called windows, using different segmentation techniques. Ideally, each window is short enough so that only one type of activity occurs during the period (Preece et al. 2009; Banos et al. 2014; Bulling et al. 2014). A window is defined by its start and its end time within the signal's time series. Several features are calculated for each window, and then they are used for the characterisation of the collected signal. After this step, algorithms are developed to classify the dataset. The inputs for these algorithms are the extracted features. Segmentation techniques can be separated in three different categories, sliding windows, event-defined windows and activity-defined windows (Preece et al. 2009; Banos et al. 2014). In a sliding window approach, the signal is separated into fixed size windows, with or without overlap windows. The event-defined window approach locates different events that are then used to segment the data. The final approach, activity-defined, separates the data when detecting variations in activity (Banos et al. 2014). Segmentation is performed to examine short sections of the time series signal. For instance, this is useful when the time series may contain multiple different activities, and you want to be able to work out when each activity happened (Bulling et al. 2014).

This process might sound straightforward and easy to perform, but it is a complicated task in practice. Activities performed by humans are diverse and complex. One activity can be executed with different approaches which can lead to ambiguity in the manual labelling of the activity (Bulling et al. 2014). Moreover, activities are carried out continuously and simultaneously. This results in signals where the activities are not clearly separated in time and makes it difficult

to isolate them. Another issue that arises is the definition of each activity. Since activities do not necessarily have start and end borderlines, an activity might be performed in two or more different ways, which makes it difficult to use a particular signal to characterise one activity. For example, an eating activity might start with reaching or holding the cutlery (Bulling et al. 2014).

2.4.2.4 Feature extraction and selection

Signals are reduced into features. Features are sets of numbers that have been derived from the time- or frequency-domain to describe the raw signal (data) in different forms. The features can then be used as input to ML algorithms. Selection of the right features is important as well because they provide different signal characteristics. Incorrect feature selection may lead to incorrect classification of activities (Lara et al. 2012). It is often necessary to simplify analysis of a highly complex signal. Further analysis is conducted on the features (signal's characteristics), rather than on the raw acceleration signal; therefore, the selection of relevant features is important (Ignatov 2018). A "good" feature needs intra-class reliability of an activity and to be robust between different people. In doing so, instances of the same activity will be closely clustered in the feature space.

A wide range of features for accelerometer data have been identified from the literature (Gjoreski et al. 2016; Hassan et al. 2018). The features may be classified as time-domain features, frequency-domain features, or other features (Lara and Labrador 2013). Some of the most common features are shown in Table 2.3.

Table 2.3: Common features used in the activity recognition literature.

Domain	Methods	References
Time	Mean	
	Standard deviation	(Suto et al. 2016)
	Variance	(Li et al. 2018)
	Interquartile range	(Waltenegus 1999)
	Entropy	(Nayak and Panigrahi 2011)
	Kurtosis	(Gupta and Dallas 2014)
	Time between peaks	
	Maximum peak amplitude	
Frequency	Spectral energy	(Nayak and Panigrahi 2011; Suto et al. 2016)
	Discrete cosine transform	(Li et al. 2018)

Often, the time-domain features represent statistical measures, such as mean, standard de-

viation, variance, etc. (Preece et al. 2009). On the other hand, frequency-domain features are derived by transforming the time series data. The reason for exploring the signal in both time- and frequency-domain is to identify different characteristics of the features. For example, time-domain analysis describes how the signal changes over time. Frequency-domain analysis describes how the energy of the signal is distributed in different frequencies. There are different methods to transform the signal from the time-domain to the frequency-domain. The methods are identified based on which category the signal of interest falls into. In general the signals can be either: a) continuous or discrete, and b) periodic or non-periodic (Smith 1999). Therefore, four categories are formed, 1) *periodic & continuous*, 2) *non-periodic & continuous*, 3) *periodic & discrete*, and 4) *non-periodic & discrete*. Depending in which category the signal is into, the appropriate methods might be used to transform the signal to the frequency-domain. For example, Fourier series methods are used for the *periodic & continuous* category, Fourier Transform methods are used for *non-periodic & continuous*. Discrete Fourier Transform (DFT) methods are used for *periodic & discrete*, and lastly Discrete Time Fourier Transform methods are used for *non-periodic & discrete* (Smith 1999). Since the signal of interest is the acceleration while performing activities that provide an almost periodic signal, the desired category is *periodic & discrete*. The static activities, sitting, standing, and lying, produce an almost flat acceleration signal. The same methods could be used however, because the frequency of the signal could still be extracted using this method. This means that DFT methods are of interest (Mertins 1999). All the different DFT methods are based on similar mathematics, therefore the Fast Fourier Transform (FFT) is often used since it is computationally efficient. FFT uses the divide and conquer algorithm to compute the DFT, hence it is faster than the original DFT. The FFT calculates the coefficients that produce the frequency components of the signal in terms of amplitude and phase of the signal (Preece et al. 2009).

Features related to spatio-temporal gait parameters, such as time between peaks or maximum peak, are often used to develop step counters (Kang et al. 2018).

2.4.2.5 Training and classification

When the features have been extracted, the next step is to train an algorithm on the labelled data to predict an outcome (e.g. typically an activity). For accelerometers, algorithms are typically used to classify the type and intensity of activities performed. Machine learning algorithms in conjunction with the advancements of inertial measurement unit (IMU) sensors

has made HAR using these techniques an increasingly popular research area. Data scientists use sensory data for analysis, and then developers use the data to develop smart-watches and mobile applications (Ogbuabor and La 2018).

In statistics and ML, classification is a problem in which we try to identify to which set of classes a new data is part of, based on a training dataset whose classes are known. These problems are solved using supervised learning algorithms. An algorithm that performs classification is known as a classifier. Several supervised learning algorithms are used for classification problems. Throughout the literature, several ML algorithms have been used for activity classification, as shown in Table 2.4.

Table 2.4: Machine Learning algorithms used for activity classification in the literature.

Algorithms	References
k-Nearest Neighbour	(Saez et al. 2016; Gjoreski et al. 2016; Ponce et al. 2016)
Gaussian naive Bayes	(Saez et al. 2016; Gjoreski et al. 2016; Ponce et al. 2016; Cleland et al. 2013)
Linear discriminant analysis	(Saez et al. 2016)
Stochastic gradient descent	(Saez et al. 2016)
Support vector machine	(Saez et al. 2016; Ponce et al. 2016; Hassan et al. 2018; Abdull Sukor et al. 2018; Gjoreski et al. 2016; Strath et al. 2015; Cleland et al. 2013)
Decision tree	(Saez et al. 2016; Ponce et al. 2016; Chernbumroong et al. 2011; Abdull Sukor et al. 2018; Gjoreski et al. 2016; Cleland et al. 2013)
Random forest	(Saez et al. 2016; Ponce et al. 2016; Gjoreski et al. 2016; Sasaki et al. 2016)
Mixture discriminant analysis	(Ponce et al. 2016)
Artificial neural network (ANN) e.g. Multilayer Percepetron (MLP)	(Ponce et al. 2016; Hassan et al. 2018; Chernbumroong et al. 2011; Abdull Sukor et al. 2018; Cleland et al. 2013; Montoye et al. 2016)
Convolutional neural network	(Jang et al. 2018; Ignatov 2018)

Each method has a set of parameters that can be used to tune the model. Depending on the parameters chosen, the model might be under- or over-fitted. When under-fitting occurs, the model cannot capture the underlying pattern of the data and the performance of the algorithm is poor. When over-fitting occurs, the model captures the noise of the data, which means that the algorithm fits the data well and the performance is good. However, when new unseen data is tested, the performance will be poor because it fits very well the trained data. To avoid

that, the data is separated into training, validation and test sets. These sets are used to avoid over-fitting because an amount of data is used to train the model, a different subset of data is used to validate the model, and different data again is used to test the model. This helps to test whether the developed algorithm is accurate or not using the parameters already set (Bulling et al. 2014; Saez et al. 2016; Korjus et al. 2016).Additionally, there are more strategies to limit the impact of overfitting. For example, it is important to train the algorithm with relevant and clean data. This ensures that the algorithm would identify general patterns that better represent the signals. Another strategy that can be used to avoid overfitting is to stop the training process of the algorithm early. This works because after a certain number of iterations, the algorithm stops being generalised and it starts to overfit the data (Kelleher et al. 2015). This can be identified by evaluating the model on the training dataset and then on a test dataset. If the performance of the training dataset is much better in comparison to the test dataset, then the model might be overfitted.

The outputs of the ML algorithms are predictions and these predictions are not 100% accurate, therefore performance metrics are used to determine to what extent the ML algorithms provide true results. Several performance metrics are available, such as confusion matrices, accuracy, sensitivity (or recall), specificity, precision, and receiver operating characteristic curves. Each metric provides different outcomes, therefore depending on the need, the most appropriate metrics can be used (Bulling et al. 2014).

A confusion matrix, as shown in Figure 2.8, is used to find the accuracy and correctness of the ML algorithm. True positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) are used in different calculations to calculate other performance metrics.

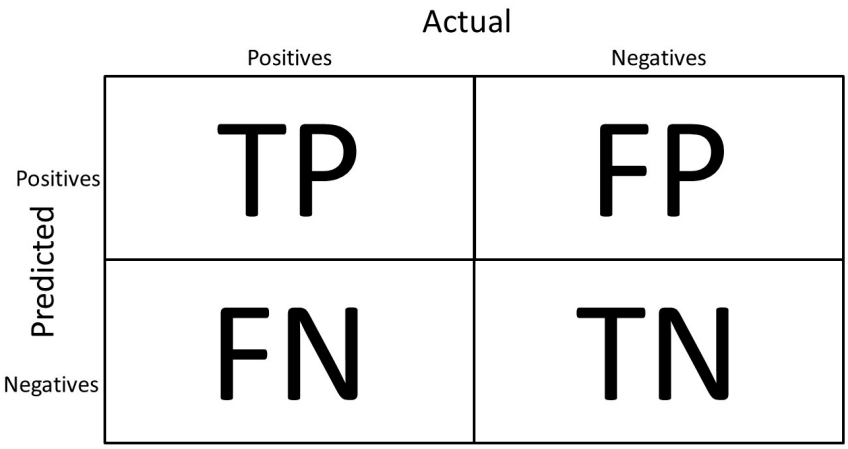


Figure 2.8: Confusion matrix

Accuracy is the number of correct predictions, positives and negatives, over all the predictions made by the model (Orphanidou and Wong 2017).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.4}$$

Sensitivity, or recall, measures the proportion between the true positives and the sum of true positives and false negatives. For example, the percentage of walking activity which is correctly identified as walking activity (Orphanidou and Wong 2017).

$$Sensitivity/Recall = \frac{TP}{TP + FN} \tag{2.5}$$

Specificity measures the proportion of true negatives that have been correctly identified as such. For example, the percentage of good quality signals which have been correctly classified as acceptable (Orphanidou and Wong 2017).

$$Specificity = \frac{TN}{TN + FP} \tag{2.6}$$

Precision measures the proportion between true positives and all the positives. For example, the percentage of good quality signals which have been classified as acceptable (Orphanidou and Wong 2017).

$$Precision = \frac{TP}{TP + FP} \tag{2.7}$$

Receiver operating characteristic curves are graphs that represent graphically the trade-offs between the specificities and sensitivities of the models (Orphanidou and Wong 2017).

2.4.3 Review of activity recognition

To perform activity recognition, it is essential to have available data first. This can be done either through primary data collection or by using a publicly available dataset. For the former, the research team recruits participants to collect their data while performing activities in the laboratory or free-living environment. For the latter, researchers have already collected raw signal data from participants, and made them publicly available to other researchers. Some of the publicly available datasets are: PAMAP2, JSI, FoS, Opportunity, WISDM, UCI and MHEALTH. Most of the data collected in the cases mentioned in Table 2.5, represent the healthy population, therefore there is a need to conduct studies that represent the patient population with physical impairments. These datasets were not considered for use in this thesis because of two main reasons: a) the number of participants was not large enough for the analysis of the pilot study, b) the datasets did not include variable walking speeds, which was a key feature in the hypothesis explored in this thesis.

Table 2.5: Information about datasets available online. Number of participants, activities performed and location of the sensors used.

Dataset	# of participants	Activities	Sensor location
PAMAP2 (Saez et al. 2016; Arif et al. 2017; Chowdhury et al. 2018)	9	Lying, Sitting, Standing still, Ironing, Vacuuming, Ascending stairs, Descending stairs, Walking outside, Nordic walking, Cycling, Running/jogging, Jumping rope, Playing soccer, Car driving, Watching TV, Folding laundry, Working on computer	Wrist(D), Chest, Ankle(D)

Continued on next page

Table 2.5 – continued from previous page

Dataset	# of participants	Activities	Sensor location
FoS (Gjoreski et al. 2016)	10	Cycling, Walking, Standing, Lying, Sitting, Running, On all fours, Kneeling, Bending, Transition	Chest, Thigh, Ankle, Wrist(R)
Opportunity (Gjoreski et al. 2016)	4	Standing, Sitting, Lying, Walking	Wrist(D),Wrist(ND)
JSI (Gjoreski et al. 2016)	5	Cycling, Walking, Standing, Lying, Sitting, Kneeling, Bending, Others	Chest, Waist, Thigh(R & L), Ankle(R & L), Upper arm(R & L), Wrist(R & L)
WISDM (Lee and Kwan 2018; Ignatov 2018)	36	Walking, Jogging, Ascending stairs, Descending stairs, Sitting, Standing	Smartphone
UCI (Ignatov 2018)	30	Standing, Sitting, Lying, Walking, Ascending stairs, Descending stairs	Waist (smartphone)
MHEALTH (Chowdhury et al. 2018)	10	Standing still, Sitting and relaxing, Lying, Walking, Climbing stairs, Waist forward bending, Frontal elevation of arms, Knees bending, Cycling, Jogging, Running, Jumping front and back	Wrist, Chest, Ankle

Most studies which collected data directly from participants had asked them to perform activities described in Table 2.5 above. The most common activities performed were: lying, sitting, standing, walking, ascending and descending stairs. The next most commonly performed activities were running, cycling and transition activities whereby participants transition from one activity to another, such as stand-to-sit.

As previously discussed, the location of the sensor plays an important role. Several studies used multiple sensor locations to identify which provides the best results in terms of activity recognition. Generally, the locations that were examined were: wrist, ankle, chest, waist, upper arm, thigh, hip and ear. The general outcome from most the studies was that activities are classified with better results if the activity monitor is placed near the body region performing the activity of interest. For instance, Arif and colleagues examined three locations, wrist, ankle and chest (Arif et al. 2017). For the walking activity, as expected, locating the sensor on the ankle resulted in the best outcomes, then chest and finally the wrist. For a complex ADL such as vacuum, cleaning, mounting the sensor on the wrist achieved more accurate results, followed by chest, while the poorest was the ankle.

A critical step for the ARC process is feature extraction. A feature is a measurable characteristic that is derived from the raw signal to provide better information about the signal of interest. Many of the articles calculated several features in the time- and frequency-domains, as shown in the Table 2.6.

Table 2.6: Features used for activity recognition using machine learning algorithms.

Time-domain		Frequency-domain
Mean	Interquartile range	Mean
Median	Pearson correlation	Median
Variance	Auto-regression coefficients	Entropy
Max	Signal magnitude area	Energy
Min	Signal vector magnitude	Mean energy
10^{th},25^{th},75^{th},90^{th} percentile	Harmonic mean	Max
Standard deviation	Number of peaks	Spectral centroid
Root mean square	Number of troughs	Skewness
Average distance between peaks and troughs	Difference in magnitude between max and min peaks	Wavelet coefficients
Skewness	Zero-crossings	
Kurtosis	Autocorrelation	
Time interval between local peaks		

In a few cases, where Convolutional Neural Networks (CNNs) were used, the researchers did not extract any features but used instead the raw acceleration signal (Jang et al. 2018; Ignatov 2018). In some other cases MLP was used which is one of the simplest Neural Network models. No complex deep learning models were used since they require a large amount of data in order to be used successfully (Jang et al. 2018). In the Attal study, two cases were examined as inputs

in the classifiers (Attal et al. 2015). In the first case, the input in the classifier was the raw acceleration signal, and in the second case, the input was a matrix of features. The performance metrics, such as accuracy, F1-score, precision and recall, achieved up to 4% better results when the input of the classifier was the features matrix rather than raw signal.

It is important to find the most appropriate features as input after their extraction because some of the features might not be informative. The features can be reduced by several methods, however one common method used is Principal Component Analysis (PCA) (Ponce et al. 2016; Hassan et al. 2018; Abdull Sukor et al. 2018). Other methods used were: Pearson's correlation (Gjoreski et al. 2016), sequential forward floating selection (Andreu-Perez et al. 2017) and wavelet transform (Arif et al. 2017).

The next step after feature extraction and selection is to train the ML algorithms. Mannini et al., Ponce et al., and Saez et al. trained more than ten different ML algorithms to examine which algorithms performed better for classifying activities (Mannini and Sabatini 2010; Ponce et al. 2016; Saez et al. 2016). Some of the approaches focused on identifying the best location to place the accelerometer to get the most accurate results. Other authors wanted to check which algorithm had the best performance for activity recognition and in some cases they used the data from all the locations together (Mannini and Sabatini 2010; Ponce et al. 2016; Saez et al. 2016). The results of the ML algorithms are influenced by several factors; a) the number of sensor locations used in the training dataset (Bao and Intille 2004; Cleland et al. 2013), b) the features used (Chernbumroong et al. 2011; Andreu-Perez et al. 2017), c) whether a feature reduction technique is used (Abdull Sukor et al. 2018), d) the validation technique (Ponce et al. 2016), e) the size of the training dataset (Saez et al. 2016) and f) the relation between the sensor location and the activities performed (Bao and Intille 2004; Cleland et al. 2013; Strath et al. 2015). Feature reduction techniques are often used to reduce the dimensions of the dataset, in order to make the data more significant and less sparse for the ML algorithms.

For instance, Saez et al., Arif et al., and Chowdhury et al. used the PAMAP2 dataset to perform activity classification (Saez et al. 2016; Arif et al. 2017; Chowdhury et al. 2018). The PAMAP2 dataset contained data from three sensor locations, nine subjects and 17 activities. Arif and colleagues developed one ML algorithm, a rotation forest, to classify all 17 activities (Arif et al. 2017). The wrist mounting location slightly outperformed the other two locations by achieving an F1-score of 93.1%. The chest location had the second highest F1-score of 92.3%

and ankle achieved a 92.2% F1-score. This paper also reported how the algorithm might perform when data from all three locations was combined. The results showed that the combination of data achieved much better outcomes, by achieving a 98.1% F1-score. The other two articles, only investigated 12 activities instead of all 17. The activities explored were: lying, sitting, standing still, ironing, vacuuming, ascending stairs, descending stairs, walking outside, Nordic walking, cycling, running/jogging and jumping rope. The two sets of results were in agreement that better performance was achieved when more than one location is used to perform activity classification (Chowdhury et al. 2018; Arif et al. 2017). For example, higher F1-score, such as 90.86%, was achieved when data from the three locations was used. Individually, the F1-scores achieved were 80.86%, 81.00%, 84.72% for wrist, chest and ankle locations respectively. In the Chowdhury study, ankle outperformed the other two locations. One of the reasons might be that when all 17 activities were used, the majority of the activities included were upper body oriented, but when the activities were reduced to 12 they were lower body oriented. Saez and colleagues demonstrated F1-scores from dataset that included all the locations (Saez et al. 2016). F1-scores between 91-96% were achieved to classify the same 12 activities that Chowdhury classified (Chowdhury et al. 2018).

Table 2.7 demonstrates the results of several studies that used ML algorithms to classify activities of daily livng.

Table 2.7: Results of studies that used ML algorithms to classify different activities

Study	ML algorithms	Activities	Sensor location	Validation	Results
(Saez et al. 2016)	kNN, Gaussian naïve Bayes, Linear Discriminant Analysis, Stochastic Gradient Descent, Support Vector Machine - linear, SVM - RBF, Decision tree, Random forest, Extra trees AdaBoost, DNN	Lying, Sitting, Standing still, Ironing, Vacuuming, Ascending stairs, Descending stairs, Walking outside, Nordic walking, Cycling, Running/jogging, Jumping rope	Wrist(D), Chest, Ankle(D)	LOSO CV	

Continued on next page

Table 2.7 – continued from previous page

Study	ML algorithms	Activities	Sensor location	Validation	Results
(Hassan et al. 2018)	ANN, SVM, DBN	Standing, Sitting, Lying down, Walking, Ascending stairs, Descending stairs, Stand-to-sit, Sit-to-stand, Sit-to-lie, Lie-to-sit, Stand-to-lie, Lie-to-stand	Smartphone		89.06%, 94.12%, 95.85%
(Chernbumroong et al. 2011)	C4.5, ANN	Standing, Sitting, Lying, Walking, Running	Wrist	5-fold CV	93.09%, 88.48%

Continued on next page

Table 2.7 – continued from previous page

Study	ML algorithms	Activities	Sensor location	Validation	Results
(Bao and Intille 2004)	Decision table, IBL, C4.5, Naive Bayes	Walking, Sitting & relaxing, Standing still, Watching TV, Running, Stretching, Scrubbing, Folding laundry, Brushing teeth, Riding elevator, Walking carrying items, Working on PC, Eating/Drinking, Reading, Bicycling, Strength-training, Vacuuming, Lying down & relaxing, Climbing stairs, Riding escalator	Upper arm, Wrist(D), Waist, Thigh, Ankle	LOSO	46.75%, 82.70%, 84.26%, 52.35%

Continued on next page

Table 2.7 – continued from previous page

Study	ML algorithms	Activities	Sensor location	Validation	Results
(Abdull Sukor et al. 2018)	Decision tree, SVM, MLP	Standing, Sitting, Lying, Walking, Ascending stairs, Descending stairs	Smartphone		96.85%, 92.87%, 100%
(Arif et al. 2017)	Rotation forest	Walking, Sitting, Standing, Laying down, Watching TV, Running, Folding laundry, Working on PC, Cycling, Strength-training, Vacuuming, Ironing, Climbing stairs, Nordic walking, Playing soccer, Car driving, Jumping rope	Wrist (D), Chest, Ankle (D)	10-fold CV,	93%, 92%, 92%

Continued on next page

Table 2.7 – continued from previous page

Study	ML algorithms	Activities	Sensor location	Validation	Results
(Lee and Kwan 2018)	RF, Gradient Boosting	Jogging, Walking & stairs, Sedentary, Standing	Smartphone, CV	10-fold	99.18%, 99.18%

2.4.4 Review of condition classification

In previous sections, we mentioned that current commercial devices can be less accurate for people with chronic conditions where their gait is affected. This is because commercial devices tend to be targeted towards large scale deployments which are based on an average user and furthermore, the accuracy does not need to be as high as it does for research devices. Additionally, the majority of the commercial devices focus on the healthy population, and there is not any evidence that they are actively developed with patient data. In this section, articles that performed activity classification by training their algorithms with one population, for example healthy, and tested the algorithm with a completely different population, for example patients with rheumatoid arthritis (RA), are discussed. The results from the majority of the articles suggested that the algorithms trained with data that was very similar to the test data yielded better performance. Lonini and colleagues developed five different classification models, which were grouped in two categories; global models and personal models (Lonini et al. 2017). The participants were healthy volunteers and a group of patients that used a wearable device (Actigraph) on their waist while wearing either an existing control or the novel knee assistive device. The aim of this study was to test the accuracy of the wearable device for people with disabilities while wearing an assistive device. Table 2.8 demonstrates the five classification models in more detail.

Table 2.8: Explanation of the five classification models, healthy, impairment specific, device specific, patient specific, patient & device specific, used by (Lonini et al. 2017).

Name of classification model	Training dataset	Test dataset
Global models		
Healthy	Healthy (no device)	Patient (novel device)
Impairment specific	Patient (control device)	Patient (novel device)
Device specific	Patient (novel device)	Patient (novel device)
Personal models		
Patient specific	Patient (control device)	Patient (novel device)
Patient & device specific	Patient (novel device)	Patient (novel device)

The *healthy* model trained on healthy participants achieved 53%, which was the lowest balanced accuracy among all the models. The second lowest balanced accuracy was achieved by the *impairment specific* model and it was 55%. Even though the training and test datasets were collected from the patient groups, the training dataset was from the patients who wore the control device. And the test dataset included data from patients who wore the novel device.

Among all the models, the *patient & device* specific model achieved the highest accuracy scores (76%). This was due to the fact that the training and test datasets included similar data from a single patient who wore the novel device. The other two models, *device specific* and *patient specific* achieved the third (61%) and second (66%) highest accuracy. The results suggested that a greater accuracy was achieved when the model was person-specific rather than group-specific.

Healthy participants and RA patients were recruited for a study conducted by (Andreu-Perez et al. 2017). One ML algorithm, Dichotomous Mapped Forest, with two different feature reduction techniques, 1) deep learning mapping and 2) metric learning mapping, were tested. Additionally, two scenarios were tested, the first was to train the algorithm with data from RA patients and test it with unseen data from RA patients. The second scenario was to train the algorithm with data from healthy participants and test it with unseen data from RA patients. Similar results to Lonini study were identified. The algorithm of the first scenario achieved higher accuracy scores than the algorithm of the second scenario. Other studies, such as Mannini and Ignatov, demonstrated the same behaviour as Lonini and Andreu-Perez studies, however they used different types of datasets from the ones already discussed (Mannini et al. 2016; Mannini et al. 2017; Ignatov 2018; Lonini et al. 2017; Andreu-Perez et al. 2017). Mannini and colleagues in 2016 compared healthy elderly, post-stroke and Huntington's disease patients (Mannini et al. 2016). Also, Mannini and colleagues in 2017 trained the algorithms with data from healthy adults and tested the algorithm with data from healthy youths (Mannini et al. 2017). Finally, the Ignatov study used two different public datasets derived from healthy participants, WISDM and UCI HAR (Ignatov 2018). Both datasets achieved high performance $> 90\%$ when the test set was a subsection of the associated original dataset. However, when the test set was from the other dataset, the performance dropped 83%. This suggested that several characteristics of the data collection, such as type of sensors, participants, lab-setting, might influence the results even if the data is collected from healthy participants. In order to reduce the likelihood that these factors influence the final outcomes, universal guidelines could be created and followed by the researchers to achieve better results. However, developing these would be beyond the remit of this thesis.

2.4.5 Review of step count literature

Most of the commercial- and consumer-grade devices for counting steps during walking use "black-box" algorithms that cannot be replicated independently. This section reviews findings

and gaps in the literature related to the step count concept. The findings discussed in this section are related to: a) participants, b) algorithms, c) sensor device and location, d) activities and e) results. Each section is discussed in detail below.

Most of the articles reviewed have targeted the healthy population, therefore there is much less literature and data related to people with chronic diseases or walking impairments. Some of the articles reviewed recruited: impaired elderly in-patients (Marschollek et al. 2008), stroke patients (Fulk et al. 2014; Klassen et al. 2016), patients with traumatic brain injury (Fulk et al. 2014), patients who underwent total joint arthroplasty (Lipperts et al. 2017), RA patients (Larkin et al. 2016), multiple sclerosis patients (Motl et al. 2011), patients with polymyalgia rheumatica (Chandrasekar et al. 2018), patients with lumbar fusion surgery (Gilmore et al. 2020) and participants with walking impairments (Treacy et al. 2017). Furthermore, Ummels and colleagues recruited patients with at least one of the following diseases: cancer, osteoarthritis, chronic pain, chronic obstructive pulmonary disease (Capela et al. 2015b), cardio-vascular diseases (Ummels et al. 2018). Additionally, patients with orthopaedic pathologies, such as lower limb osteoarthritis and cruciate ligament injury, as well as patients with neurological diseases, such as radiation induced leukoencephalopathy, Parkinson's disease (Lamont et al. 2018), hemispheric stroke and toxic peripheral neuropathy were recruited by (Oudre et al. 2018).

Many of the articles which examined patient groups used commercial- and consumer-grade monitors to test how well they perform in terms of counting the number of steps. However, Marschollek and colleagues tested four available algorithms used in the literature using data in free-living from healthy and impaired participants (Marschollek et al. 2008). The step detection algorithms used were pan-Tompkins, dual-axis, wolf and autocorrelation. The results suggested that the algorithms yielded greater error for the impaired participants in comparison to the healthy participants. Additionally, Capela study developed a proprietary algorithm, which was based on adaptive locking period and adaptive signal shape template (Capela et al. 2015b). The algorithm was tested in 15 healthy participants during a 6-minute walk test, and it achieved accuracy greater than 99.4% in all participants. Another template-based matching algorithm was developed by Oudre and colleagues to count the number of steps of group of patients (Oudre et al. 2018). This suggested that in general there is a need to test commercial devices in a wider range of patient groups, and also that there is a need to develop step count algorithms that target specific patient groups, since each disease affects differently the movement of patients.

The device and its location used for step count played an important role in terms of how well the device performed in a number of studies. The majority of the commercial- and consumer-grade devices were located mostly on the wrist, hip and/or waist (Fortune et al. 2014; Fulk et al. 2014; Storm et al. 2015; Klassen et al. 2016; Larkin et al. 2016; Chu et al. 2017; Genovese et al. 2017; Feng et al. 2017; Treacy et al. 2017; Chow et al. 2017; Alinia et al. 2017; Chandrasekar et al. 2018; Tophøj et al. 2018; Lamont et al. 2018; Wong et al. 2018; Toth et al. 2018; Bunn et al. 2018; Ummels et al. 2018; Bunn et al. 2019; Gilmore et al. 2020; Montes et al. 2020). From those articles, only Fortune and Genovese studies have compared commercial devices with their own step count algorithm (Fortune et al. 2014; Genovese et al. 2017). The remaining articles only compared the commercial- and consumer-grade devices in terms of how well they perform to measure the number of steps. Moreover, Huang et al. developed their own algorithm, but the algorithm was integrated in a smartphone device rather than in a wearable device (Huang et al. 2012). In most of the reviewed articles that developed their own step count algorithm, they used a smartphone device (Mikov et al. 2013; Chandel et al. 2014; Seo et al. 2015; Zeng et al. 2015; Capela et al. 2015b; Lee et al. 2015; Gu et al. 2017; Thanh et al. 2017; Dirican and Aksoy 2017; Ao et al. 2018; Kang et al. 2018; Rodríguez et al. 2018; Pham et al. 2018).

Based on these findings, it has been identified that it is desirable to develop step count algorithms that can be used for people with walking impairments and for slow walking speeds. Additionally, it would be beneficial if the algorithms were publicly available in order to help other researchers to strengthen the literature and to develop high performance algorithms. Lastly, the wrist has become one of the top options for wearables, hence it might be beneficial to appropriately value that specific location and develop accurate algorithms. Since, the wrist is one of the most user-friendly and unobtrusive wearables, as well as acceptable location (Cola et al. 2017).

The most common activity performed while calculating the number of steps was walking. Additionally, ascending stairs and descending stairs were also examined. It is essential to have step count algorithms that can accurately measure the number of steps undertaken during these activities. These activities are some of the most common locomotion activities undertaken throughout the day for people with chronic conditions and walking impairments. Moreover, each of these activities can be further categorised into different speeds. The most obvious example is walking with slow, normal or fast speeds. Even though not all of the articles have examined the different speeds for each activity, it is very important to do so (Motl et al. 2011;

Mikov et al. 2013; Klassen et al. 2016; Ho et al. 2016; Cho et al. 2016; Beevi et al. 2016; Chow et al. 2017; Feng et al. 2017; Alinia et al. 2017; Genovese et al. 2017; Wong et al. 2018; Tophøj et al. 2018; Lamont et al. 2018; Pham et al. 2018; Toth et al. 2018; Rhudy and Mahoney 2018). The reason for this is because the algorithms of the commercial-grade devices are developed based on healthy population data. Therefore, to count a step within the device, the acceleration value should pass a threshold, which is based on healthy data (Walker et al. 2016). Often, the elderly, people with walking impairments and patients with chronic conditions walk slower in comparison to the healthy population (Mancuso et al. 2014) with the resulting accelerations far smaller than those seen in the healthy population. This can be confirmed from the results of the articles that examined several walking speeds. It is demonstrated that walking at slow speed yielded the greatest error when counting the number of steps (Motl et al. 2011; Beevi et al. 2016; Feng et al. 2017; Pham et al. 2018).

An observation made regarding step count algorithms related to the performance evaluation of those algorithms. For example, when ML algorithms are used, they are usually validated using universal performance metrics as the ones already described in section 2.4.2.5 and which are: accuracy, recall, precision, F1-score, etc. It is essential to have such performance metrics for validating the step count algorithms. In the majority of studies reviewed, the algorithms were validated with several metrics, such as: intraclass correlation (ICC[2,1]), mean percentage error, accuracy, mean absolute percentage error and root mean square error. Even though there was not a common metric to measure the accuracy of the step count algorithms, the general outcome of the results was similar in the majority of the articles. The general outcome was that in the studies that compared the same device or algorithm for healthy and patient groups, the accuracy scores of the patient groups were lower in comparison to the healthy group (Marschollek et al. 2008; Motl et al. 2011; Larkin et al. 2016; Lipperts et al. 2017; Ummels et al. 2018; Gilmore et al. 2020). For example, the Lipperts study compared the error created from a device attached on the thigh of healthy participants and orthopaedic patients while walking and climbing up and down the stairs (Lipperts et al. 2017). The results suggested that the mean percentage error was always higher for the patient group in comparison to the healthy group. For instance, the errors for the healthy group were 1.7%, 6.4% and 5.4% for walking, ascending and descending stairs respectively. For the patient group, the errors were 3.4%, 6.9% and 8.2% for walking, ascending and descending stairs respectively. Additionally, it has also been observed, in studies which examined several commercial- and consumer-grade devices, that the location, walking

speed and devices plays an important role in the accuracy of the step count results (Chow et al. 2017; Bunn et al. 2019).

2.4.6 Generation of synthetic signal

In sections 2.4.3 and 2.4.5, we have seen that there is a limited number of available datasets online for both healthy and patient populations. Additionally, there are several challenges to collect data, especially during the pandemic. For instance, the researchers encounter multiple funding, bureaucratic and regulatory challenges. To address these issues, a possible option is to generate synthetic data. This concept is used to artificially create data rather than actually collecting the data. This is done using different types of algorithms to create test data for model validation, as well as for new tools and products. Additionally, synthetic data can also be used in Artificial Intelligence (AI) model training. This method provides several benefits, for example: a) the data is easily accessible; b) the method is cost-effective; c) the data is available quickly; d) the privacy of patients is protected; e) the data can be used as a benchmark while comparing different methods; and f) the data can be used to supplement real data (Wang et al. 2019).

Since there is limited available data that represents the activities of interest as they have been collected from an accelerometer, three different methods are discussed in this section that can be potentially used to generate walking synthetic data. The methods are: 1) coupled differential equations, 2) Generative Adversarial Networks (GANs) and 3) the pendulum approach.

The first approach was developed by McSharry and colleagues to generate synthetic ECG signals (McSharry et al. 2003). This method has since been used by other authors to generate different types of synthetic signals, such as phonocardiogram (Almasi et al. 2011), force during running (Racic and Morin 2014), jumping (Racic and Pavic 2010a) and walking (Racic and Brownjohn 2012). This method has not previously been used yet to generate acceleration signals. The basic idea of this approach is to use three dynamic coupled equations to represent the morphology of the desired signal. The signals represented by this approach are all periodic or near-periodic. Consequently, by using the coupled equations, the morphology of the signal for one period is described and this period is repeated the chosen amount of times.

The second approach, GANs, is a deep-learning approach that is mainly used to generate synthetic data. It uses two neural networks, generative model and discriminative model, to

generate new, synthetic signals that can be classified as real data (Goodfellow et al. 2014). The two models operate against each other, hence it is described as an adversarial process. Therefore, the generative model, as the name suggests, generates new data by capturing the distribution of data. The discriminator model then evaluates the probability that the new data came from the generative model or the training dataset which includes real data. The generator is updated until the discriminator cannot tell the difference between real and generated data. Alzantot and Hassouni studies used the GANs approach to generate sensory data and both used publicly available datasets as their training datasets (Alzantot et al. 2017; Hassouni et al. 2018). The former study used Human Activity Recognition using a smartphone data set, and the latter used the WISDM dataset. Both datasets contained 3D accelerometer data from six activities, walking, walking upstairs, walking downstairs, sitting, standing and lying, all performed by healthy participants. The Human Activity Recognition dataset also contained 3D gyroscope data.

The third and final approach is based on the pendulum system, which can be used mathematically to model the human arm configuration. The mathematical approach is based on the Lagrange equation which describes a triple pendulum system (Al-zu et al. 2012; Agarana and Akinlabi 2018). This method separated the human arm into three segments, upper arm, lower arm and palm. To successfully use this method, the length, mass and position of the three arm segments should be known. This mathematical model is used to calculate the angular displacement, period and frequency of the pendulum.

Table 2.9 describes the advantages and disadvantages of the three aforementioned approaches.

Table 2.9: Advantages and disadvantages of three approaches to generate synthetic data.

Approach	Advantage	Disadvantage
Coupled differential equations	Can be used in multiple scenarios	Has not been developed yet for acceleration data
GANs	High accuracy	Requires large dataset and real dataset
Pendulum system	Simple method	Requires information about human body segments, Final result is not acceleration

A recent study, by Alharbi and colleagues, developed a model that generated several types of human activity sensor data using a Wasserstein GAN method. For the study two real

datasets were used for the generation of the synthetic dataset, Sussex-Hawei Locomotion and Smoking Activities Dataset. In order to test whether the synthetic datasets were similar to the originals, two deep learning classifiers were used, convolutional neural network and long short-term memory. Two models were developed for each algorithm, one model was trained with the original dataset and tested with the synthetic dataset, and vice versa. An F1-score was used to evaluate the performance of the classification, and scores between 0.59 and 0.99 were achieved (Alharbi et al. 2020). The results of this study suggest that it is possible to develop representative synthetic datasets.

Synthetic data might be a useful alternative for real data, however its use still faces some challenges. A few of the major challenges identified are: a) the requirement for collection of realistic datasets in order to generate synthetic data, b) the degree to which realistic data represents a ground truth, particularly in relation to including the unreliability and uncertainty of sensors, c) the acceptance of the synthetic data from the potential users, d) the time and effort to build a representative model.

2.5 Summary and aims

Physical activity (PA) has been identified as essential for our health and well-being. Wearables are tools that can be attached on the human body in various locations in order to measure PA objectively. The wrist is one of the most common and popular locations used to attach the wearable device. These devices contain several sensors for different usages. In order to measure PA objectively, the accelerometer sensor is the most essential. By collecting acceleration data, an effective approach to measure PA is using ML algorithms to identify whether the user is active or sedentary, or to even identify the specific activities that the user performed. Step count is an objective measurement that can be used to inform clinicians about the overall activity of their patients. This literature review has shown the differing performance of available algorithms in people with impaired gait, compared to healthy populations. Regardless of the metric used to quantify PA, it is essential that algorithms are developed and trained using data from the target population. For example, if the intended users are patients with RA, the algorithm should be trained with data collected from patients with RA rather than a healthy population.

Although studies regarding activity classification have been conducted by many authors, there are still some problems that need to be explored. For example, limited studies were conducted

to classify ADLs from people with walking impairments and slow pace. The majority of studies explored the healthy population. Additionally, only a few studies have analysed the classification between different sub-population groups. Hence, to fill these literature gaps, this thesis explores whether the classification of daily activities between two different groups can be achieved with high performance. Then the next step is to explore whether each group might need a population-specific ML algorithm to achieve high performance.

It was also observed that the majority of the step count algorithms were developed to work in healthy populations. Therefore, these algorithms might not be very representative for any other sub-population group, such as elderly, people with walking impairments. In order to understand and address this problem, an adaptive algorithm to different sub-population groups was developed.

If this is achieved, then the wearable device can be used to firstly understand whether the person who wears it is healthy or patient. The next step, according to the outcome of the first classification, is to classify the activities that the user performed using the population-specific ML algorithm. The last step is to count the number of steps for the appropriate walking speed.

To achieve this, it is essential to have available a large dataset that can be used to train the ML algorithms, and also to be used as a good representation of the desired sub-population group while developing a novel step count algorithm. The data collection is time-consuming, and it is important to ensure that the data is secured. Due to this, there are not many datasets available online that use acceleration for daily activities, and the majority of these datasets represent the people from the healthy population. There are a few methods in the existing research that can be used to generate realistic synthetic data, however this problem is still insufficiently explored especially for accelerometer signals that represent activities such as walking. Therefore, to fill this gap, an algorithm was developed to examine whether it is possible to generate synthetic walking acceleration signals. This will benefit the research community and the people who need data in order to develop and/or analyse algorithms.

The following chapters will present the approaches taken in order to fill the literature gaps and to answer the proposed research questions.

2.5.1 Research questions

1. Is a wrist mounted activity monitor suitable for condition classification and activity recognition? (Chapter 3)
2. Can we automatically identify whether a patient is moving normally? (Chapter 3)
3. Can we automatically identify different types of physical activity in healthy participants in normal and simulated-pathological conditions? (Chapter 3)
4. Can we accurately measure step count in healthy participants in normal and simulated-pathological gaits? (Chapter 4 & 5)
5. Can we accurately generate synthetic acceleration data that represent normal and relevant atypical walking patterns? (Chapter 6)

Chapter 3

Using machine learning for activity and condition classification

3.1 Introduction

Chapter 3 describes a pilot study that was carried out with healthy participants. The purposes of the study were: a) to create a baseline for the ML algorithms based on healthy participants and b) to test step count algorithms from the literature on healthy participants. This study was an essential first step in confirming whether the differences in gait between healthy and functionally compromised persons were sufficiently pronounced to develop tuneable algorithms.

3.2 Methodology

A prospective pilot study was conducted with healthy participants recruited at Chapel Allerton Hospital. Participants performed a range of ADLs under normal and simulated-pathological conditions. We used simulated-pathological condition, since recruiting actual patients was considered infeasible and impractical, especially given the exploratory nature of the pilot work. Participants wore two accelerometer-based devices to measure their acceleration while performing the activities.

The study was reviewed and approved by the School of Medicine Ethics Committee (Ref #: MREC16-172). Each participant provided written informed consent before participating in the study.

3.2.1 Study design

The research questions, mentioned in section 2.5.1, were answered using the acceleration data collected from healthy participants.

For each participant the same approach was followed to collect the accelerometer data. The data collection was performed at a single visit for each participant.

3.2.2 Sampling plan

3.2.2.1 Sample size

For this pilot study, a sample of 30 healthy participants, of 18+ years of age, was recruited based on published guidance for pilot studies which indicate that 30 participants are appropriate for a pilot study (Julious 2005). In comparison to other databases, which were mentioned in Table 2.5, the database collected for this thesis had recruited larger number of volunteers. Additionally, the volunteers specifically emulated the movement of people with a compromised gait.

3.2.2.2 Inclusion/exclusion criteria

Participants were considered eligible for inclusion if they could walk freely without pain for two minutes. All participants were healthy, without any musculoskeletal condition or any other condition which would have affected their gait.

3.2.2.3 Recruitment process

Participants were recruited via email and word of mouth from the staff and students of the Leeds Institute of Rheumatic and Musculoskeletal Medicine, the Leeds Institute of Health Sciences and the Institute of Medical and Biological Engineering.

3.2.2.4 Data collection

Data was collected at the Gait Lab of the NIHR Leeds Biomedical Research Centre based at Chapel Allerton Hospital, Leeds, UK. Each potential participant was asked to come for a single visit at the hospital where they were screened prior to providing written informed consent. Before collecting the accelerometer data, participant demographics were collected via interview. The details collected were: age, sex, height, weight, dominant hand and dominant leg. Clear

instructions for the data collection process were given verbally and in writing (see Appendix C).

3.2.2.5 Data acquisition

Activities were captured using two MOX accelerometers (Maastricht Instruments, Maastricht, The Netherlands). The accelerometers were initialised using the IDEEQ software (Maastricht Instruments, Maastricht, The Netherlands) before attaching them on the participants. The acceleration data is sampled by a 12 bit analog to digital converter at the rate of 100 Hz. It is then stored in a raw, non-filtered format in the units of gravity (G's). This data is stored directly into an embedded SD memory card in the MOX sensor. The data can be transferred via a USB connection to a personal computer with the dedicated MOX software. Devices were placed on the: non-dominant wrist of the participant because in real life there is a greater chance to perform extra activities with your dominant hand, and on the dominant ankle of the participant. Both monitors were attached using elasticated straps. Figures 3.1 and 3.2 demonstrate the position of the device and the device itself respectively.

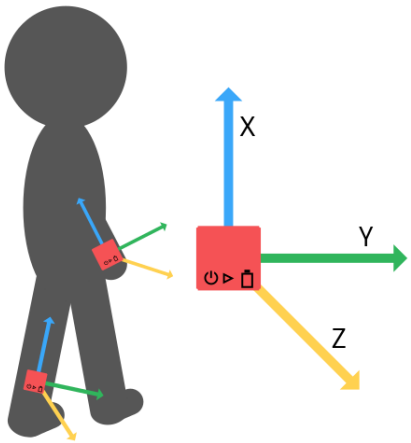


Figure 3.1: Two activity monitors placed on the wrist and ankle of the healthy participant

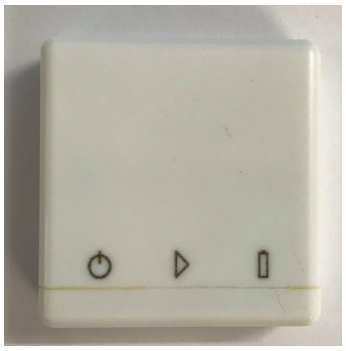


Figure 3.2: MOX accelerometer

It is important to mention that commercial devices have been used in research for activity classification of healthy individuals. However, it is already known from previous studies that step count results are much less accurate in people with reduced mobility. Additionally, prior to any analysis, it was decided by the research team that the product used would need to make the raw acclerometry data available. The raw signal was essential for the development of machine learning algorithms, as well as the step count algorithms. By using the raw signal, the research

team would have full signal information, where any processing technique can be implemented. Also, the MOX device had the potential to integrate the algorithms developed, which means that the algorithms can be used directly when the device is worn. Since the commercial devices did not allow raw data output or algorithm integration, they were excluded in the early scoping.

Participants were instructed verbally by command on how to perform the activities prior to attaching the accelerometers, and this was reinforced while they performed the activities. The nine pre-determined activities were: (1) lie down, (2) sit, (3) stand, (4) stand-to-sit, (5) slow walk, (6) normal walk, (7) fast walk, (8) stair ascent and (9) stair descent.

A camera was used to video record the participants while performing the activities which then provided a gold standard during subsequent analysis. This reference standard enabled retrospective labelling of the accelerometer data, to define the start and end of each activity, as well as the number of steps.

Participants were first instructed to perform a single jump prior to starting the activity sequence. The jump provided a recognisable peak in the signals that could be used to synchronise the accelerometer signal with the gold standard video recording. After performing the jump, participants performed the nine activities sequentially. After the end of the activity sequence, the participants were asked to jump again to provide a clear signal at the end of the activity sequence.

Each set of activities was performed twice, under normal and simulated-pathological conditions. To simulate pathological gait, participants were asked to repeat the series of activities via a slow, short-step, shuffling gait mimicking that of someone impaired by a condition such as severe RA. A shuffling gait is defined as when the foot is moving forward at the time of initial contact or during mid-swing, with the foot either flat or at heel strike, usually accompanied by shortened steps, reduced arm swing and forward flexed posture (Whittle 2007).

A description and a video of the shuffling gait was provided to the participants prior to data collection (see Appendix C). Participants were free to try the simulated pathological gait before the monitors started collecting data. At the end of each set of activities the data was analysed and then saved using IDEEQ software to provide a proprietary format ".bin" file. The file was observed, to ensure that all three axes were collected, when the participant completed each set of activities. Additionally, the researcher was observing the movement of each participant while

they were performing the activities to assess visually the quality of the simulated-pathological signals. The limitations of the collecting simulated-pathological signal were discussed in detail in sections 3.4 and 7. Figure 3.3 demonstrates the three main characteristics of the pathological gait.

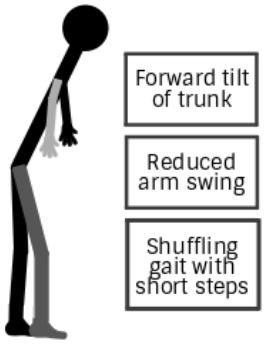


Figure 3.3: Characteristics of a patient with shuffling gait

3.2.3 Data analysis

The data analysis followed in this chapter is based on the Activity-Recognition-Chain process (ARC). Generally, it is used as a generic framework to design and evaluate activity recognition systems. As discussed in chapter 2, ARC is a series of techniques, such as signal processing, pattern recognition, and machine learning. It consists of six steps: (1) data acquisition, (2) data pre-processing, (3) data segmentation, (4) feature extraction and selection, (5) training and classification and (6) performance evaluation.

This process was used to analyse the accelerometer data, and develop machine learning models to answer the first three research questions (see secction 2.5.1), which are related to:

- Location of the monitor (wrist and ankle)
  - Condition classification using the dataset that includes all the activities
  - Activity classification using only the general classes (type) of activities
- Condition classification
  - the dataset that includes all the activities
  - the sub-dataset that includes the dynamic activities
  - the sub-datasets that includes the dynamic activities individually

- Activity classification
  - Healthy training set vs healthy test set
    - Activity type
    - Activity task
  - Simulated-pathological training set vs simulated-pathological test set
    - Activity type
    - Activity task
  - Healthy training set vs simulated-pathological test set
    - Activity type
    - Activity task

Owing to the broad applicability of ARC process, some of the ARC steps have been also employed to assist the process for answering the last two research questions. This was done in order to prepare the input data that would be used for answering the questions. Figure 3.4 shows analytically the steps of the ARC process followed and used for answering all the research questions.

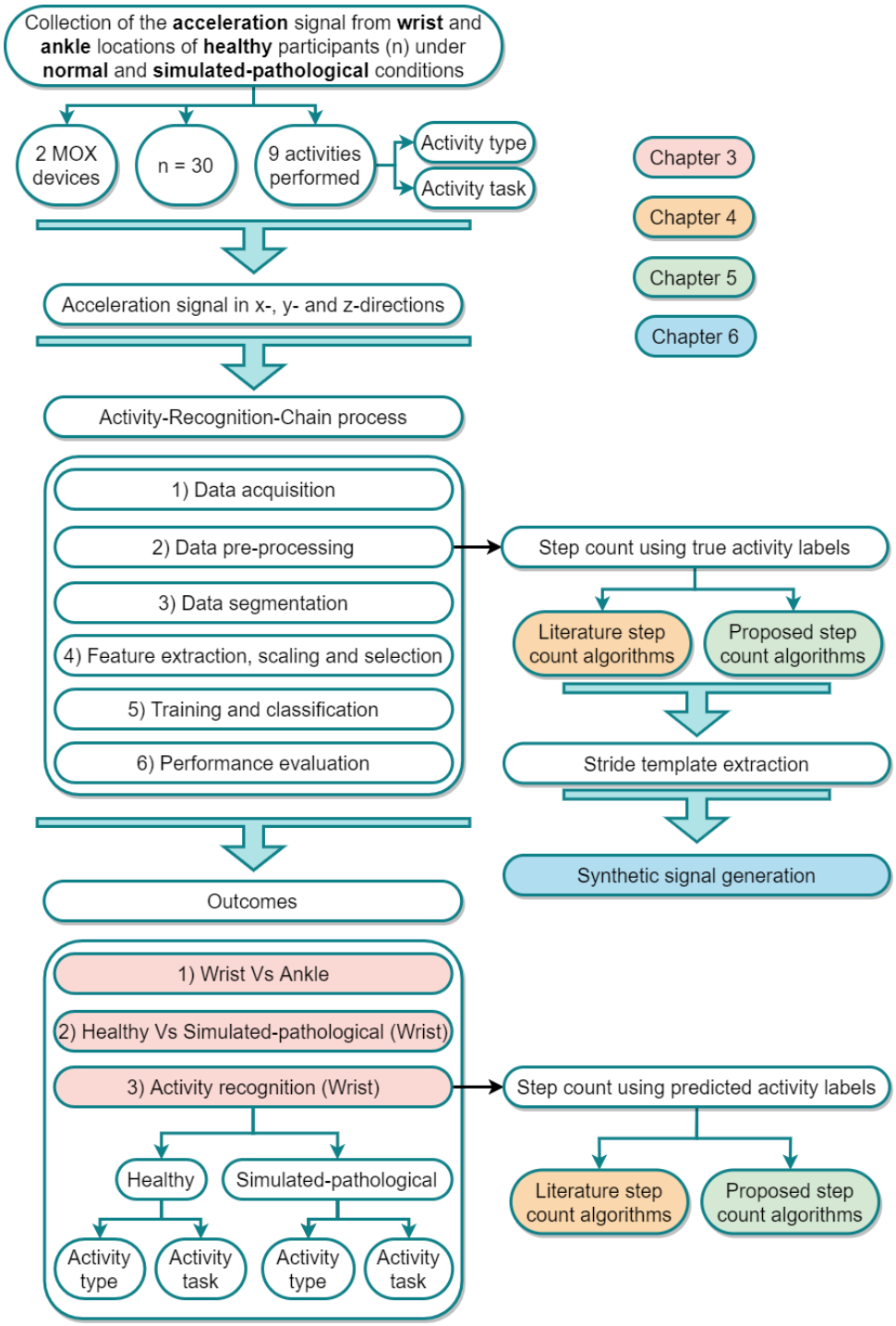


Figure 3.4: Flowchart describing the flow of each chapter in the PhD

3.2.3.1 Data pre-processing

Data extraction The binary files from the accelerometer were imported to MATLAB (MATLAB R2017a) and saved as text files. The reason for saving the files through MATLAB was that the proprietary script developed by Maastricht Instruments was written in MATLAB. The text files were then imported into Python^{TM} (v3.6) for analysis. The extracted text files con-

tained three columns of acceleration data, representing acceleration along the three principal axes. We derived five types of discrete signals from the tri-axial accelerometer data: (i) dynamic accelerations, (ii) total magnitude (M), (iii) jerk, (iv) angular velocity and (v) inclination angles.

Dynamic acceleration was calculated by averaging the readings, and then subtracting the corresponding average value from the raw acceleration signal. This was done for all three directions.

$$dynamic\ accel\ (x,t) = accel(x,t) - sum\ for\ all\ in\ t\ accel(x,t)/n \tag{3.1}$$

Total magnitude was calculated using the magnitude formula used for 3D vectors. It combines all three directions, therefore it is not affected by the orientation of the device.

$$total\ magnitude\ (M,t) = \sqrt{acc_x^2 + acc_y^2 + acc_z^2} \tag{3.2}$$

Jerk is the rate of change of acceleration. It was calculated using the collected acceleration signal and its sampling time on each direction. Jerk is created when an object accelerates so rapidly, and hence the acceleration is also increasing.

$$jerk(x,t) = \frac{acc_{x(t+T)} - acc_{x(t)}}{T} \tag{3.3}$$

where T is the sampling period.

Angular velocity was identified by calculating the angle between the acceleration vectors in the current and the previous point. The accelerometer registers its data at equal time intervals, therefore the angle between the vectors provides the angular velocity. This signal refers to how fast an object rotates relative to another point.

$$\cos(i, i+1) = \frac{x_i x_{i+1} + y_i y_{i+1} + z_i z_{i+1}}{\sqrt{x_i^2 + y_i^2 + z_i^2 + x_{i+1}^2 + y_{i+1}^2 + z_{i+1}^2}} \tag{3.4}$$

Inclination angle was calculated for each direction. It is used to measure the tilt of the accelerometer in all three directions. Figure 3.5 demonstrates the inclination angle of the three

directions.

$$\phi_x = \arccos \frac{acc_x^2}{acc_t} \tag{3.5}$$

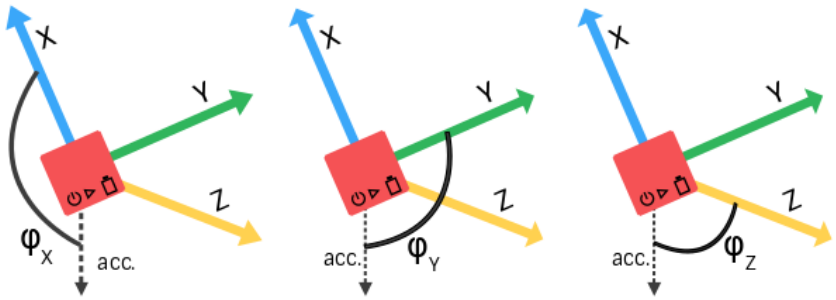


Figure 3.5: Inclination angles of x, y and z directions of the accelerometer

Data labelling The raw acceleration signal was manually labelled, with each condition and activity identified, and labelled based on the graphs and on the time of the gold standard video. The reason for doing this manually was because no automated algorithm had yet been built to automatically label the signal with the correct activities. Figure 3.6 demonstrates a labelled acceleration signal.

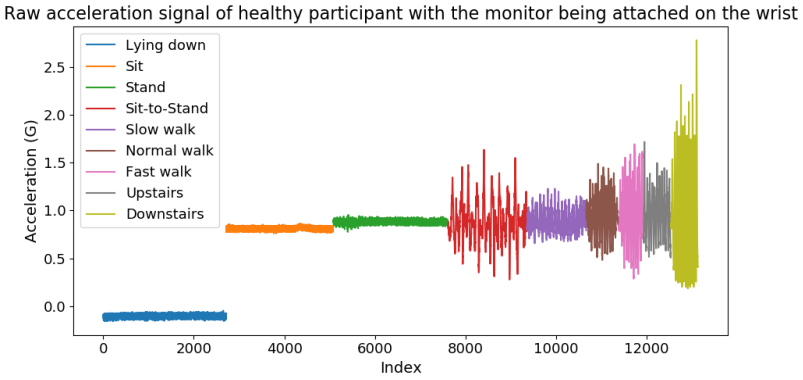


Figure 3.6: Acceleration signal showing nine activities of daily living

The signal was classified by activity-types, activity-tasks and conditions. The first group represented general types of activities, such as static, transition, and dynamic. The second group represented the precise activities or tasks, such as lying, sitting, standing, stand-to-sit, walking slowly, walking normally, walking fast, stair ascent and stair descent as demonstrated in Figure 3.7.

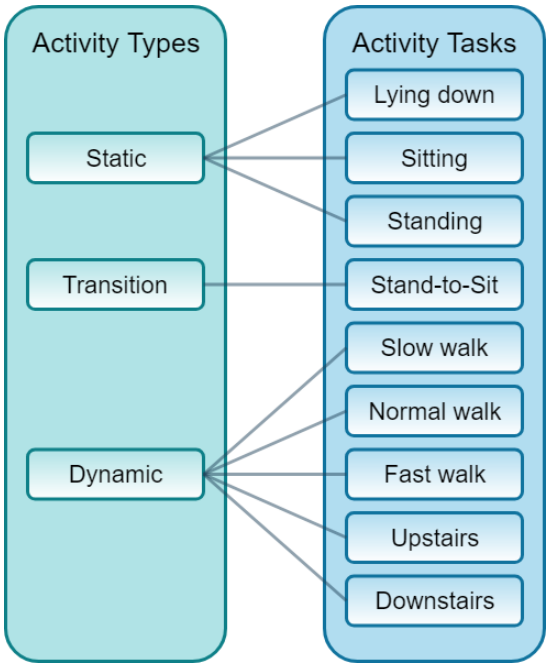


Figure 3.7: Relation between activity types and activity tasks performed

The last group represented the condition under which the activity had been performed, i.e. either the normal condition or simulated-pathological condition. Table 3.1 demonstrates the labels used to represent each category.

Table 3.1: Manual labels given for each classification.

Condition classification		Activity type classification		Activity task classification	
Normal	0	Static	1	Lying	1
Simulated-pathological	1	Transition	2	Sitting	2
		Dynamic	3	Standing	3
				Stand-to-sit	4
				Slow walk	5
				Normal walk	6
				Fast walk	7
				Stair ascent	8
				Stair descent	9

Data filtering A low-pass Butterworth filter was used to filter the acceleration data of this pilot study. This filter has a flat amplitude response within the pass-band, which means that it is ripple-free as demonstrated in Figure 3.8. Equation 3.6 is the definition of a Butterworth

filter in terms of the magnitude of amplitude response:

$$|H(j\omega)| = \frac{1}{\sqrt{1 + (\frac{\omega}{\omega_c})^{2n}}} \tag{3.6}$$

where ω_c is the filter cut-off frequency and n is the filter order.

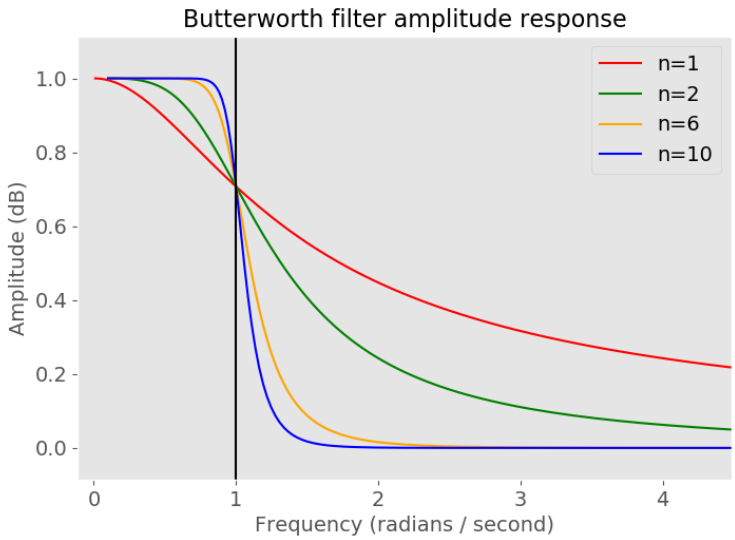


Figure 3.8: Response of Butterworth filter of four different orders

The cut-off frequency and order parameters were set to optimise the filter. We applied a sixth-order Butterworth filter with a cut-off frequency of 3 Hz for all the groups. The chosen cut-off frequency was fixed for all the activities and both groups to have a consistency in the filtering method. In general, the human movement has a frequency between 0-20 Hz (Wang et al. 2011). Walking has the greatest frequency among the activities performed in this study, hence its frequency was chosen to be the threshold for the cut-off frequency. This was done in order to ensure that no important signal characteristics were lost after the filtering process. Therefore, walking has a frequency between 0.6-2 Hz, and by applying the Nyquist theorem a minimum cut-off frequency of 4 Hz it is suggested. Hence, a credible cut-off at 6Hz was selected. Therefore by applying firstly 6 Hz as cut-off frequency, the data was checked visually to make sure no important information, such as activity peaks, were missing from the filtered signal, as well as to ensure that the noise in data was reduced. After the visual inspection, the cut-off frequency was chosen as 3 Hz to reduce the extra noise that looked like small peaks attached on the bigger peaks.

Figure 3.9 represents the acceleration signal at its raw and filtered formats.

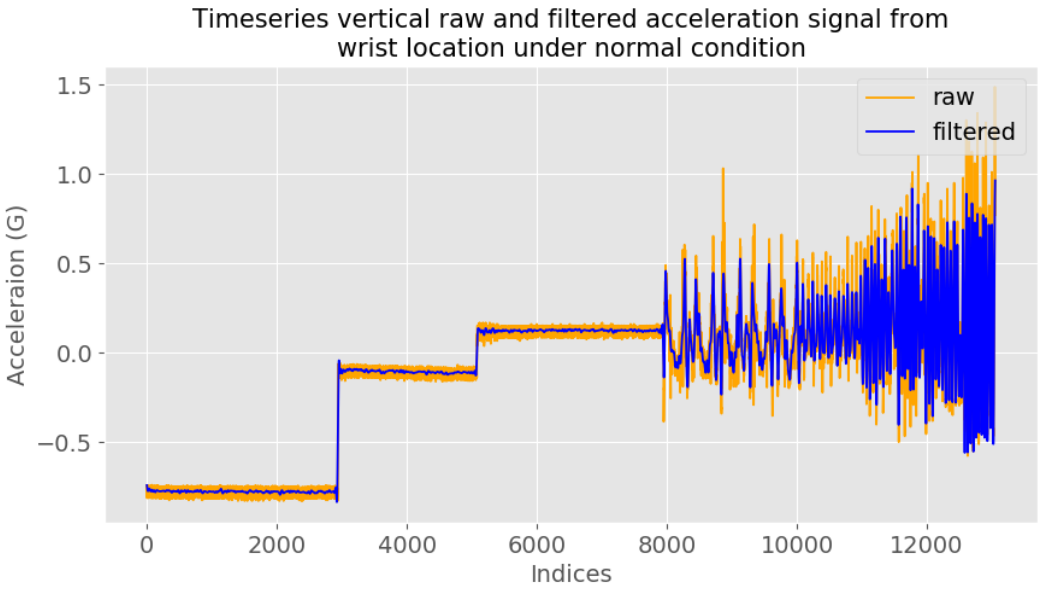


Figure 3.9: Raw and filtered vertical acceleration signal representing all nine activities

3.2.3.2 Data segmentation

The acceleration data was split into a series of short time windows, in which the signal may be approximated as stationary. A signal is characterised stationary when the its frequency does not change with respect to time. We used windows of 200 samples (with 50 samples overlap), corresponding to a time period of 2 seconds (Banos et al. 2014). Figure 3.10 demonstrates the schematic of this technique.

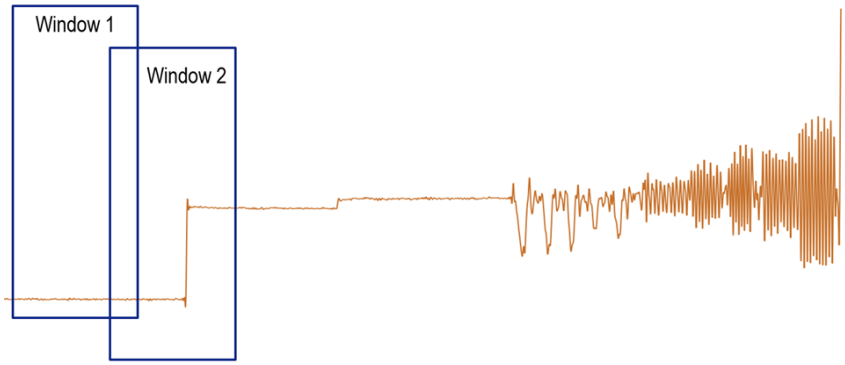


Figure 3.10: Schematic of overlap windowing technique

3.2.3.3 Feature extraction, scaling and selection

From the acceleration time series in each window, we extracted a set of 120 summary features to represent the acceleration (x, y, z, M), jerk (x, y, z, M), angular velocity (v) and inclination

angle (x, y, z) signals. M is the total magnitude of the 3D acceleration. Table 3.2 demonstrates all 120 features.

Table 3.2: Features used for activity and condition classification.

Time-domain		Frequency-domain
Mean acc (x-y-z-M)	Std acc (x-y-z-M)	Energy acc (x-y-z-M)
Median acc (x-y-z-M)	Skewness acc (x-y-z-M)	Max frequency 1 acc (x-y-z-M)
Kurtosis acc (x-y-z-M)	IQR acc (x-y-z-M)	Max frequency 2 acc (x-y-z-M)
RMS acc (x-y-z-M)	Median A acc (x-y-z-M)	Mean frequency acc (x-y-z-M)
MPSD acc (x-y-z-M)		Entropy acc (x-y-z-M)
Mean jerk (x-y-z-M)	Std jerk (x-y-z-M)	Energy jerk (x-y-z-M)
Median jerk (x-y-z-M)	Skewness jerk (x-y-z-M)	Max frequency 1 jerk (x-y-z-M)
Kurtosis jerk (x-y-z-M)	IQR jerk (x-y-z-M)	Max frequency 2 jerk (x-y-z-M)
RMS jerk (x-y-z-M)	Median A jerk (x-y-z-M)	Mean frequency jerk (x-y-z-M)
MPSD jerk (x-y-z-M)		Entropy jerk (x-y-z-M)
Mean ang.velocity	Std ang.velocity	
Mean inc.angle (x-y-z)	Std inc.angle (x-y-z)	

Extracted features might vary in terms of units, range and magnitudes, thus those with high magnitudes will stand out more than features with low magnitudes. Figure 3.11 demonstrates a matrix of the features created.

Feature Matrix										
Participants Windows	**F_1**	**F_2**	**F_3**	...	...	...	...	...	...	**F_n**
P1_W1	-0.772	-0.219	0.637	...	...	...	...	...	...	1.913
P1_W2	-0775	-0.221	0.637	...	...	...	...	...	...	1.681
P1_W3	-0.775	-0.221	0.637	...	...	...	...	...	...	1.710
P1_W4	-0.774	-0.220	0.637	...	...	...	...	...	...	1.783
P1_W5	-0.775	-0.220	0.636	...	...	...	...	...	...	2.003
P1_W6	-0.776	-0.220	0.636	...	...	...	...	...	...	1.961
P1_W7	-0.777	-0.221	0.635	...	...	...	...	...	...	2.030
P1_W8	-0.776	-0.222	0.634	...	...	...	...	...	...	1.681
P1_W9	-0.776	-0.222	0.634	...	...	...	...	...	...	1.761
Pm_Wn	0.422	0.052	-0.383	...	...	...	...	...	...	2.035

Figure 3.11: Feature matrix where each row represents one window for each participant, and each column represents a feature calculated.

To limit this, features were scaled using Min-Max, and then fed into principal component analysis (PCA), a dimensionality reduction technique (see Figure 3.12) which acts by performing

a linear data mapping to a space with lower dimensions where the variance of the data at low dimensional space is maximised. The data fed into PCA was a matrix with 120 columns representing the features and the rows represented the signal segmented into smaller windows for each participant. This was done prior to the cross-validation and therefore the PCA technique was applied to the whole dataset. In order to perform PCA, the PCA function from sickit-learn was used. Additionally, Min-Max was applied to the same data format as that described for PCA.

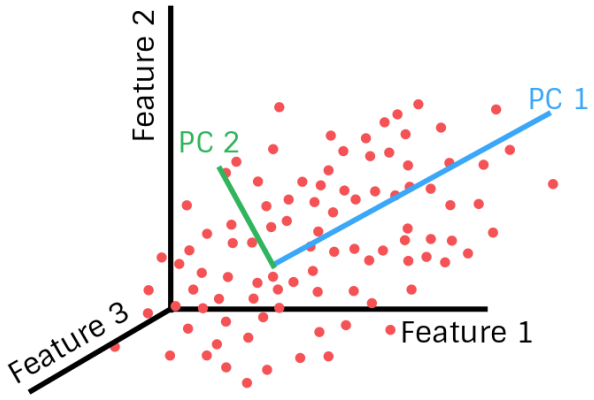


Figure 3.12: Dimensionality reduction and principal component analysis

Figure 3.13 demonstrates a matrix of the principal components created.

Principal Component Analysis Matrix										
Participants Windows	**PC_1**	**PC_2**	**PC_3**	...	...	...	...	...	...	**PC_n**
P1_W1	1.218	0.468	-0.494	...	...	...	...	...	...	0.232
P1_W2	1.239	0.440	-0.515	...	...	...	...	...	...	0.132
P1_W3	1.243	0.437	-0.516	...	...	...	...	...	...	0.120
P1_W4	1.241	0.440	-0.515	...	...	...	...	...	...	0.133
P1_W5	1.226	0.454	-0.518	...	...	...	...	...	...	0.109
P1_W6	1.230	0.453	-0.515	...	...	...	...	...	...	0.108
P1_W7	1.220	0.463	-0.516	...	...	...	...	...	...	0.111
P1_W8	1.221	0.461	-0.517	...	...	...	...	...	...	0.147
P1_W9	1.219	0.464	-0.518	...	...	...	...	...	...	0.121
Pm_Wn	-0.383	0.423	-0.104	...	...	...	...	...	...	-0.085

Figure 3.13: Principal component analysis matrix where each row represents one window for each participant, and each column represents a principal component calculated.

A reduced number of linear combinations of these features were selected using PCA. A 95% threshold was set on the explained variance. The principal components represent the same amount of information as the original features, meaning that it is possible to restore the original features from the transformed principal components. In addition, the total variance remains the same, although it is distributed differently than in the original case. The first principal component explains the most variance among the new principal components, and also the most variance a singular principal component can explain. In general, the first principal component explains the most variance out of all principal components, and the last principal component explains the least variance out of all principal components. By reducing the dimensionality of the feature set, we limited the risk of over-fitting subsequent classification models. The PCA feature set was then used as input to a selection of ML classifiers.

3.2.3.4 Training and classification

Four scenarios were conducted to answer the first three research questions. The scenarios were:

1. algorithm trained on combined normal and simulated-pathological data; tested on combined normal and simulated-pathological data (condition classification)
2. algorithm trained on normal data; tested on normal data (activity classification)
3. algorithm trained on simulated-pathological data; tested on simulated-pathological data (activity classification)
4. algorithm trained on normal data; tested on simulated-pathological data (activity classification).

For this study, these specific classification models were chosen to detect human activity: Multilayer Perceptron Neural Network (NN), Random Forest (RF), k-Nearest Neighbour (kNN), Gaussian Naïve Bayes (GB) and Support Vector Machine (SVM) (Preece et al. 2009; Saez et al. 2016).

Artificial neural networks There are several types of NNs. In this thesis, we consider only the simplest type, the multilayer perceptron (MLP). There are two perceptron types, single layer and multilayer. A single perceptron can be used only for linearly separable data. MLP, also known as a feedforward NN, can deal with non-linearly separable data. This means that a set of linear classifiers could model a non-linear model, and this is done by using additional

layers that allow combination of the linear classifiers from the first layers to create a non-linear decision boundary.

A single perceptron is a NN unit that consists of (i) weights, w, and bias, b, (ii) a combination function and (iii) an activation function. Specifically, the perceptron accepts several input variables, such as x_1, x_2 and so on, as demonstrated in Figure 3.14. These n inputs, x, are then multiplied with the weights and the net sum of all these multiplications is calculated (equation 3.7).

$$y = (\sum_{i=1}^{n} w_i x_i + b) \tag{3.7}$$

Then, an activation function checks whether the net sum exceeds a certain threshold. Based on this decision, an appropriate output is resulted.

$$y = \begin{cases} 1 & \text{if } \phi(w \cdot x + b) > 0 \\ 0 & \text{if } \phi(w \cdot x + b) \leq 0 \end{cases} \tag{3.8}$$

where w is a vector of weights, b is the bias, x is a vector of input variables and ϕ is the activation function.

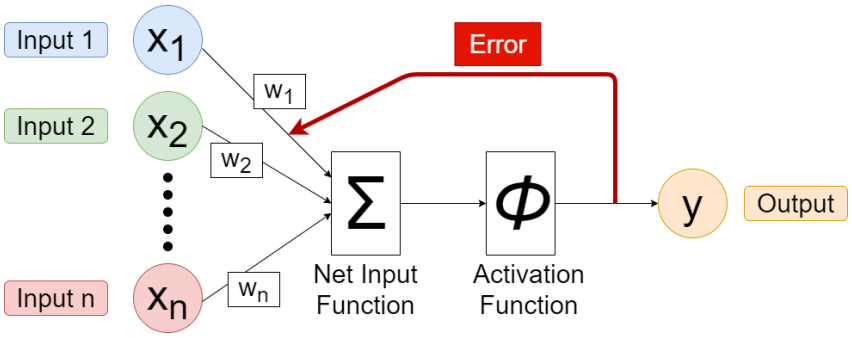


Figure 3.14: Diagram of a perceptron

Figure 3.15 demonstrates an artificial neural network. MLP has similar structure with a single layer perceptron with one or more hidden layers. Hidden layers are layers between the input and the output layers that allow the neural network to classify non-linearly separable data.

An MLP algorithm performs three steps when training the model. At the beginning, when the model is initiated, the weights and bias are set at random. The optimum weights and bias are learned from the data set. Subsequently, the steps followed are: 1) forward pass, 2) calculate

error and 3) backward pass. In the first step, the output of the model is calculated, which is the predicted value. In the second step, the error between the predicted and the true values is calculated using the loss function. Lastly, the error value is back-propagated using gradient and the weights of the model and the bias are updated to different values that aim to minimise the error. In the first phase, the non-linear activation function (ϕ) is used to describe the relationship between inputs and outputs in a non-linear way. The weights and bias keep changing in the training process until the output is accurately classified. The error in the output is back-propagated and weights updated to minimize the error. For this thesis, the ReLU activation function was used because the neurons tend to show good convergence performance. In terms of the solver algorithm that updates the weights, adam was used because it works well on datasets with one thousand or more training samples, as suggested by the sickit-learn documentation. To ensure convergence as much as possible, at each step the weights are changed in the direction of the gradient of the error, with respect to that weight. This process is stochastic gradient descent and is guaranteed to converge to a local minimum. Global convergence is not guaranteed, i.e. we can never be sure that we have the best solution, only a good solution.

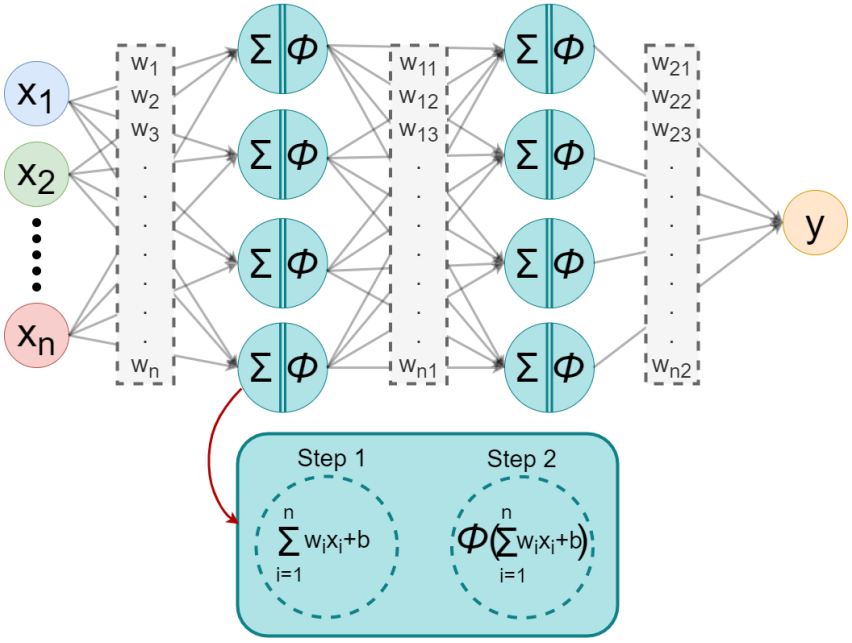


Figure 3.15: Diagram of a neural network

Random forest The RF algorithm is a type of parallel ensemble method. Ensemble methods are a set of techniques that combines multiple ML algorithms into one predictive model. This is done either to decrease bias (boosting), variance (bagging), or improve predictions (stacking). RF is an ensemble of decision trees, which means that RF builds several decision trees and

combines them by a voting process to get a more stable and accurate prediction. RF falls in the family of bagging algorithms.

In this section, decision trees are described since they are the basis of the RF algorithm. A decision tree is a tree where nodes represent features, branches represent decisions, and leaves represent the outcomes as shown in Figure 3.16. The main idea is that the dataset is separated into smaller chunks of data based on the features until all the data points have a final label.

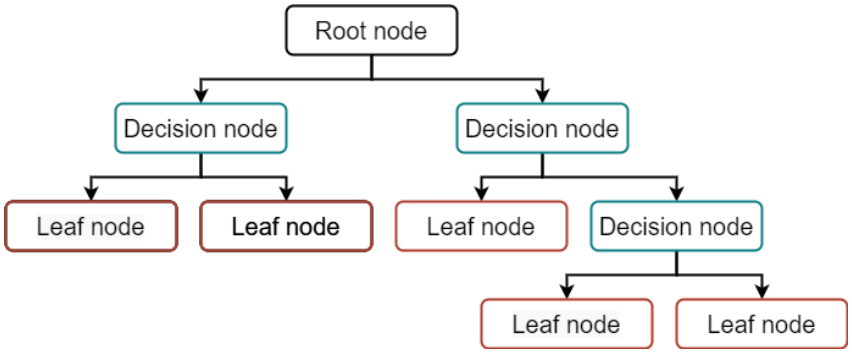


Figure 3.16: Decision tree learning

Trees can be built using different algorithms, such as CART (Classification and Regression Trees). This algorithm works with Gini impurity, which is used to decide the optimal split from the root node, and the subsequent splits. The CART algorithm was used because it does not require to compute a logarithmic function and it is not computationally intensive.

$$Gini = 1 - \sum_{i=1}^{c} p_i^2 \tag{3.9}$$

Where p_i is the probability of class i in a node. Gini impurity measures the quality of the split to select the best possible split from a root node and subsequent splits. A gini impurity of zero is the lowest and best possible impurity. It is achieved when all data points have the same label.

RF produces a great amount of decision trees, however it changes the way the trees are built up. Instead of selecting the question and the threshold that maximally reduces the impurity, a randomness of sampling the training data, in this case the features, is added to the process, which is called bagging. And then, the information from all the trees is combined, and the most common answer is selected as the final label.

k-nearest neighbour kNN algorithm is a non-parametric algorithm, which means that the structure of the model is solely determined from the data. The algorithm stores any available

data and then classifies any new data using a majority vote of its k neighbours. The class that the new point is classified to is the most common among the k neighbours measured by a distance function. Several distance functions are available, such as Euclidean, Minkowski, Hamming, Manhattan, etc. Euclidean and Manhattan distances are demonstrated in Figure 3.17. For this pilot study, the default scikit-learn distance metric was used which is the Euclidean (d2) distance.

$$d_2(X, Y) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{3.10}$$

where x_i and y_i are the coordinates of the points, such as A and B as shown in figure 3.17 that demonstrates two common distance metrics, the Euclidean and the Manhattan.

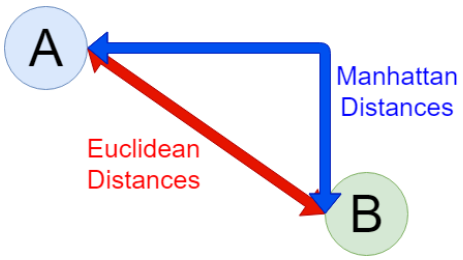


Figure 3.17: Example of Euclidean and Manhattan distances between two points A and B

Gaussian naive bayes This classifier is a probabilistic model that is used for classification problems, and it is based on Bayes theorem. By using this theorem, we can find the probability of y happening, given that X has occurred. y is the hypothesis and X is the evidence. It is assumed that features are independent, so that one particular feature does not affect the other. Bayes theorem is written as:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \tag{3.11}$$

For this problem class probabilities and conditional probabilities are used. Variable X represent the features, which can be mapped to mean, standard deviation, energy, etc.

$$X = (x_1, x_2, x_3, \ldots, x_n) \tag{3.12}$$

Variable y represents the target class, for example the activity performed (e.g. slow walk).

When we substitute for X, and expand using the chain rule we get,

$$P(y|x_1,\ldots,x_n) = \frac{P(x_1|y)P(x_2|y)\ldots P(x_n|y)P(y)}{P(x_1)P(x_2)\ldots P(x_n)} \tag{3.13}$$

For all the dataset, the denominator remains static, and therefore it is removed.

$$P(y|x_1,\ldots,x_n) \propto P(y)\Pi_{i=1}^{n} P(x_i|y) \tag{3.14}$$

In multiclass classification problems,the aim is to find the class with the maximum probability:

$$y = argmax_y P(y)\Pi_{i=1}^{n} P(x_i|y) \tag{3.15}$$

Gaussian Naïve Bayes model uses all the above, and it assumes that the dataset associated with each label follows a Gaussian distribution.

Support vector machine The main idea of the SVM algorithm is that it finds a hyperplane that optimally divides the dataset into two classes by maximising the separation between the two classes. Support vectors are the data points that are nearest to the hyperplane. In the case that these points are removed, the position of the hyperplane will change. Therefore, they can be considered as the critical points of a dataset.

Figure 3.18 demonstrates a classification problem between two features using SVM.

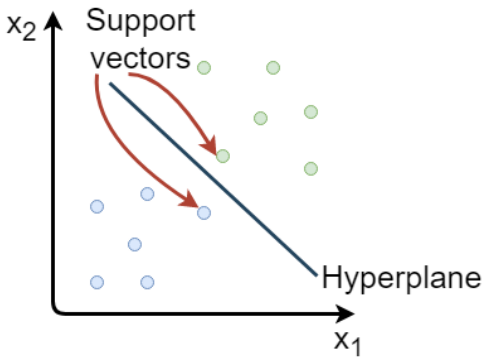


Figure 3.18: Support vector machine for classification using a single hyperplane

The example above is a two feature classification problem. In this case, the hyperplane can be represented as a line that linearly separates data points. For classification with n features, an *n-1* dimensional hyperplane is required. For example, for a 2D (feature) space, hyperplane is

1D, which is a line. For a 3D space, hyperplane is 2D, which is a plane. The hyperplane can be written as:

$$\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_n \cdot x_n = 0 \tag{3.16}$$

if a new data point satisfies: $\beta_0 + \overrightarrow{\beta} \cdot \overrightarrow{x} = 0$, it lies on the hyperplane

$\beta_0 + \overrightarrow{\beta} \cdot \overrightarrow{x} < 0$, it lies above the hyperplane

$\beta_0 + \overrightarrow{\beta} \cdot \overrightarrow{x} > 0$, it lies below the hyperplane

Using only this relation, we cannot find the $\overrightarrow{\beta}$ components and β_0, which are important to calculate the equation of the hyperplane. To find the equation, Maximal Margin Hyperplane (MMH) was used. This hyperplane is the farthest from any training data point, and hence it is the "optimal" as demonstrated in Figure 3.19. To find the MMH, the first step is to calculate the perpendicular distance for each training data point for a given separating hyperplane. The smallest perpendicular distance to a training data point from the hyperplane is called the margin. The MMH is the hyperplane where the margin is the largest.

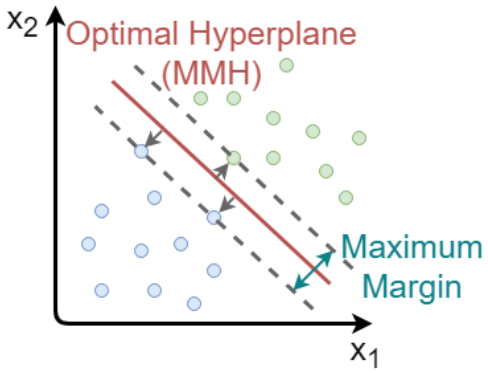


Figure 3.19: Support vector machine for classification using optimal hyperplane and maximum margin

The ideal case is that the data is perfectly separable. However, in real-world problems this is not the case. Therefore, the idea of soft margin classifier, is introduced. This classifier allows some of the data points to be on the incorrect side of the hyperplane, hence it provides a soft margin. For non-linear problems the kernel-trick is used (Boser et al. 1992). The kernel trick is used to transform the data into a higher-dimensional feature space in which the data are linearly separable as shown in Figure 3.20. In this problem, a radial kernel is often used which

is written as:

$$K(\overrightarrow{x_i}, \overrightarrow{x_k}) = exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{kj})^2) \tag{3.17}$$

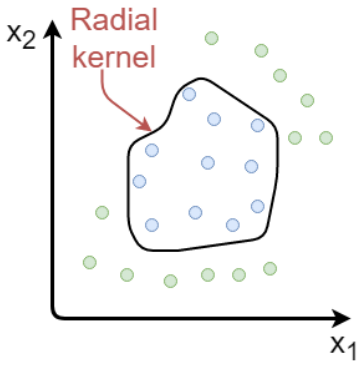


Figure 3.20: Support vector machine: Kernel trick

Cross-validation methodology

For the first three scenarios, a 10-fold cross-validation (CV) was used to evaluate the models. This procedure is called k-fold CV, where k represents the number of groups a given dataset is to be split into. Figure 3.21 demonstrates an example of a 5-fold CV. For this example, the dataset was randomly shuffled and split into five groups. Each group was separated into different training and test sets. Using the CV method, overfitting is avoided because it uses the available training data to generate several mini train-test splits. This is used to test the model with unseen data several times. All the available data was used to make predictions, and the CV method can also be used for hyper-parameter tuning, which is explained in section 3.3.1.1. In this project, stratified k-fold CV was used. Stratification is the process of rearranging the data as to ensure each fold is a good representative of the whole.

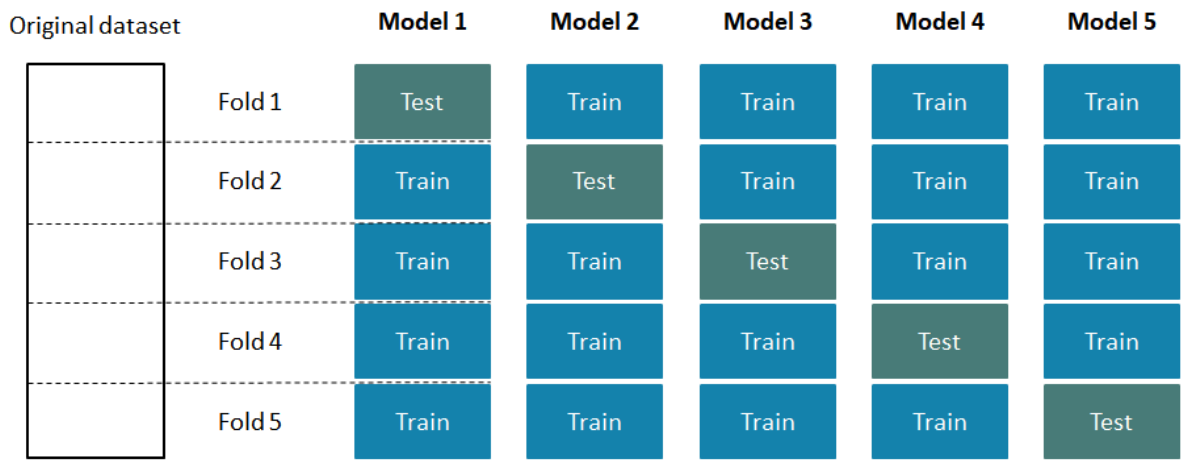


Figure 3.21: Schematic of 5-folds cross-validation

In total, five types of supervised ML algorithms were used in this thesis for condition and activity classification; back-propagation NN, RF, SVM, kNN, GB. These ML algorithms have been commonly used for clinical classification problems (Cleland et al. 2013; Erdaş et al. 2016; Wu et al. 2008). For scenario 1, an SVM classifier was used because this algorithm is useful for non-linear binary classifications, by being able to identify an optimal hyperplane using the kernel trick. For scenarios 2 and 3, all five previously mentioned algorithms were used to identify which one provides the best classification of the activities. Lastly, for scenario 4 only SVM and kNN algorithms were used since they were the best algorithms for activity-type and activity-task classifications respectively. Regarding CV, it was used in all scenarios except scenario 4. The reason for that is because in scenario 4 the training data was completely different from the test data. For example, the training data represented the healthy group, and the test data represented the simulated-pathological group. To perform the ML algorithms, scikit-learn library written in Python was used (Pedregosa et al. 2011).The following Table 3.3 demonstrates the ML algorithms used, and for which scenarios CV technique was used.

Table 3.3: Information about each scenario.

	Condition classification	Activity classification		
	Scenario 1	Scenario 2	Scenario 3	Scenario 4
	N+S/N+S	N/N	S/S	N/S
Cross-Validation	✓	✓	✓	
kNN		✓	✓	✓
NN		✓	✓	
RF		✓	✓	
SVM	✓	✓	✓	✓
GB		✓	✓	

Scenario 1: Condition classification The SVM algorithm was used for the binary condition classification. Normal and simulated-pathological activities were classified into two groups. The data collected from each group was analysed similarly. Following that, all features from both groups were scaled, and PCA technique was performed. The principal components and labels [0, 1] were imported in the SVM algorithm. The algorithm was trained using data from both conditions and tested using unseen data from both normal and simulated-pathological conditions (Figure 3.22).

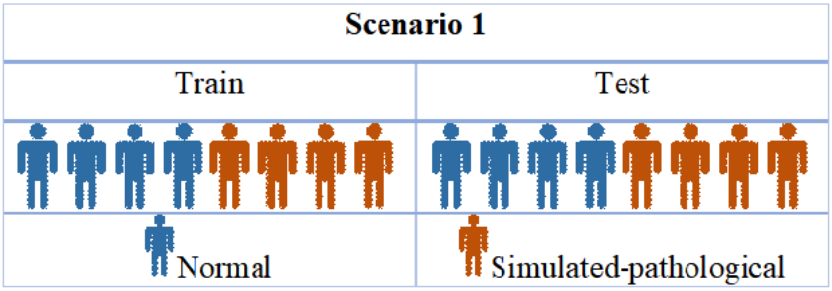


Figure 3.22: Schematic of training and test datasets of scenario 1

The results showed whether new data could be correctly classified as either normal or simulated-pathological.

This binary classification was implemented twice, separately for both wrist and ankle accelerometer data. The classification outcomes for each location were compared to identify which of the two locations provided better results for the condition classification. The location with the best results was intended to be used for further analysis.

Scenarios 2, 3 & 4: Activity classification Activity classification was performed using all five ML algorithms, NN, RF, SVM, kNN and GB. Each ML algorithm was assessed on

its ability to classify both activity type and activity task. For scenario 2 and scenario 3, the classifiers were trained and tested with data from the same group. However, for scenario 4, the classifier was trained with normal data and tested with the simulated-pathological data. This was demonstrated in Figure 3.23.

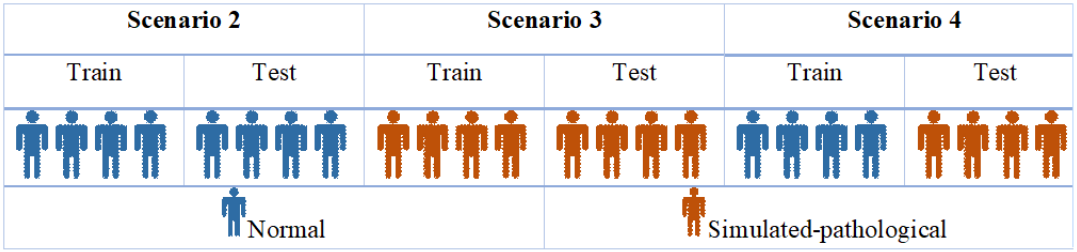


Figure 3.23: Schematic of training and test datasets of scenarios 2, 3 and 4

The theory behind each ML algorithm differs, hence each algorithm works according to its own parameters. The parameters should be selected carefully in order to build a successful classifier that results in high performance, and also avoids over- and under-fitting. To do this, hyper-parameter tuning was used to aid in the selection of the best parameters for each algorithm. Grid search, a hyper-parameter optimisation technique, was used to complete the hyper-parameter tuning process. This means that all the combinations of the different parameters are tested, and the performance of each combination is calculated. Then, the parameters that produced the best performance, for example high accuracy, are selected. The results of the hyper-parameter tuning performed for this thesis will be presented in the following sections.

Table 3.4 shows the parameters for each algorithm that were tuned in order to find their optimal value.

Table 3.4: Parameters tuned for each machine learning algorithm.

Machine learning algorithm	Parameter(s)
k-Nearest Neighbour	Number of neighbours (k)
	Distance measure
Neural Network	Number of layers and neurons per layer
Random Forest	Trees
	Minimum number of samples to split internal node
Support Vector Machine	Regularisation parameter (C)
	Gamma

Each algorithm has several parameters that can be tuned, however the above were selected as they are in principle the most important ones. Table 3.5 includes the definition of each

parameter and how it influences the classifier.

Table 3.5: Definition and influence of each parameter tuned for the machine learning algorithms.

Parameter(s)	Definition	Influence
k-Nearest Neighbour		
Nearest neighbours	The number of nearest points taking into consideration to make a prediction for a new data point	The model is too complex when a single neighbour is used. It becomes simpler when more neighbours are considered
Distance metric	The distance function used to provide a relationship metric between each element in the dataset	
Neural Network		
Neurons (and hidden layers)	The computational unit that have weighted input signals and produce an output using an activation function	The greater number of neurons (and hidden layers), the smoother the decision boundary would be
Random Forest		
Trees	The number of decision trees used	The larger the number of trees used, the better the model will be
Minimum samples to split internal node	The minimum number of classes required to split the internal node of a decision tree	It stops the creation of the tree early, hence avoids overfitting
Support Vector Machine		
Regularisation parameter (C)	The strength of regularisation, which is used to limit the importance of each point	The model is restricted with small C value. Increasing C allows the misclassified points to have a stronger influence on the model, and the decision boundary bends to correctly classify them
Gamma	The amount of curvature of the decision boundary. It is used only with Gaussian RBF kernel	The decision boundary varies slowly with small gamma value, hence yielding a model of low complexity. Increasing gamma, the model becomes more complex

3.2.3.5 Performance evaluation

Classification performance metrics were used to evaluate the performance of each classifier. The metrics used for this study were: (i) accuracy (training and test sets), (ii) F1-score, (iii) precision, (iv) recall, and (v) confusion matrix. Scenarios 1, 2 and 3 used CV, hence the results

reported are the mean and confidence intervals (CIs). CIs represent a range of values that the true parameter can be among those values with an associated confidence level. In this case, 95% CIs were used.

3.3 Results

Participants′ ages range from 22 to 66 years (32.7±12.7) with 14 identified as female, 16 identified as male. Their height range from 1.60 to 1.90 meters (171.5±7.1) and weight range from 49 to 105 kilograms (69.2±13.6). In terms of the dominant hand, the right hand is dominant for 29 participants and the left hand is dominant for one participant. In terms of the dominant leg, the right leg is dominant for 25 participants and the left leg is dominant for five participants.

In this section, two analyses are presented. Analysis 1 is performed to identify the best sensor location (wrist or ankle). Analysis 2 is an in-depth analysis for the best location identified in analysis 1. Table 3.6 describes in detail the two analyses.

Table 3.6: Information about the two analyses presented in results section.

Scenario	Analysis 1 (wrist or ankle)	Analysis 2 (wrist)
Scenario 1	Whole set	Whole set Dynamic subset Slow walk subset Normal walk subset Fast walk subset Stair ascent subset Stair descent subset
Scenario 2	activity-*type* classification using SVM	activity-*type* classification using SVM, NN, RF, kNN and GB activity-*task* classification using SVM, NN, RF, kNN and GB
Scenario 3	activity-*type* classification using SVM	activity-*type* classification using SVM, NN, RF, kNN and GB activity-*task* classification using SVM, NN, RF, kNN and GB
Scenario 4		activity-*type* classification using SVM activity-*task* classification using kNN

The following three figures, 3.24, 3.25 and 3.26, demonstrate the range of the three acceleration axes.

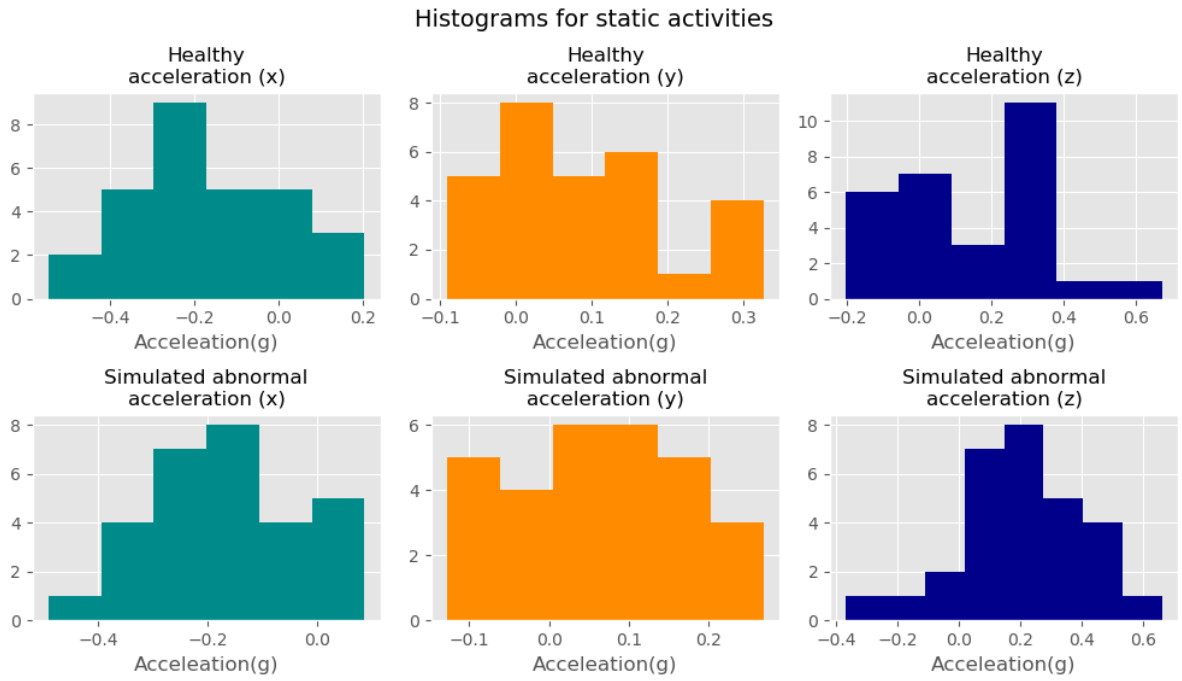


Figure 3.24: Histogram plots for static activities of healthy and simulated-pathological acceleration

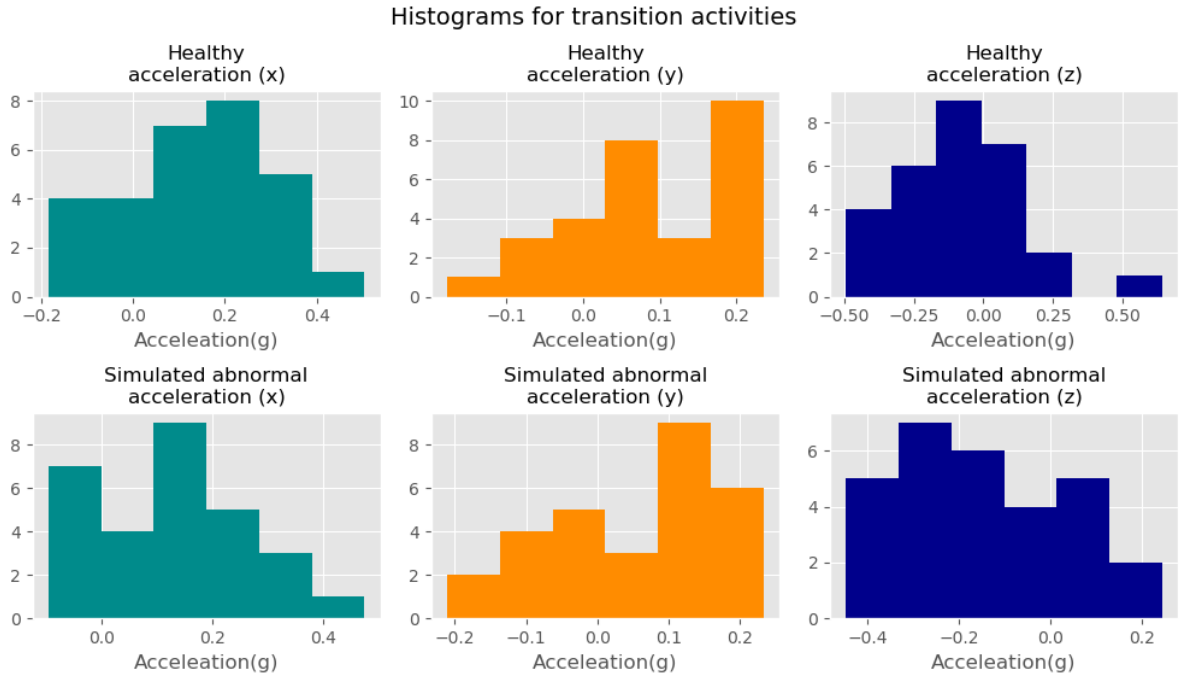


Figure 3.25: Histogram plots for transition activity of healthy and simulated-pathological acceleration

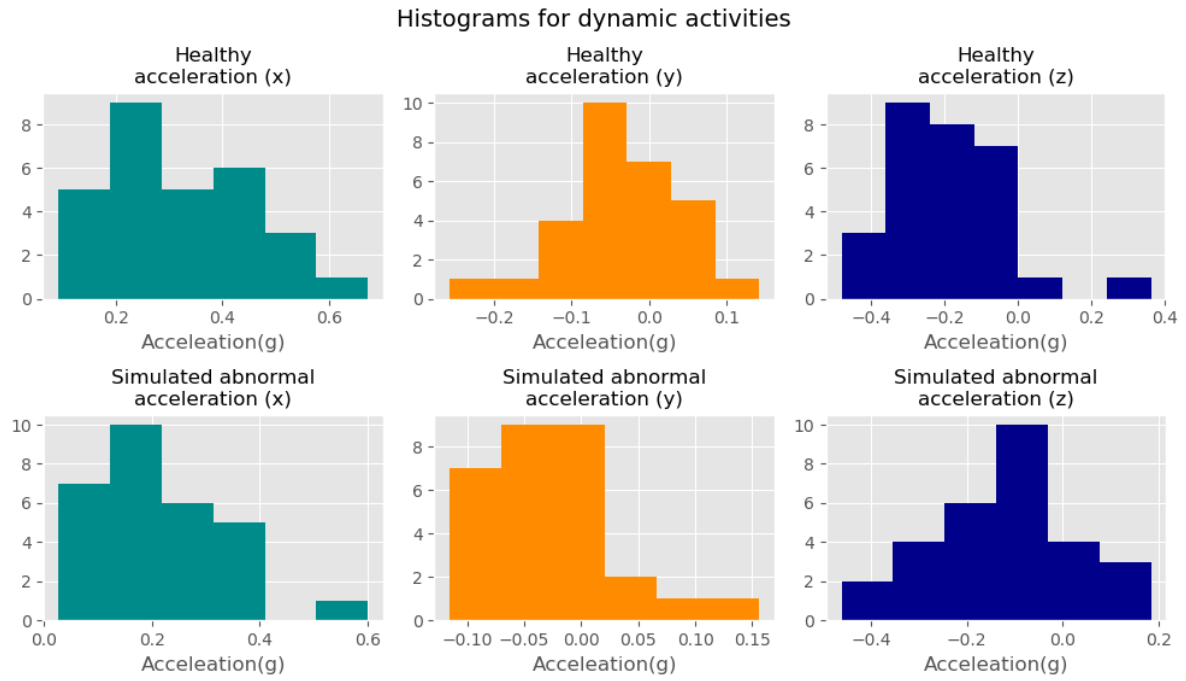


Figure 3.26: Histogram plots for dynamic activities of healthy and simulated-pathological acceleration

The following figures, 3.27, 3.28 and 3.29, demonstrate the mean and standard deviation of the features calculated based on the data of all volunteers.

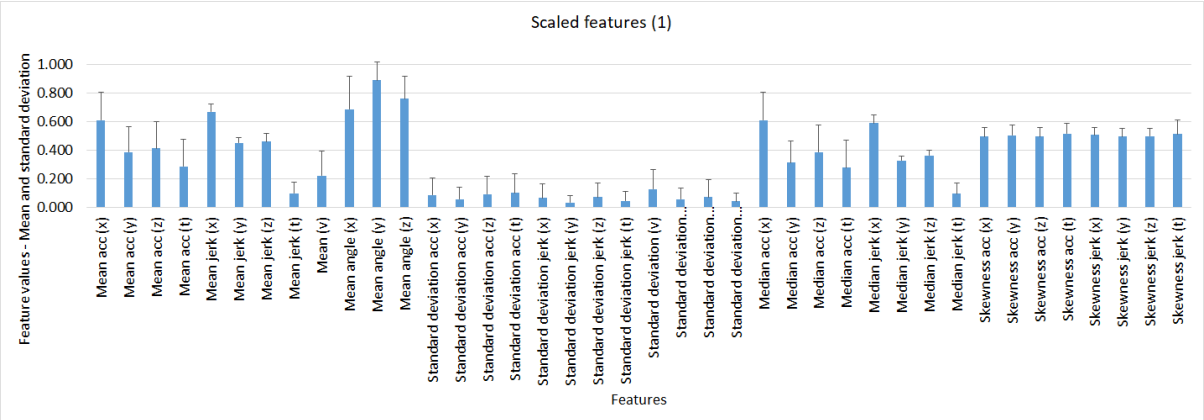


Figure 3.27: Scaled features before applying principal component analysis

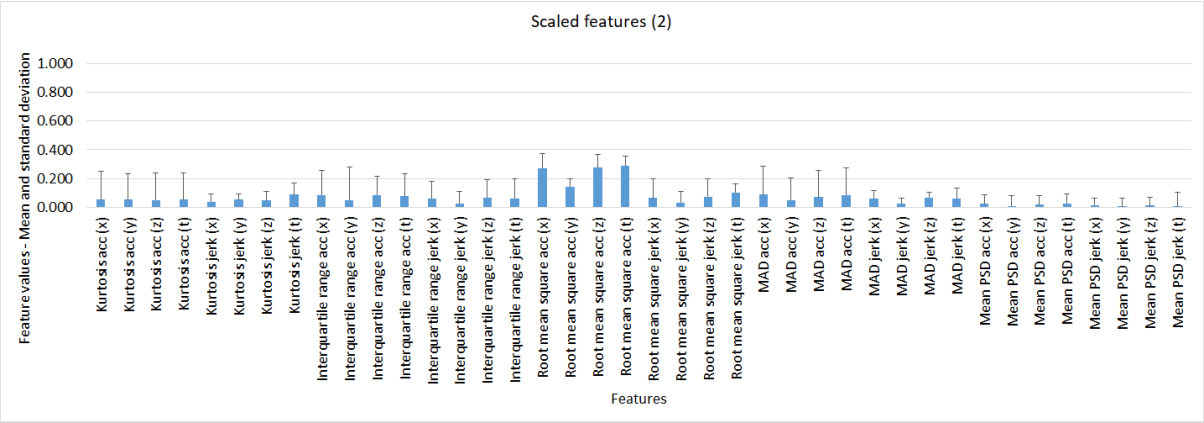


Figure 3.28: Scaled features before applying principal component analysis

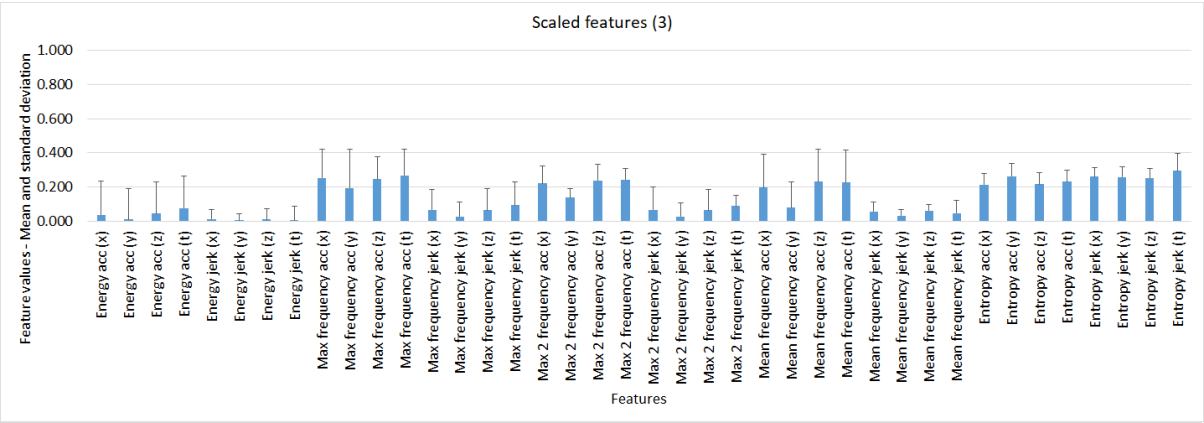


Figure 3.29: Scaled features before applying principal component analysis

3.3.1 Analysis 1: Identify the best location (wrist or ankle)

3.3.1.1 Scenario 1: Condition classification using the whole dataset that includes all the activities

Principal component analysis The number of principal components selected was based on a threshold of 95% of explained variance of the new derived principal components. Figure 3.30 demonstrates the number of components needed to explain variance for each location performing a binary classification. Data recorded from wrist and ankle locations needed 28 and 22 principal components respectively. These results suggested that data from the wrist location had greater variability than data from the ankle.

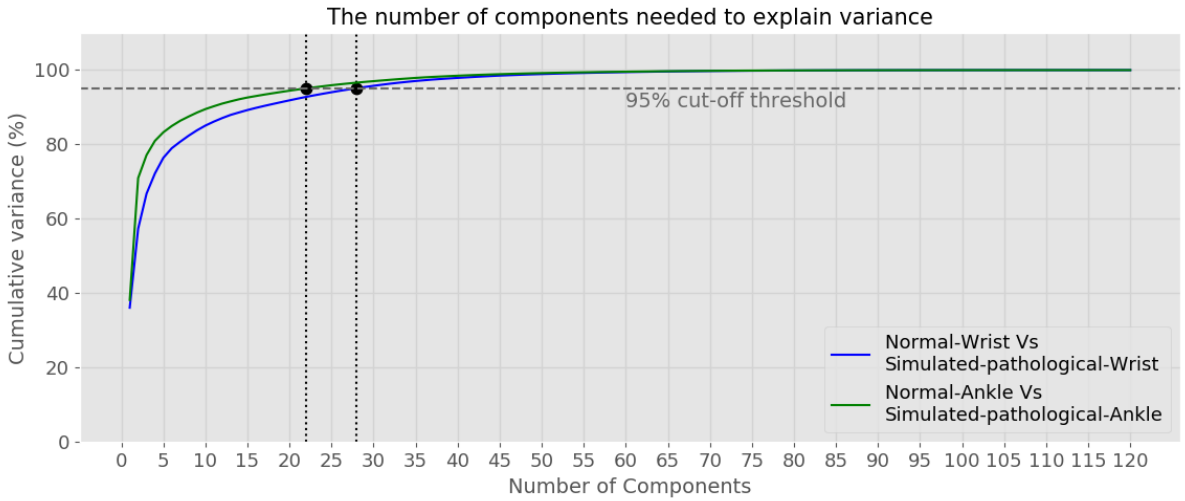


Figure 3.30: Number of components required for 95% explained variance for normal against simulated-pathological conditions in both wrist and ankle locations

The following two Tables 3.7, 3.8 demonstrate the top five features that most influenced the first three principal components. In the wrist location, the first principal component had a strong association with features that represent the total acceleration magnitude. The second component had a strong association with the entropy and standard deviation features of different acceleration axes. The third component had an association with features from the acceleration in y-axis and the mean angle of z-axis. The ankle location results differed in that the first component had a strong association with three features from the total acceleration magnitude and the RMS of y- and x-axis of acceleration. The second component had an association with the acceleration entropy in all directions, and the standard deviation of the angular velocity. The third component had associations with several acceleration features in the z-axis and x-axis.

Table 3.7: Top three principal components with the top five features for wrist location for condition classification.

PC1	PC2	PC3
Acc: Mean (M)	Acc: Entropy (M)	Acc: Max F (y)
Acc: Median (M)	Acc: Entropy (x)	Angle: Mean (z)
Acc: Max F (M)	Acc: Std (M)	Acc: RMS (y)
Acc: RMS (M)	Acc: Entropy (z)	Acc: Mean (y)
Acc: Max 2nd F (M)	Acc: Std (z)	Acc: Max 2nd F (y)

Table 3.8: Top three principal components with the top five features for ankle location for condition classification.

PC1	PC2	PC3
Acc: Median (M)	Acc: Entropy (M)	Acc: Max F (z)
Acc: RMS (y)	Acc: Entropy (y)	Acc: RMS (z)
Acc: Max F (M)	Acc: Entropy (z)	Acc: Median (x)
Acc: RMS (x)	Acc: Entropy (x)	Acc: Mean (x)
Acc: Mean (M)	Ang. vel: Std	Acc: Max 2nd F (z)

Hyper-parameter tuning Hyper-parameter tuning was performed by checking all the possible combinations among the values presented in Table 3.9. This step was performed in order to build a model that will be able to generalise to new data.

Table 3.9: Gamma and C parameters of support vector machine used for hyperparameter tuning for condition classification.

	Parameters			
Gamma	0.01	0.1	1	10
C	0.01	0.1	1	10

Figures 3.31 and 3.32 were used to aid in the decision of the optimal parameters to build a model that is able to generalise. This means that the model is able to make accurate predictions on unseen data. There is a trade-off of model complexity against training and test accuracy. If the model is simple, there is a chance that it will under-fit data, if the model is too complex then there is a chance for over-fitting data, hence it is crucial to find the "sweet spot". The following figures demonstrate the results from the Grid search. For the specific cross-validation, K was equal to 10, hence there are 10 small upward triangles to demonstrate the result of each subset. Additionally, the empty downward triangle represents the mean accuracy of the 10 K-folds. This is done for both training (blue) and test (grey) sets. Based on the results from both figures, the optimal parameters found were gamma equals to 1 and C equals to 10. The reason for selecting these parameters is because the accuracy of the training dataset was not 100%, but instead it was 96.5% and 93.5% for the wrist and ankle locations respectively. Additionally, the accuracies of the test dataset were 94.9% and 92.6% for the wrist and ankle locations respectively, which are slightly lower than the accuracy of the training dataset. For these reasons, there is high possibility that we might have avoided the problem of under- and over-fitting.

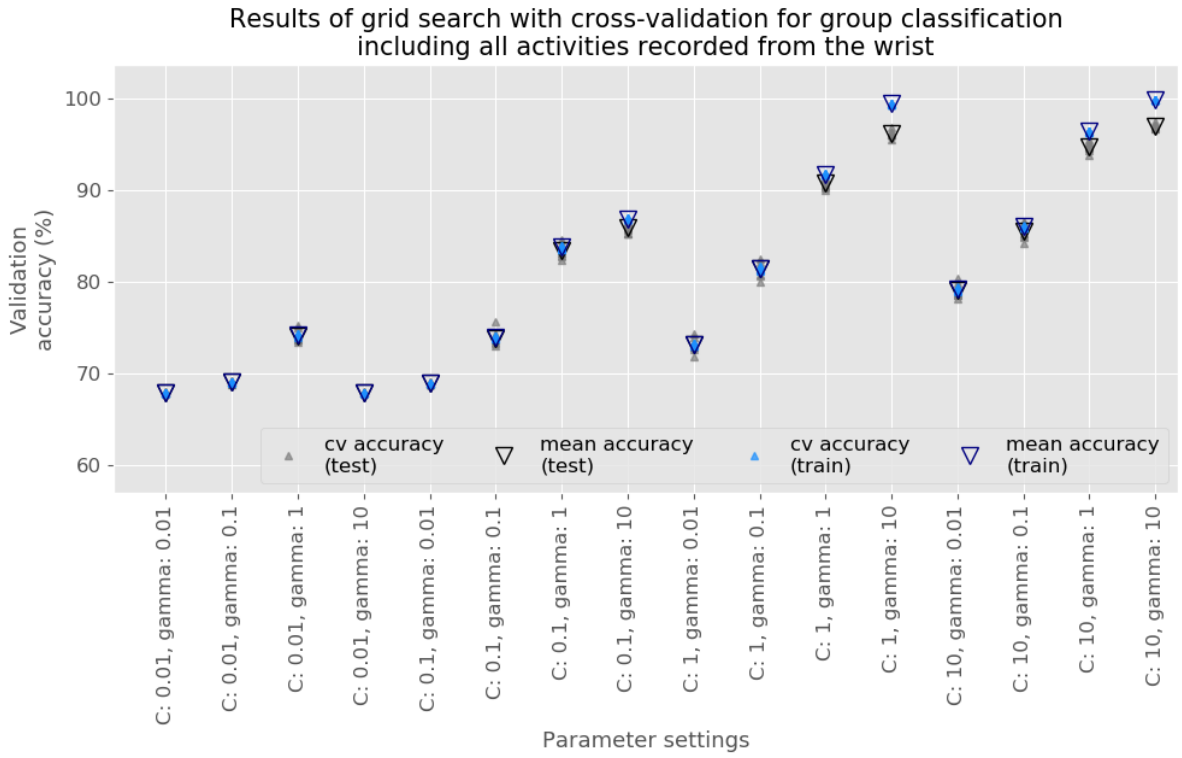


Figure 3.31: Hyperparameter tuning results of support vector machine for condition classification using all activities for wrist location using 10 K-fold CV

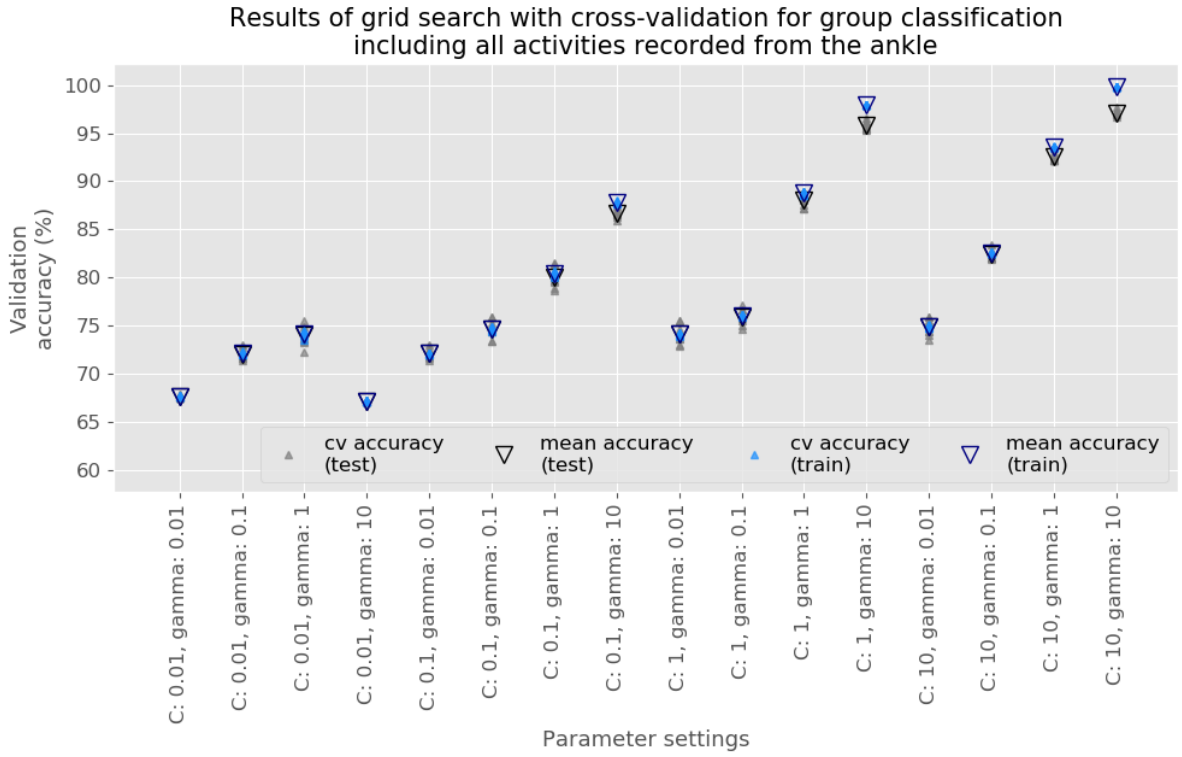


Figure 3.32: Hyperparameter tuning results of support vector machine for condition classification using all activities for ankle location using 10 K-fold CV

Performance evaluation Table 3.10 demonstrates all performance metrics regarding the normal and simulated-pathological groups with the accelerometer worn at the wrist and ankle location. Across both locations, the classifier algorithm correctly classified the conditions (normal and simulated-pathological) with an accuracy of more than 92.5%. High degrees of recall (avoiding false negatives) and precision (limiting false positives) scores were obtained for both locations. A high F1-score (balance between recall and precision) was noted as well, since it is calculated based on precision and recall scores.

Table 3.10: Results for condition classification for both wrist and ankle locations [Mean (95% CI)].

Performance metrics	Wrist	Ankle
Accuracy (test set)	0.949 (0.949-0.950)	0.926 (0.925-0.927)
Accuracy (training set)	0.965 (0.965-0.965)	0.935 (0.935-0.936)
F1-score	0.942 (0.941-0.942)	0.915 (0.914-0.916)
Precision	0.943 (0.942-0.944)	0.922 (0.921-0.923)
Recall	0.940 (0.940-0.941)	0.909 (0.908-0.910)

3.3.1.2 Scenarios 2 and 3: Activity-*type* classification

Activity type clasification was performed to classify the general types of activities, such as static, transition, and dynamic.

Principal component analysis A threshold of 95% of explained variance was used to identify the number of principal components. For this case, each individual group was of interest, since activity classification was performed for each group separately. The groups were: a) normal – wrist, b) simulated-pathological – wrist, c) normal – ankle, d) simulated-pathological – ankle.

The required number of principal components were 25, 30, 17 and 26 for groups normal – wrist, simulated-pathological – wrist, normal – ankle and simulated-pathological – ankle respectively. The results are demonstrated in Figure 3.33. These results suggest that data that represented the simulated-pathological groups was more varied than the associated data from the normal groups.

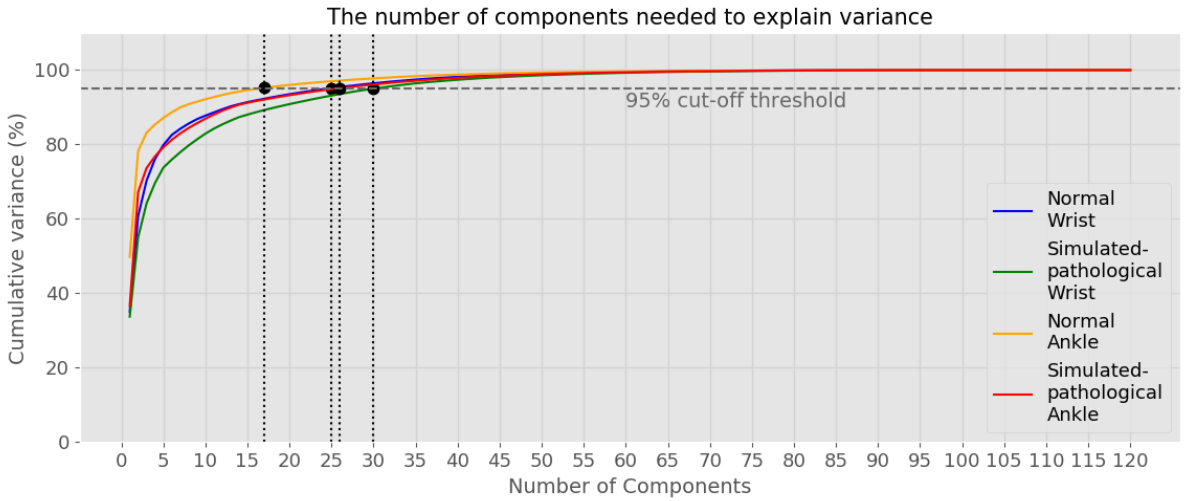


Figure 3.33: Number of components required for 95% explained variance for activity classification of normal condition and simulated-pathological condition in both wrist and ankle locations

Again, the following four Tables 3.11, 3.12, 3.13, 3.14 demonstrate the top five features that formed the first three principal components. In this case, the components were used to reduce the number of features to classify the activities for each group individually. The first two principal components, of the normal group with the wrist-worn device, had a strong association with the total acceleration magnitude. The third component had a strong association with features from the y-axis acceleration. On the contrary, the first component of the normal group with ankle-worn device had a significant association with features from the jerk signal in x-axis. The second component had a strong association with the first maximum frequency and RMS of acceleration in y- and x-axis. The third component had an association with the acceleration signal in z-axis with features from frequency-domain.

Table 3.11: Top three principal components with the top five features for wrist location for activity classification under normal condition.

PC1	PC2	PC3
Acc: Median (M)	Acc: RMS (M)	Acc: Max F (y)
Acc: Mean (M)	Acc: RMS (x)	Acc: Mean (y)
Acc: Max F (x)	Acc: Mean (M)	Acc: RMS (y)
Acc: Max F (M)	Acc: Mean F (M)	Acc: Median (y)
Acc: RMS (M)	Acc: Median (M)	Angle: Mean (z)

Table 3.12: Top three principal components with the top five features for ankle location for activity classification under normal condition.

PC1	PC2	PC3
Jerk: Mean F (x)	Acc: Max F (y)	Acc: Max F (z)
Jerk: STD (x)	Acc: Max F (x)	Acc: RMS (z)
Jerk: RMS (x)	Acc: Median (M)	Acc: Max 2nd F (z)
Jerk: Max 2nd F (x)	Acc: RMS F (y)	Acc: Mean F (z)
Jerk: Max F (x)	Acc: RMS (x)	Acc: Mean (x)

The first principal component of the simulated-pathological group with the wrist-worn device, had a strong association with the total acceleration magnitude. Similar results were observed for this location in the normal group. The second component had a large association with features such as standard deviation and entropy from the acceleration signal. The third component had a strong association with features from the y-axis acceleration. The first component of the simulated-pathological group with ankle-worn device also had a large association with total acceleration magnitude features. The second component had a large association with acceleration mean frequency in y- and M-axis, as well as features such as standard deviation and RMS of the jerk signal. The third component had a strong association with the acceleration signal in z- and x-axis.

Table 3.13: Top three principal components with the top five features for wrist location for activity classification under simulated-pathological condition.

PC1	PC2	PC3
Acc: Mean (M)	Acc: Entropy (M)	Acc: Max F (y)
Acc: Median (M)	Acc: Std (y)	Acc: RMS (y)
Acc: Max F (M)	Acc: Entropy (x)	Acc: Max 2nd F (y)
Acc: RMS (M)	Acc: Std (z)	Angle: Mean (z)
Acc: Max F (z)	Acc: Std (M)	Acc: Mean (y)

Table 3.14: Top three principal components with the top five features for ankle location for activity classification under simulated-pathological condition.

PC1	PC2	PC3
Acc: Median (M)	Acc: Mean F (y)	Acc: Median (x)
Acc: Max F (M)	Acc: Std (M)	Acc: Mean (x)
Acc: Mean (M)	Jerk: Std (y)	Acc: Max F (z)
Acc: Max F (y)	Jerk: RMS F (y)	Acc: RMS (z)
Acc: RMS (M)	Acc: Mean F (M)	Acc: Max 2nd F (z)

Hyper-parameter tuning The same approach was used to find the optimal parameters for each mounting location group for the activity classification. Based on the results from both figures, the optimal parameters found were gamma equals to 1 and C equals to 10.

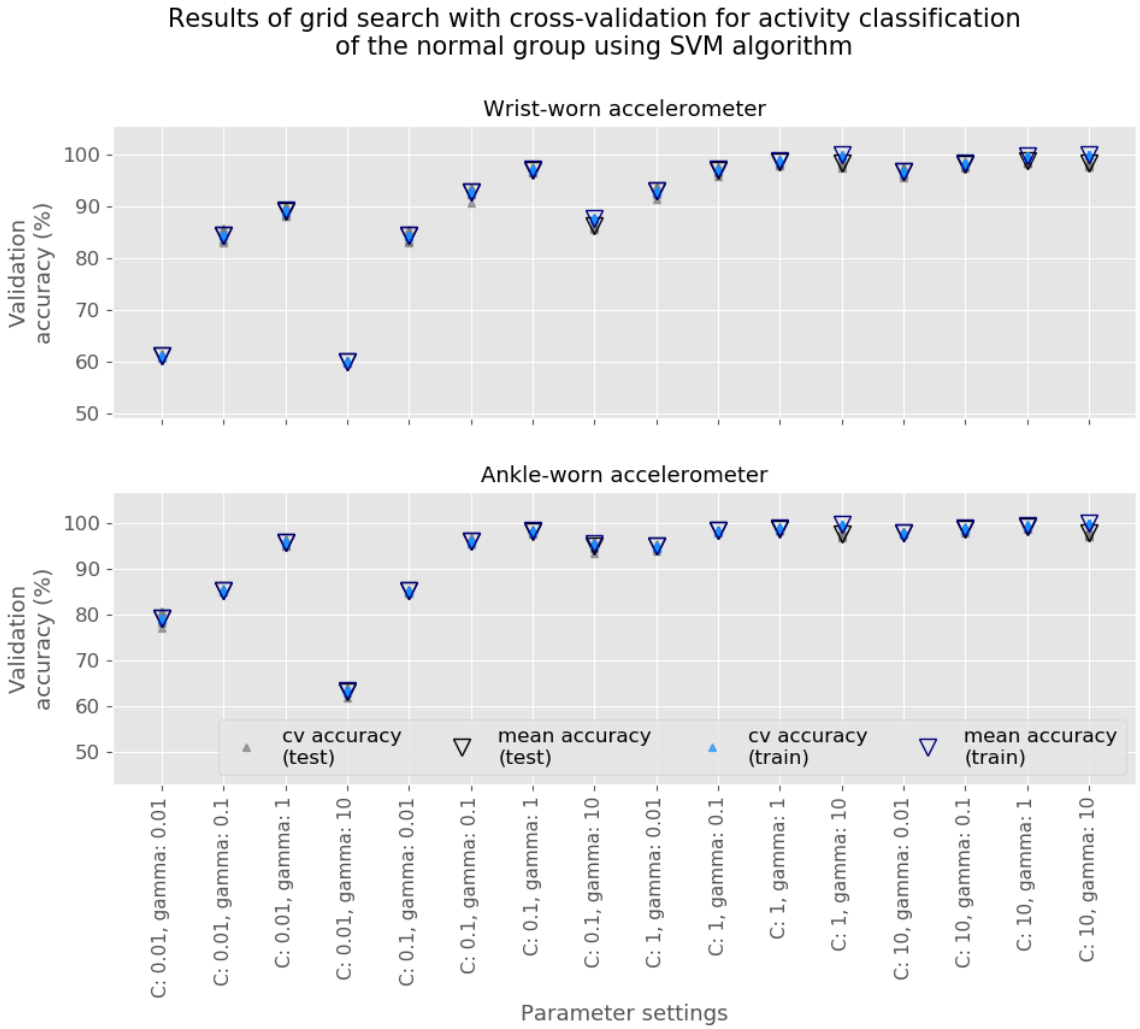


Figure 3.34: Hyperparameter tuning results of support vector machine for activity classification under normal condition for both wrist and ankle locations using 10 K-fold CV

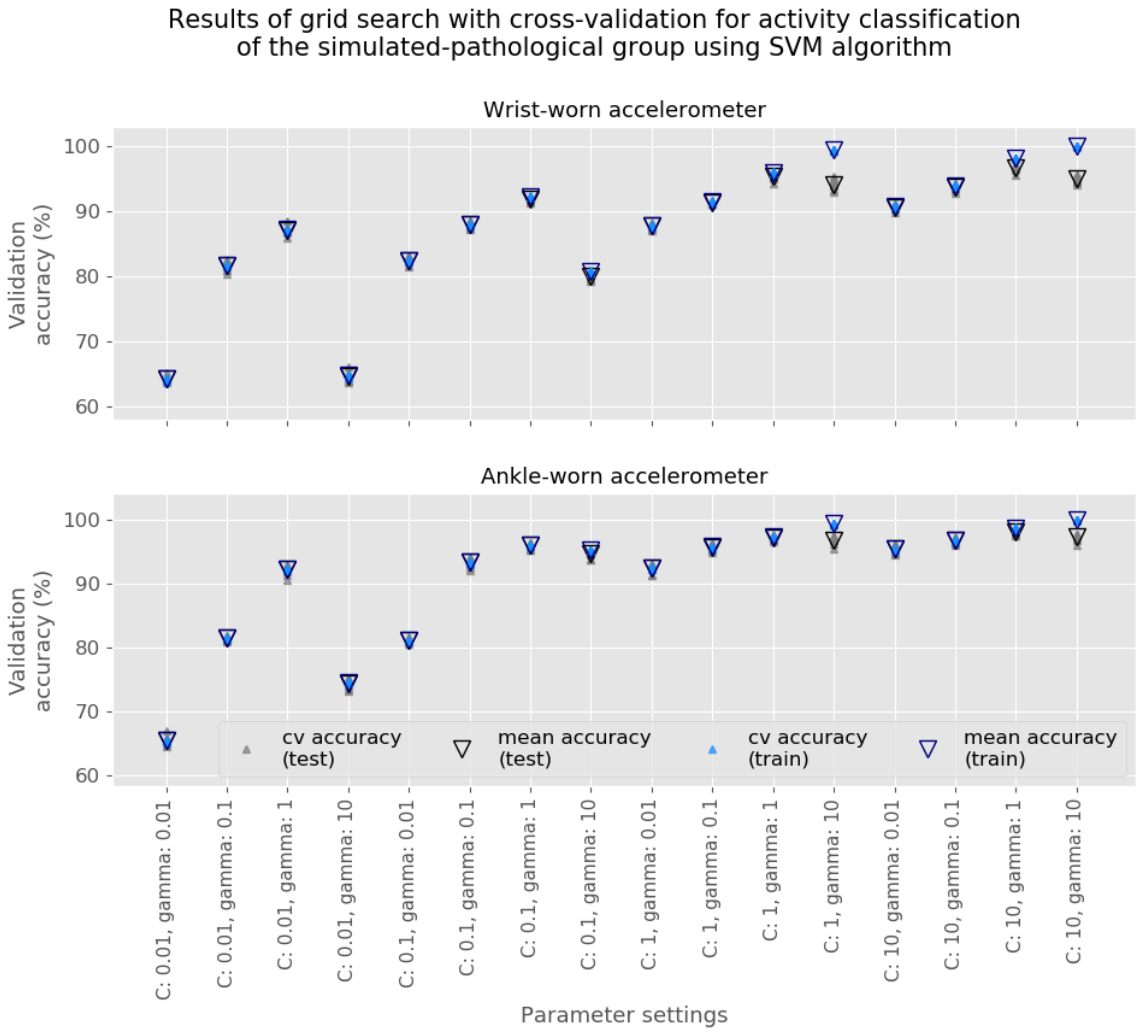


Figure 3.35: Hyperparameter tuning results of support vector machine for activity classification under simulated-pathological condition for both wrist and ankle locations using 10 K-fold CV

Performance evaluation For the scope of this thesis, activity recognition was one of the most important aspects. Tables 3.15 and 3.16 demonstrate the results generated from the activity-*type* recognition of the four groups individually (normal – wrist, simulated-pathological – wrist, normal – ankle, simulated-pathological – ankle). In section 3.2.3.4, four algorithmic scenarios were described and scenario 2 and 3 were applied for the normal and simulated-pathological conditions respectively.

The following performance metrics represent the results from both locations of the normal group. Both locations achieved very high performance in all the tested metrics ($> 97\%$). The performance metrics from the ankle location were slightly better than the metrics of the wrist location, with differences between 0.003 and 0.008.

Table 3.15: Results for activity classification under normal condition for both wrist and ankle locations [Mean (95% CI)].

Performance metrics	Wrist	Ankle
Accuracy (test set)	0.984 (0.983-0.984)	0.987 (0.987-0.988)
Accuracy (training set)	0.989 (0.989-0.989)	0.989 (0.989-0. 989)
F1-score	0.971 (0.970-0.973)	0.979 (0.978-0. 979)
Precision	0.971 (0.970-0.973)	0.978 (0.977-0.978)
Recall	0.971 (0.969-0.973)	0.979 (0.978-0.981)

A similar pattern of results was observed for both locations for the simulated-pathological group. The performance of both locations was very good overall, although the scores were slightly lower than seen in the normal group. The ankle location again provided slightly superior scores than those obtained from the wrist location with differences between 0.014 and 0.018.

Table 3.16: Results for activity classification under simulated-pathological condition for both wrist and ankle locations [Mean (95% CI)].

Performance metrics	Wrist	Ankle
Accuracy (test set)	0.967 (0.966-0.968)	0.981 (0.980-0.982)
Accuracy (training set)	0.982 (0.982-0.982)	0.987 (0.987-0.987)
F1-score	0.958 (0.957-0.959)	0.971 (0.970-0.972)
Precision	0.964 (0.963-0.965)	0.972 (0.970-0.973)
Recall	0.952 (0.951-0.954)	0.970 (0.969-0.971)

The ankle location performed slightly better than the wrist in all performance metrics. While the ankle was slightly better however, the performance at both locations was high (>95%) enough that either could potentially be used for activity recognition in any of the groups, and condition classification.

This was an important finding because previous discussions with patients through patient and public involvement (PPI) work indicated that patients had a strong preference for a wrist worn device rather than ankle worn (See appendix A). As performance was similar between locations, but patients would only wear one device in the real world scenario, it was decided to focus subsequent analysis on the data from wrist worn devices.

3.3.2 Analysis 2: In-depth analysis for wrist location

3.3.2.1 Scenario 1: Condition classification

Principal component analysis The number of essential principal components was calculated for the dataset including all the activities, the sub-dataset including only the dynamic activities, and the sub-dataset for each individual dynamic activity. As already mentioned, a threshold of 95% of explained variance was used to derive the new principal components. Table 3.17 demonstrates the number of essential principal components.

Table 3.17: Number of components required for 95% explained variance for activity classification under normal condition in different activity datasets.

Dataset	Principal components
Whole	28
Dynamic	24
Slow walk	34
Normal walk	31
Fast walk	30
Stair ascent	33
Stair descent	32

Hyper-parameter tuning The range of hyper-parameters for the whole dataset and dynamic sub-dataset was slightly different from the individual activity sub-datasets. The difference was that C and gamma were up to two instead of ten. This is because the results were poor when C=10 and gamma=10, therefore smaller values were used instead. Figures 3.36 to 3.41 display the results from the hyper-parameter tuning of the SVM algorithm.

CV was used to calculate the results, for the whole set and the dynamic subset the data was split into 10-folds. This means that the dataset was separated into ten smaller groups. However, for the individual activity subsets, the data was split into 5-folds since the amount of data was small.

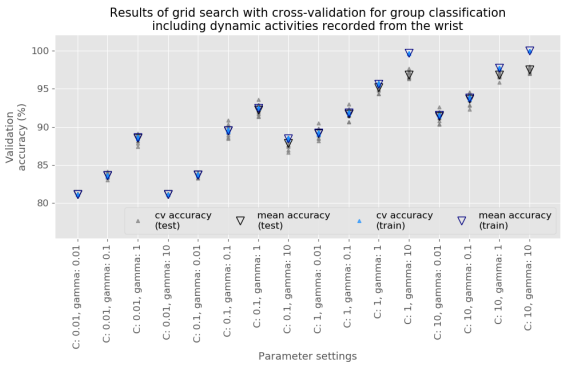


Figure 3.36: Hyperparameter tuning results of SVM for condition classification using dynamic activities for wrist location using 10 K-fold CV

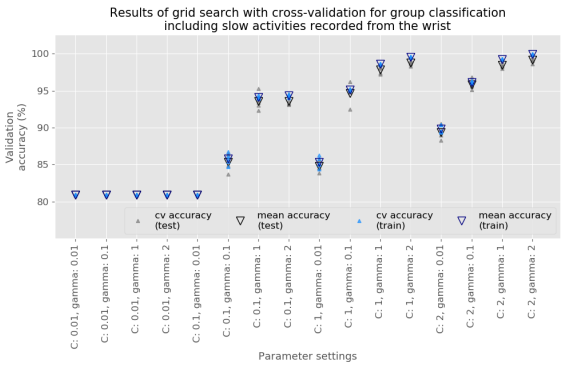


Figure 3.37: Hyperparameter tuning results of SVM for condition classification using slow walk activity for wrist location using 5 K-fold CV

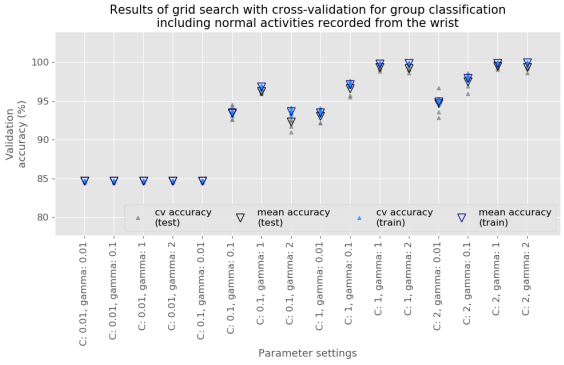


Figure 3.38: Hyperparameter tuning results of SVM for condition classification using normal walk activity for wrist location using 5 K-fold CV

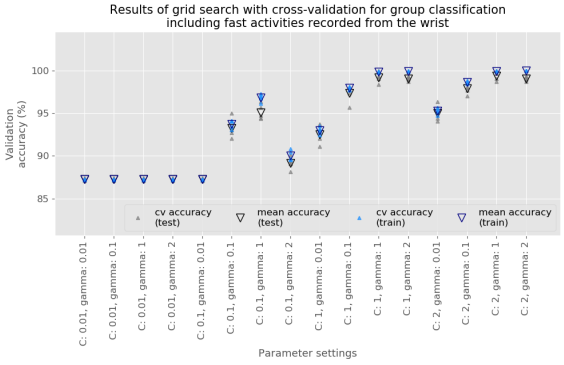


Figure 3.39: Hyperparameter tuning results of SVM for condition classification using fast walk activity for wrist location using 5 K-fold CV

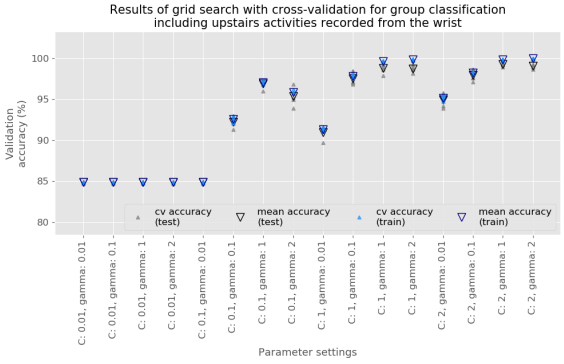


Figure 3.40: Hyperparameter tuning results of SVM for condition classification using ascending stairs activity for wrist location using 5 K-fold CV

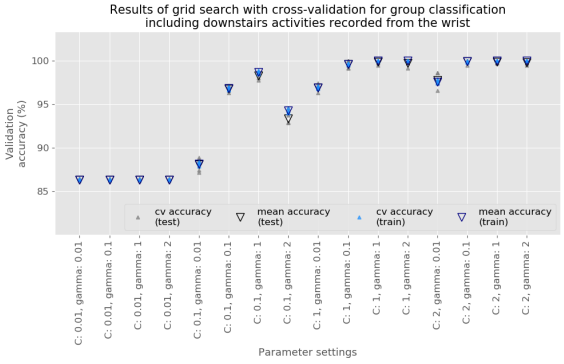


Figure 3.41: Hyperparameter tuning results of SVM for condition classification using descending stairs activity for wrist location using 5 K-fold CV

Table 3.18 demonstrates the required parameters for all the cases examined.

Table 3.18: Best gamma and C parameters for support vector machine for condition classification of different activity datasets.

Dataset	Gamma	C
Whole	1	10
Dynamic	1	10
Slow walk	1	2
Normal walk	1	1
Fast walk	0.1	2
Stair ascent	1	1
Stair descent	0.1	1

Performance metrics

Whole set Table 3.10 demonstrates the results for the condition classification of the wrist-worn device including all nine activities. The classifier was fed with features representing all nine activities into multiple windows of two seconds. Stratified (shuffled) CV was used to ensure that the training and test sets have the same proportion of the feature of interest as in the original dataset. The shuffling of data was used to mix the dataset, however this might be a problem if data from the same participant exists in the train fold as well as the test fold. This might be an issue since the model might not be generalisable. As already mentioned, the SVM classifier achieved high performance scores in order to differentiate the activities performed from the two conditions using the whole dataset of activities. The classifier achieved an accuracy score of 94.9% and an F1-score of 94.2%.

A confusion matrix is a useful way to visually represent the results from the classification. The values on the main diagonal of the confusion matrix represent the correct classifications, while other values show how many samples of one class were misclassified as another class.

Figure 3.42 demonstrates the classification results of the wrist location. The SVM algorithm was confused by predicting the normal condition as simulated-pathological condition for 684/8099 times. The algorithm also misclassified the simulated-pathological condition as normal condition for 595/17125 times. The individual precision and recall scores between the two conditions were very close. However, the simulated-pathological condition had greater precision and recall scores compared to the normal condition. The precision scores of the normal and simulated-pathological conditions were 0.926 and 0.960 respectively and the recall scores of the normal and simulated-pathological conditions were 0.916 and 0.965.

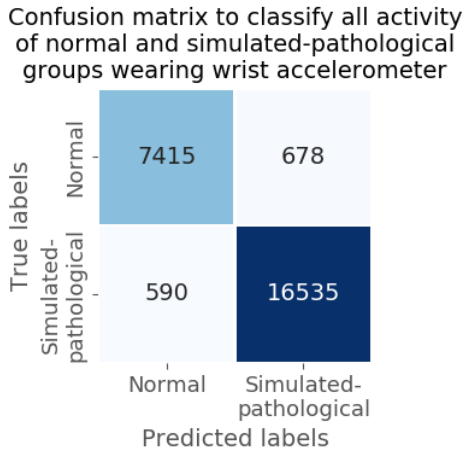


Figure 3.42: Confusion matrix for condition classification of the dataset including all nine activities

Dynamic subset Table 3.19 demonstrates the results for the condition classification of the wrist-worn device, however using only the activities from the dynamic category. The classifier achieved slightly higher performance scores using only the dynamic activities instead of all the activities. Accuracy and F1-score were achieved as 96.7% and 94.3% respectively.

Table 3.19: Results for condition classification of the dataset including dynamic activities [Mean (95% CI)].

Performance metrics	Dynamic subset
Accuracy (test set)	0.967 (0.966-0.968)
Accuracy (training set)	0.977 (0.977-0. 977)
F1-score	0.943 (0.942-0.945)
Precision	0.961 (0.959-0.962)
Recall	0.928 (0.926-0.930)

The SVM algorithm misclassified the normal condition as simulated-pathological condition for 291/2184 times as shown in Figure 3.43. Additionally, the algorithm misclassified the simulated-pathological condition as normal condition for 93/9399 times. The individual precision and recall scores of the simulated-pathological condition had greater precision and recall scores from the normal condition. The precision scores of the normal and simulated-pathological conditions were 0.952 and 0.970 respectively and the recall scores of the normal and simulated-pathological conditions were 0.866 and 0.990.

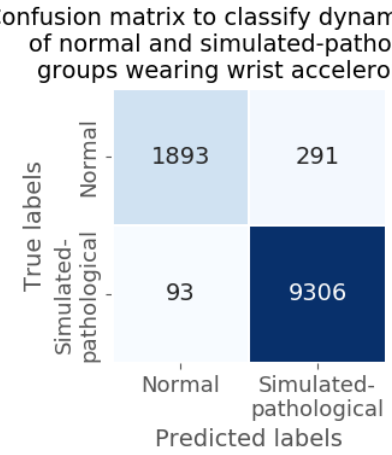


Figure 3.43: Confusion matrix for condition classification of the dataset including dynamic activities

Individual dynamic subsets For this case, each individual activity within the dynamic category was tested individually. This was done to understand whether a specific task is better or worse at differentiating the two conditions. The results in Table 3.20 demonstrate that all the dynamic activities performed well.

Table 3.20: Results for condition classification of the dataset for individual dynamic activities [Mean (95% CI)].

Performance metrics	Slow walk	Normal walk	Fast walk	Stair ascent	Stair descent
Accuracy (test set)	0.984 (0.983-0.985)	0.994 (0.993-0.996)	0.981 (0.977-0.986)	0.984 (0.981-0.998)	0.991 (0.990-0.992)
Accuracy (training set)	0.993 (0.992-0.993)	0.998 (0.998-0.999)	0.986 (0.986-0.987)	0.997 (0.996-0.997)	0.998 (0.997-0.998)
F1-score	0.974 (0.972-0.976)	0.989 (0.986-0.992)	0.958 (0.948-0.967)	0.968 (0.961-0.976)	0.991 (0.990-0.992)
Precision	0.979 (0.977-0.981)	0.990 (0.987-0.994)	0.970 (0.958-0.982)	0.982 (0.974-0.989)	0.991 (0.990-0.992)
Recall	0.969 (0.965-0.974)	0.988 (0.984-0.991)	0.947 (0.937-0.957)	0.956 (0.948-0.964)	0.991 (0.990-0.992)

The following Figures 3.44 to 3.48 depicts the confusion matrix for each of the five individual activities. In all cases, the algorithm misclassified more times the normal condition compared to the simulated-pathological condition than vice versa. Moreover, there were higher individual precision and recall scores in the simulated-pathological (0.987, 0.993) condition than the normal condition (0.970, 0.946).

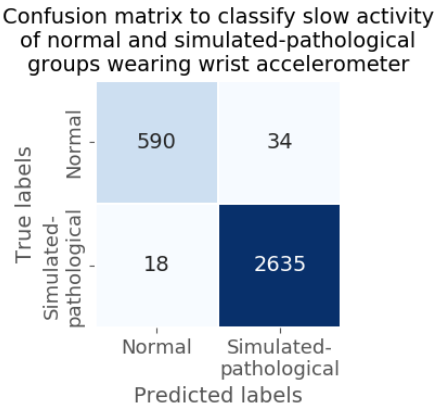


Figure 3.44: Confusion matrix for condition classification of the dataset for slow walk activity

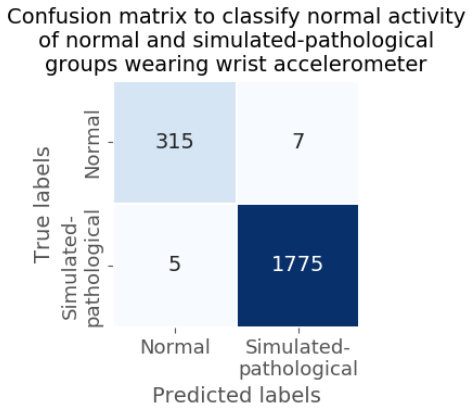


Figure 3.45: Confusion matrix for condition classification of the dataset for normal walk activity

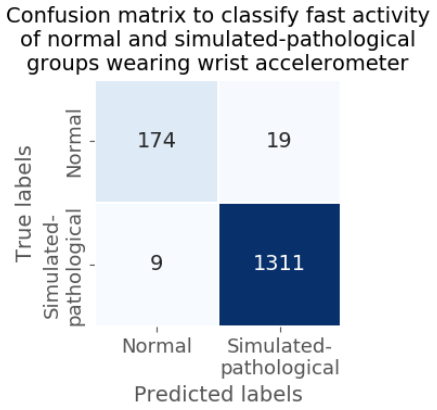


Figure 3.46: Confusion matrix for condition classification of the dataset for fast walk activity

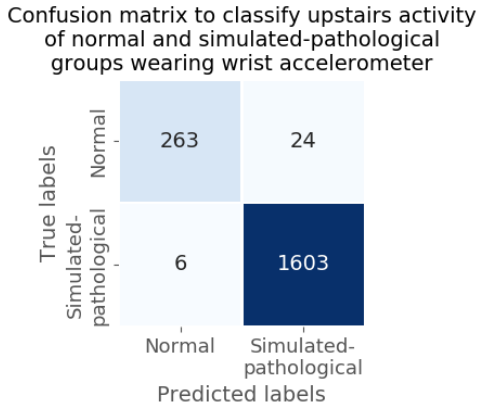


Figure 3.47: Confusion matrix for condition classification of the dataset for ascending stairs activity

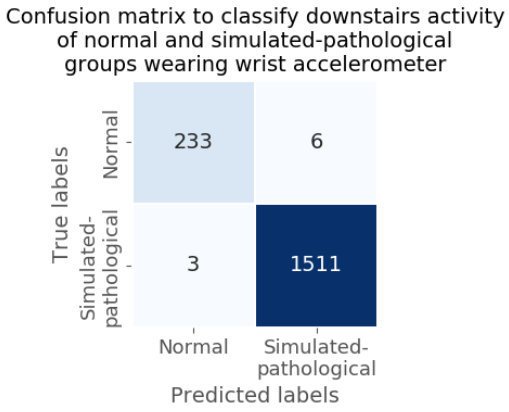


Figure 3.48: Confusion matrix for condition classification of the dataset for descending stairs activity

3.3.2.2 Scenarios 2, 3 and 4: Activity classification

Hyper-parameter tuning Often, the model can suffer from under- or over-fitting and therefore, a CV technique was used to reduce the likelihood this occurring. This was achieved by visualising the accuracy of both training and test datasets. The model is likely to be under-fitted when the accuracy of both training and test sets are both low. On the other hand, the model is likely to be over-fitted when the accuracy of the training set is much larger than the accuracy of the test set. Therefore, the overall performance of each algorithm was taken into consideration in order to identify the optimal hyper-parameters.

Five different ML algorithms were trained to perform activity classification on normal and simulated-pathological conditions. However, hyper-parameter tuning was performed only on four of those algorithms, kNN, NN, RF and SVM because with the GB algorithm there were no parameters to test. This is because the only parameter that the GB algorithm accepts is the number of classes of probabilities. However, if this is set manually, the classes are not adjusted based on the data.

The following eight figures demonstrate the change of accuracy score according to the parameters used for each ML algorithm. Each figure shows the accuracy achieved when using the training set and the test set. Additionally, each combined figure features two graphs, the top graph describes the activity-*type* classification, and the bottom graph describes the activity-*task* classification.

The results show that the accuracy of the training set was always better than the accuracy of the test set. Additionally, activity-type classification showed better outcomes of accuracy than for specific activity-task classification.

k-Nearest Neighbour For the kNN classifier, the number of neighbours and the distance metric were of high importance and therefore, for the hyper-parameter tuning, these two parameters were considered.

Overall, the accuracy achieved by the test set was higher ($> 98\%$ for activity-type and $> 96\%$ for activity-task) when the number of neighbours was low. This suggested that with low number of neighbours, the data was over-fitted. On the other hand, the accuracy of the test and the training sets dropped when the number of neighbours was high. This suggested that the data was under-fitted. Hence, for all four cases (see Figures 3.49 and 3.50), four neighbours and Euclidean distance were chosen as the optimal parameters. The reason for choosing four

neighbours was that the range of CV accuracy of test set was not very close to the CV accuracy of the training set. For example, for the activity type of the normal group, the CV accuracy of the test set when $k = 5$ was close to the accuracies of the test set.

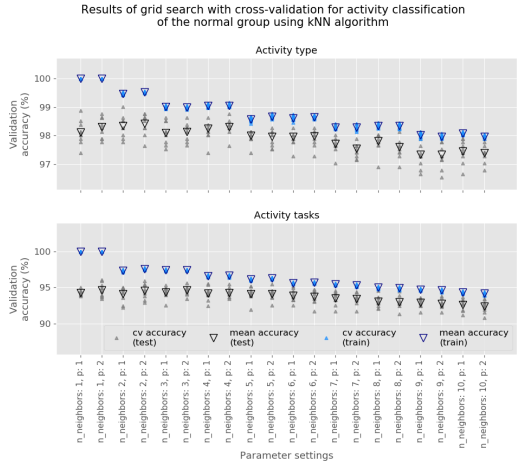


Figure 3.49: Hyperparameter tuning results of k-Nearest Neighbour for activity classification under normal condition using 10 K-fold CV. The parameters chosen were: neighbours=4 and distance=2 for both activity type and task

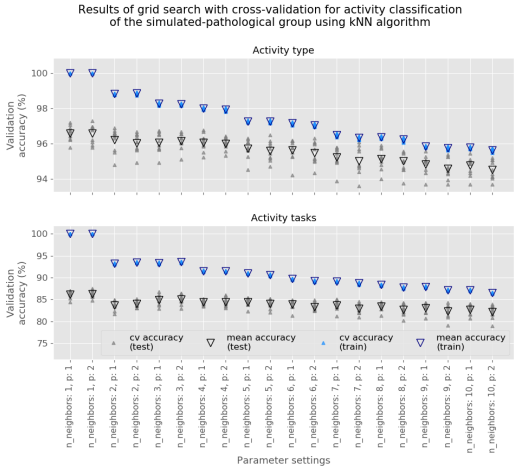


Figure 3.50: Hyperparameter tuning results of k-Nearest Neighbour for activity classification under simulated-pathological condition using 10 K-fold CV The parameters chosen were: neighbours=4 and distance=2 for both activity type and task

Neural Network For the NN classifier, the only parameter tuned was the size of hidden layers. Only one hidden layer was chosen, and the number of neurons selected was the parameter altered. A general trend was demonstrated for all four cases as shown in Figures 3.51 and 3.52. For the training set, the accuracy increased as the number of neurons increased. For the test set, the accuracy remained steady as the number of neurons increased. This suggests that at that point the model started to become overfitted because the accuracy of the test set was not increasing, which means the model had completed using the training set.

Considering the activity types in normal group, 35 neurons were selected as the optimal value because of the range of CV accuracy of the test set and the fact that training and test set accuracies did not overlap. For the activity task, 60 neurons were chosen as the optimal value because the accuracy of the test set was similar to most of the other results, however the range of the CV accuracy of the test set was small. Also the CV accuracy of the test set did not overlap with any CV accuracy of the training set. Regarding the activity types of the simulated-pathological group, 55 neurons were selected. The test set accuracy was steady between 40 and 80 neurons. The accuracy of the training set though was increasing, hence 55 neurons was

chosen as the optimal value. Another reason for this choice was the range of the CV accuracy of the test set with 55 neurons showing the smallest range. For similar reasons, 75 neurons was selected as the optimal parameter for activity task classification in the simulated-pathological group.

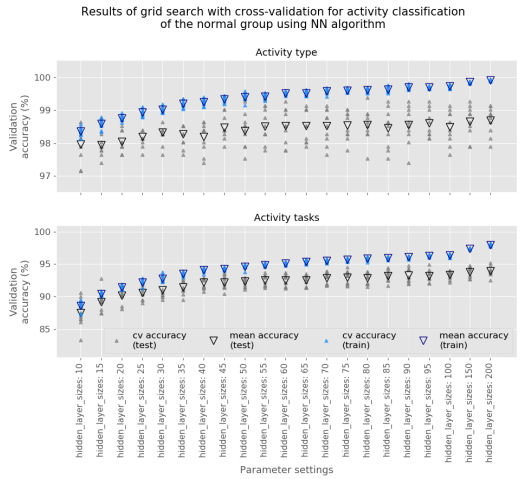


Figure 3.51: Hyperparameter tuning results of Neural Network for activity classification under normal condition using 10 K-fold CV. The parameters chosen were: neurons=35 for activity type and neurons=60 for activity task

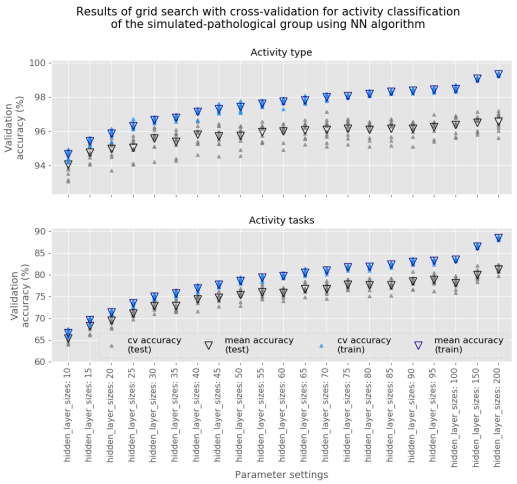


Figure 3.52: Hyperparameter tuning results of Neural Network for activity classification under simulated-pathological condition using 10 K-fold CV. The parameters chosen were: neurons=55 and neurons=75 for activity task

Random forest The number of trees and the value of minimum sample split were two of the most important parameters when tuning the RF classifier. The following Figures 3.53 and 3.54 demonstrate the changes of the accuracy according to the two parameters.

Considering all four cases, there was a general trend towards increasing accuracy with increasing numbers of trees. Considering the minimum sample split parameter, the accuracy slightly dropped when the number of min_sample_split increased.

When the number of trees was five or more, the data started to become over-fitted and so in all cases, four trees and 12 min_sample_split were selected. The test set and training set accuracy was still high with these two parameters as demonstrated in Figures 3.53 and 3.54, but not the highest in order to avoid over-fit.

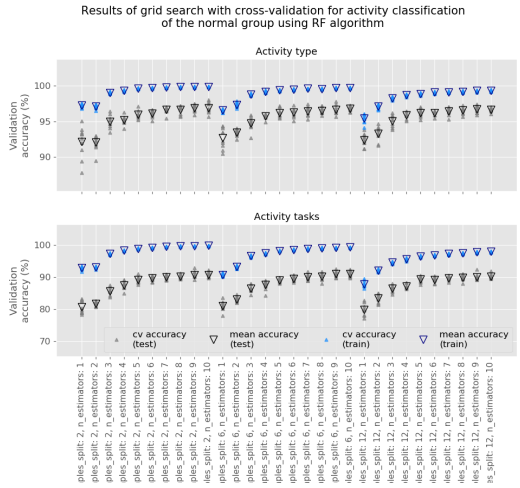


Figure 3.53: Hyperparameter tuning results of Random Forest for activity classification under normal condition using 10 K-fold CV. The parameters chosen were: trees=4 and minimum sample split=12 for both activity type and task

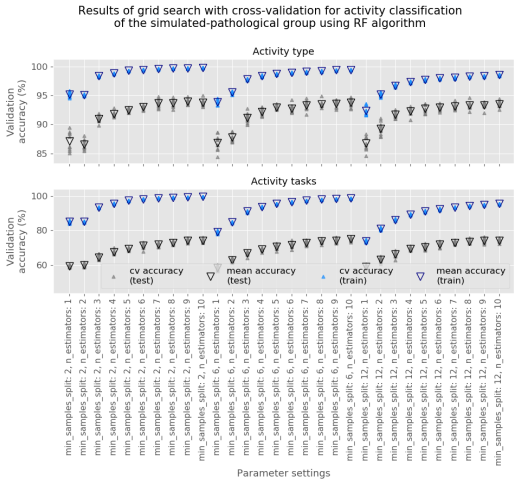


Figure 3.54: Hyperparameter tuning results of Random Forest for activity classification under simulated-pathological condition using 10 K-fold CV. The parameters chosen were: trees=4 and minimum sample split=12 for both activity type and task

Support Vector Machine Gamma and C parameters were used to perform hyper-parameter tuning for the SVM classifier.

The following Figure 3.55 and 3.56 demonstrate the more varied results. Considering the normal group, the optimal parameters chosen were C and gamma equal to one. Most of the results, when C was equal to 0.01 and 0.1, suggested that the data was under-fitted because the accuracy was low ($< 60\%$) in comparison to the other results, and the training and test set accuracies were equal. The chosen parameter once again was selected because both training and test sets showed high accuracies ($> 96\%$), but not the highest of training set. A similar pattern was followed in the simulated-pathological group, however the optimal parameters chosen were C equals to 10 and gamma equals to 1.

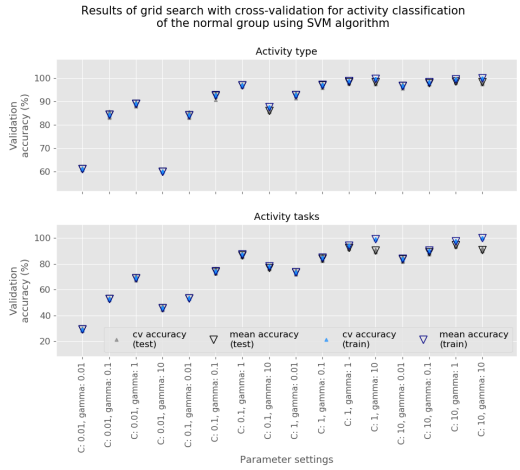


Figure 3.55: Hyperparameter tuning results of Support Vector Machine for activity classification under normal condition using 10 K-fold CV. The parameters chosen were: C=1 and gamma=1 and C=1 and gamma=1 for activity task

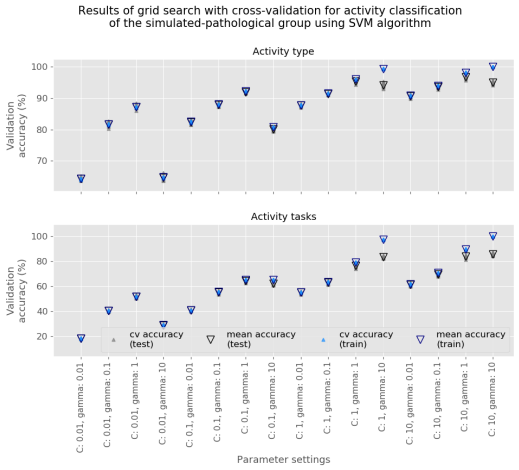


Figure 3.56: Hyperparameter tuning results of Support Vector Machine for activity classification under simulated-pathological condition using 10 K-fold CV. The parameters chosen were: C=10 and gamma=1 for activity type and C=10 and gamma=1 for activity task

Table 3.21 demonstrates all the chosen hyper-parameters for all groups.

Table 3.21: Best parameters for activity classification under normal and simulated-pathological conditions.

	Normal		Simulated-pathological	
	Activity type	Activity task	Activity type	Activity task
kNN (neighbours, distance)	4, 2	4, 2	4, 2	4, 2
NN (neurons)	35	60	55	75
RF (trees, min_sample_split)	4, 12	4, 12	4, 12	4, 12
SVM (C, gamma)	1, 1	1, 1	10, 1	10, 1

Performance metrics The third research question was related to whether it was possible to identify the different types and tasks of physical activity in normal and simulated-pathological conditions. Five Machine Learning algorithms, NN, RF, kNN, GB and SVM were trained to classify the activities. This section demonstrates the results for each activity classification group when the monitor was attached to participant's wrist. The performance metrics of accuracy, F1-score, precision and recall are presented and the confusion matrix of the selected algorithm is also presented in each case.

Scenario 2: Normal training set Vs Normal test set

Activity type Overall, all algorithms, except GB, demonstrated excellent performance. They reached accuracies between 0.953 and 0.984. Additionally, the other three performance metrics also showed high performance, which was above 0.919.

Table 3.22: Results for activity type classification under normal condition using five machine learning algorithms [Mean (95% CI)].

Performance metrics	kNN	NN	RF	SVM	GB
Accuracy (test set)	0.983 (0.982-0.983)	0.983 (0.982-0.983)	0.953 (0.952-0.954)	**0.984** (0.983-0.984)	0.897 (0.896-0.898)
Accuracy (training set)	0.991 (0.991-0.991)	0.992 (0.991-0.992)	0.986 (0.985-0.986)	0.989 (0.989-0.989)	0.898 (0.898-0.898)
F1-score	0.970 (0.969-0.971)	0.971 (0.969-0.973)	0.920 (0.918-0.922)	**0.971** (0.970-0.973)	0.846 (0.844-0.848)
Precision	0.970 (0.969-0.972)	0.969 (0.968-0.970)	0.917 (0.915-0.917)	**0.971** (0.970-0.973)	0.843 (0.840-0.845)
Recall	0.970 (0.969-0.972)	0.970 (0.969-0.972)	0.919 (0.917-0.922)	**0.971** (0.969-0.973)	0.850 (0.848-0.852)

Based on its slightly better performance than kNN and NN, the SVM algorithm was explored further in a confusion matrix. The Figure 3.57 below demonstrates that the algorithm misclassified by predicting static and dynamic activities as transition activity, sit-to-stand, for 23/4845 and 34/2230 times respectively. The greatest misclassification though was when transition activity was predicted as dynamic 54/1018 times. Static activity had the highest recall and precision values among the activity types, achieving scores of 0.994 and 0.997 respectively. The worst performance, among the activity types, was obtained by the transition activity with 0.935 and 0.944 values for recall and precision respectively. Sit-to-stand was the only transition activity as demonstrated in Figure 3.7.

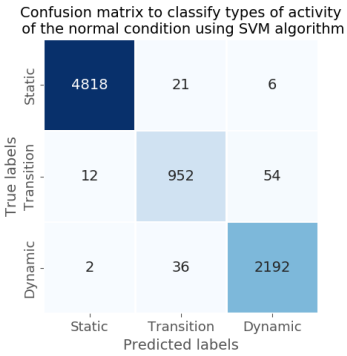


Figure 3.57: Confusion matrix for activity type classification under normal condition using support vector machine

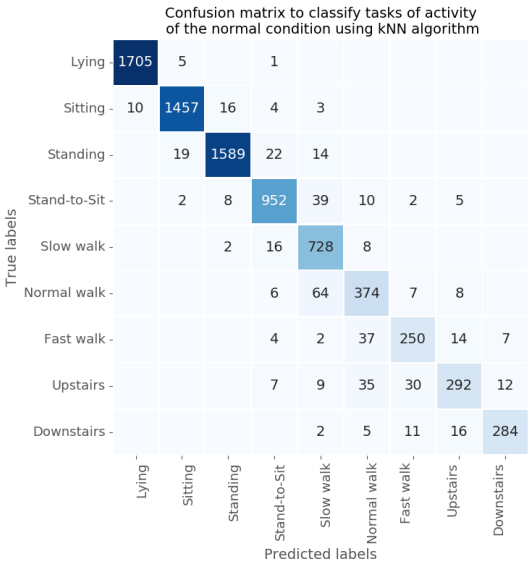


Figure 3.58: Confusion matrix for activity task classification under normal condition using k-Nearest Neighbour

Activity task For classification of specific activity tasks, the overall performance of kNN was excellent in comparison to the other four algorithms. The NN and SVM algorithms had high accuracy scores but the scores for the other performance metrics were poorer. GB showed the worst overall performance for the activity task classification.

Table 3.23: Results for activity task classification under normal condition using five machine learning algorithms [Mean (95% CI)].

Performance metrics	kNN	NN	RF	SVM	GB
Accuracy (test set)	**0.943** (0.941-0.944)	0.926 (0.924-0.927)	0.873 (0.871-0.875)	0.926 (0.925-0.928)	0.749 (0.746-0.751)
Accuracy (training set)	0.967 (0.966-0.967)	0.951 (0.951-0.952)	0.956 (0.956-0.957)	0.943 (0.943-0.944)	0.754 (0.754-0.754)
F1-score	**0.904** (0.902-0.906)	0.871 (0.868-0.874)	0.871 (0.868-0.874)	0.863 (0.860-0.866)	0.651 (0.647-0.655)
Precision	**0.911** (0.909-0.913)	0.871 (0.868-0.874)	0.797 (0.792-0.802)	0.874 (0.871-0.877)	0.675 (0.671-0.679)
Recall	**0.901** (0.899-0.903)	0.861 (0.858-0.864)	0.781 (0.778-0.784)	0.858 (0.855-0.861)	0.655 (0.651-0.658)

Again using the best performing algorithm, this time kNN, a confusion matrix was developed. The classification of specific activity tasks was more challenging than the classification of broader activity types. This was demonstrated with the results of the confusion matrix in Figure 3.58

above. In general, misclassification occured among tasks that were part of the same activity type. For example, regarding dynamic activities, normal walk was predicted 62/459, 7/459 and 9/459 times as slow walk, fast walk and upstairs. Another example considering static activities, sitting was confused as lying and standing 11/1490 and 16/1490 times respectively. Activities such as lying, sitting, standing, stand-to-sit and slow walk achieved individual recall scores above 0.933, with lying achieving the highest recall score of 0.996. The other four activities achieved recall scores between 0.764 and 0.893, with fast walk having the worst performance. In terms of the precision score, lying, sitting, standing, stand-to-sit and stair descent achieved scores greater than 0.940. Normal walk obtained the worst precision score of 0.798.

Scenario 3: Simulated-pathological training set Vs Simulated-pathological test set

Activity type Similar to activity type classification, the SVM algorithm showed the best overall performance of the five algorithms. kNN and NN algorithms had slightly worse performance than the SVM, hence SVM was selected for the misclassification analysis.

Table 3.24: Results for activity type classification under simulated-pathological condition using five machine learning algorithms [Mean (95% CI)].

Performance metrics	kNN	NN	RF	SVM	GB
Accuracy (test set)	0.960 (0.959-0.961)	0.957 (0.956-0.958)	0.921 (0.920-0.923)	**0.967** (0.966-0.968)	0.834 (0.832-0.836)
Accuracy (training set)	0.979 (0.979-0.979)	0.975 (0.974-0.975)	0.972 (0.972-0.972)	0.982 (0.982-0.982)	0.834 (0.834-0.834)
F1-score	0.948 (0.948-0.949)	0.948 (0.947-0.949)	0.892 (0.890-0.894)	**0.958** (0.957-0.959)	0.799 (0.796-0.801)
Precision	0.954 (0.953-0.955)	0.952 (0.951-0.954)	0.904 (0.901-0.906)	**0.964** (0.963-0.965)	0.793 (0.791-0.796)
Recall	0.944 (0.943-0.945)	0.943 (0.941-0.945)	0.889 (0.887-0.890)	**0.952** (0.951-0.954)	0.806 (0.804-0.808)

Scenario 3 yielded more misclassifications than Scenario 2. The greatest number of misclassifications made by the SVM algorithm was due to prediction of transition activity as dynamic activity 204/2746 times as demonstrated in Figure 3.59. This was the greatest misclassification in the activity type classification of the normal group as well. Dynamic activity had the highest individual recall score of 0.984, and transition activity had the lowest recall score of 0.900. In terms of the individual precision scores, the range was between 0.948 and 0.974, achieved by transition and static activities respectively.

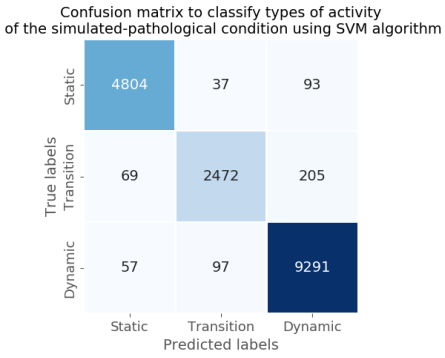


Figure 3.59: Confusion matrix for activity type classification under simulated-pathological condition using support vector machine

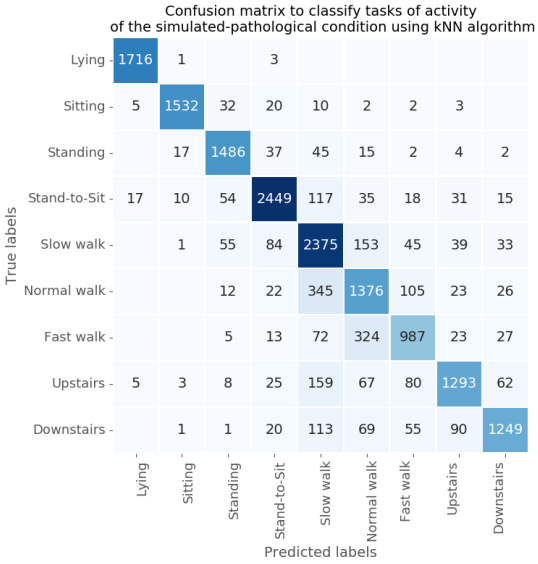


Figure 3.60: Confusion matrix for activity task classification under simulated-pathological condition using k-Nearest Neighbour

Activity task The classification of activity tasks for the simulated-pathological groups showed decreased overall performance in comparison to the results for activity type. kNN and SVM had similar performance, with kNN having slightly better performance metrics.

Table 3.25: Results for activity task classification under simulated-pathological condition using five machine learning algorithms [Mean (95% CI)].

Performance metrics	kNN	NN	RF	SVM	GB
Accuracy (test set)	**0.845** (0.843-0.846)	0.770 (0.767-0.772)	0.689 (0.687-0.691)	0.838 (0.836-0.840)	0.516 (0.514-0.518)
Accuracy (training set)	0.915 (0.915-0.915)	0.816 (0.815-0.817)	0.896 (0.895-0.896)	0.897 (0.896-0.897)	0.519 (0.519-0.520)
F1-score	**0.846** (0.844-0.848)	0.772 (0.770-0.774)	0.687 (0.685-0.689)	0.839 (0.836-0.841)	0.510 (0.508-0.512)
Precision	**0.855** (0.853-0.857)	0.777 (0.776-0.778)	0.692 (0.690-0.694)	0.846 (0.844-0.848)	0.551 (0.548-0.553)
Recall	**0.840** (0.838-0.842)	0.773 (0.771-0.776)	0.684 (0.681-0.686)	0.834 (0.832-0.837)	0.518 (0.516-0.520)

There were many misclassifications among the different activity tasks, however the larger amount of misclassifications was occured among the tasks that were part of the same activity type. Overall, static activities, such as lying, sitting and standing, had the three highest recall scores respectively as shown in Figure 3.60. They achieved recall scores greater than

0.924. Stand-to-sit activity achieved a recall score of 0.891, which was highest than any of the scores of any of the dynamic activities. Fast walk obtained the least recall score with a value of 0.680. Considering individual precision scores, lying had the greatest score once again, which was 0.985. Sitting and stand-to-sit activities achieved high scores as well, obtaining values of 0.980 and 0.916 respectively. Standing, upstairs and downstairs activities achieved individual prediction scores between 0.860 and 0.899. The worst performance in terms of prediction was obtained by normal walk activity which was 0.675.

Scenario 4: Normal training set Vs Simulated-pathological test set

For this particular scenario, there was no need to use CV because the training and test sets were completely different. Additionally, this suggests that there was minimal risk for overfitting the data, since the model was trained and tested with two different datasets. In this investigation the SVM and kNN algorithms were trained with normal dataset and tested with the simulated-pathological dataset. The reason for testing only the SVM algorithm and kNN algorithm for activity-type and activity-task respectively was because they had been previously shown to consistently outperform the other approaches. Since these algorithms had the best performance when the training and test data were from the same group, it was assumed that they will continue have the best performance in comparison to the other algorithms because now the data would be from a completely different group. Additionally, in this scenario the features calculated were reduced to principal components. This was done in the same way with both scenarios 2 and 3, where the 120 features are calculated, scaled and then used as an input in the principal component analysis function. This step was done prior the usage of a classifier.

Activity type The SVM algorithm had shown excellent overall performance for the activity type classification when the algorithm was both trained with normal dataset and tested with a normal dataset. The performance metrics obtained were: 0.984, 0.971, 0.971, and 0.971 for accuracy, F1-score, precision and recall respectively. The performance dropped when the normal data set was used to train the algorithm for use with the simulated-pathological data as demonstrated in Table 3.26.

Table 3.26: Results for activity type classification using normal data as training set and simulated-pathological data as test set.

Performance metrics	SVM
Accuracy (test set)	0.528
F1-score	0.535
Precision	0.638
Recall	0.528

Even though the overall performance of this algorithm was poor, some of the individual recall and precision scores were reasonably high. This is demonstrated in Figure 3.61. For example, the individual recall score of the static activity was 0.969 and the individual precision score of the dynamic activity was 0.793.

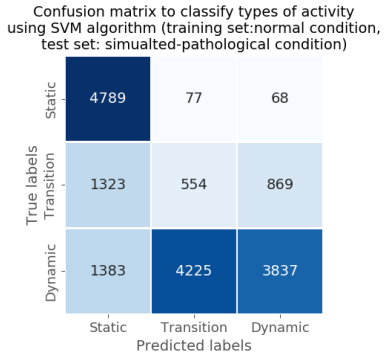


Figure 3.61: Confusion matrix for activity type classification using normal data as training set and simulated-pathological data as test set

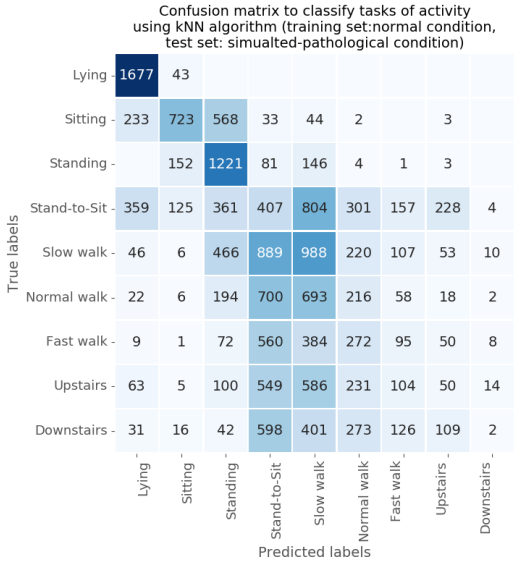


Figure 3.62: Confusion matrix for activity task classification using normal data as training set and simulated-pathological data as test set

Activity task For the classification of activity tasks, kNN algorithm demonstrates the best performance when the algorithm was trained with normal dataset. Hence, the same algorithm was used for this case. When the normal data set was used to train the algorithm for use with the simulated-pathological data however, the kNN algorithm had the worst performance among all the algorithms tested in this chapter. Table 3.27 demonstrates the performance of the kNN algorithm.

Table 3.27: Results for activity task classification using normal data as training set and simulated-pathological data as test set.

Performance metrics	kNN
Accuracy (test set)	0.313
F1-score	0.274
Precision	0.270
Recall	0.313

The confusion matrix in Figure 3.62 showed that many of the activity tasks were misclassified by the algorithm. In the majority of the cases, the numbers of false predictions were even larger than the number of correct predictions. For example, normal walk activity was predicted as stand-to-sit and slow walk activities 714/1909 and 686/1909 times respectively. Normal activity was correctly predicted as its true class only 215/1909 times. Almost all dynamic activities and stand-to-sit activity showed similar pattern for their results. In terms of the individual precision and recall scores, the lying activity showed the best results, which were 0.975 and 0.687 respectively. The second highest recall score was achieved by standing activity, and the second highest precision score was achieved by sitting activity. All the other activities yielded recall and precision scores between 0.001 and 0.441. The stairs descent activity achieved the worst scores of all activities.

3.4 Discussion

This chapter presents the analysis of data from both wrist and ankle worn activity monitors in volunteers in both their usual normal and simulated pathological states. The results presented here answer the first three research questions posed in section 2.5.1; 1) Can we identify whether locating a wrist mounted activity monitor is suitable for condition classification and activity recognition? 2) Can we automatically identify if a patient is moving normally? and 3) Can we automatically identify different types of physical activity in healthy participants under normal and simulated-pathological conditions? These questions provided an answer for two classification cases; whether the wrist-worn device was able to differentiate the two different conditions, and whether successful activity classification was performed individually in each case.

Earlier studies have tested either both locations or at least one of the two. In general, studies in the literature have used multiple accelerometers attached at several different body locations.

Although this might be beneficial for the researcher, attaching accelerometers in real-time on multiple locations on the user is inefficient and can be inconvenient for him or her. Most studies that attached accelerometers on both wrist and ankle, demonstrated that devices attached to the ankle achieved better outcomes of accuracy than those worn on the wrist (Bao and Intille 2004; Mannini et al. 2013; Gjoreski et al. 2016). One study showed that wrist mounting achieved better results than the ankle (Sasaki et al. 2016) although this might be because the activities tested were more upper limb focussed.

Our data demonstrated slightly better results for the wrist location for the binary condition classification when compared to the ankle location. However, our data demonstrated slightly better results for the ankle location for the multi-class classification than the results from the wrist. The differences were not large (max difference: 1.8%) and both were considered acceptable and, based on patient preferences elicited during our stakeholder engagement work, the wrist location was selected for the remaining analysis in this thesis.

Previous researchers have developed classifiers to differentiate healthy population from patient population with pathological gaits. ML algorithms such as NN, SVM and kNN were used for the differentiation. For example, Mannini and colleagues differentiated the gaits of people with Huntington's disease, post-stroke condition and healthy elderly (Mannini et al. 2016). As expected the results suggested that each gait can be classified with high accuracy since they produce different acceleration signals.

The results presented in this chapter are consistent with previously published studies, in that the algorithms successfully differentiated between the normal and pathological gaits. There are however, a number of differences from previous studies. One difference was the use of more than one type of sensors, for example some studies used both accelerometer and gyroscope sensors on the participants while performing the activities. In this chapter, only accelerometer sensors were used because the MOX device used contained only accelerometer sensor. Another difference between the work described in this chapter and other studies was that the pathological gait was simulated by healthy participants. This was not the case for many of the literature studies (Del Rosario et al. 2014; Capela et al. 2015a; Lonini et al. 2016), however there were a few recent studies that also recruited healthy volunteers to simulate pathological gaits (Cola et al. 2015; Esfahani and Nussbaum 2019). The approach of collecting simulated pathological gait provides the opportunity to prepare and make any improvements in the experimental protocol ahead of

collecting data from actual patients. While real-world patient data is the gold standard it si not always possible to obtain, for example during a pandemic. Furthermore patients who are usually by definition, frailer and in more pain, carry an extra ethical imperative to ensure that any data collected from patients is useful and there may be circumstances in which synthetic data carries less ethical burden. It is also faster to collect data and modify collection processes in this way as it removes the need to deal with the complexity of clinical settings and patient-related governance.

In terms of differentiating the conditions, other measures such as ground reaction force, pressure and planar motion, as well as motion analysis were used (Alaqtash et al. 2011; Pogorelc et al. 2012; Zeng et al. 2016; Esfahani and Nussbaum 2019). These measures are all lab-based and they are not representative of real-life data.

For the purpose of this thesis, a machine learning, SVM algorithm was used to perform the binary classification between normal and simulated-pathological conditions. The classifier was able to differentiate the gait patterns between the two conditions successfully. Detection of all activities displayed high levels of accuracy, achieving greater than 96.8%. The results of the other performance metrics, F1-score, precision and recall, showed some minor differences between the activities. Recall had slightly lower values than precision values (1.2% max difference), meaning that there were more false-negatives than false-positives. False-negative means that an activity, for instance slow walk, was predicted as a different activity like normal walk. This was observed in all the individual activities and dynamic set of the wrist location. The SVM algorithm is a popular algorithm used in literature to differentiate multi-class problems because of its properties that enables to differentiate data in more than one plane. The reason for that is because of its kernel property that turns the model into non-linear and enables the classification of non-linear problems. In comparison to results from literature, our algorithm showed excellent performance. The literature results ranged from 80.96% (Chowdhury et al. 2018) to 96% (Sasaki et al. 2016).

In general, the results suggested that the acceleration signal had differences when the physical activities were performed under the two conditions. The reason for that was because the movement was different when someone was performing an activity under normal and pathological conditions. For instance, people with Rheumatoid Arthritis (RA) have limited joint motion, therefore their arm swings in lower range in comparison to healthy people (Weiss et al. 2008).

Additionally, people with RA might have reduced movement in the lower limbs due to pain (Weiss et al. 2008). Regarding the work occurring at specific joints, it was suggested that the most noticeable difference between healthy people and RA patients is the reduced work at the ankle (Weiss et al. 2008). This might be due to the reduced internal plantar-flexor moments. This can be associated with reduced walking speed in the plantar-flexor muscle group. Since the arms swing out of phase with the legs (Whittle 2007), when the leg movement becomes slower the arms swing becomes slower as well with a reduced range of motion. Additionally, lower intensity activities are not measured easily hence they are not detected. This might be because of their arrhythmic and intermittent nature (Calabró et al. 2014). RA patients move slower because of the pain and stiffness associated with the condition resulting in lower accelerations.

Earlier studies have assessed the physical activity classification of healthy population. However, few studies have assessed explicitly whether algorithms trained on data from healthy populations were suitable for pathological populations. Those that have made this comparison conclude, like us, that large differences between groups means that algorithms will perform better when trained for specific target groups (Del Rosario et al. 2014; Capela et al. 2015a; Lonini et al. 2016).

Several studies performed activity classification in healthy populations using ML classifiers and a small number have used a similar pattern of activity task classification as was performed in this chapter. Activity type describes the activities that were classified in general groups, and activity task describes the activities that were classified specifically. Machine learning classifiers, such as ***SVM*** (Mannini and Sabatini 2010; Zhang et al. 2012; Mannini et al. 2013; Cleland et al. 2013; Sasaki et al. 2016; Saez et al. 2016; Gjoreski et al. 2016), ***RF*** (Sasaki et al. 2016; John et al. 2013; Sasaki et al. 2016; Saez et al. 2016; Gjoreski et al. 2016; Lee and Kwan 2018; Twomey et al. 2018), ***kNN*** (Bao and Intille 2004; Mannini and Sabatini 2010; Saez et al. 2016; Gjoreski et al. 2016), ***NN*** (Mannini and Sabatini 2010; Zhang et al. 2012; Cleland et al. 2013; Twomey et al. 2018) and ***GB*** (Bao and Intille 2004; Mannini and Sabatini 2010; Atallah et al. 2011; Zhang et al. 2012; Cleland et al. 2013; Saez et al. 2016; Gjoreski et al. 2016) have been used for physical activity classification. All of these studies tested the algorithms on a healthy population only, and hence high levels of accuracy were achieved (>90%). On the other hand, some other studies have developed algorithms using healthy data and tested whether they could be applied accurately to pathological populations. Most studies using this approach

showed that algorithms trained with data from the healthy population cannot be applied to pathological populations (Del Rosario et al. 2014; Capela et al. 2015a; Lonini et al. 2016). These studies identified a need for population-specific models, or even patient-specific models to classify physical activities in pathological populations. ML approach has the potential to develop models specific to the data that is given to train the algorithm. Therefore, in general, there is a need to collect large amounts of data from a specific population and develop the appropriate models.

One aspect of the pilot study was to act as a baseline for developing activity classifiers for the healthy population. These classifiers will be further updated to suit the pathological population with walking impairments. Before refining them for the pathological population, classifiers need to be validated. Overall this chapter's findings for the healthy population were in accordance with findings reported by (Montoye et al. 2016; Abdull Sukor et al. 2018). The main difference between the conclusions of this chapter and the studies in the state of the art section, is the application to classification of the simulated-pathological gait. The methods used were similar to those described in the state of the art.

Within this chapter, several algorithms have been developed for each group. There were four scenarios for how those algorithms were used, as demonstrated in Table 3.3. For scenario 2 and 3, 10-fold CV was used, hence the mean of the 10-folds was presented for the performance evaluation measures. This was because the training and test sets were taken from the same dataset, while for scenario 4, the training and test sets were completely different as the former was the normal dataset and the latter was the simulated dataset. It is important to mention that the training set always showed higher accuracy than the test set because the model better fitted to the data that it has trained with rather than to unseen data. Additionally, activity type classification had better performance than activity task classification because the classes of activity types were more distinct than the classes of activity tasks.

In all cases the normal group achieved higher levels of accuracy than the simulated-pathological group. The difference in mean accuracy is likely due to the fact that volunteers were asked to make significant changes to their motions under simulated-pathological conditions. Although we attempted to train participants to replicate compromised motion, we could not be certain that their movements accurately reflected real pathological motion. Indeed, participants may have interpreted the instructions on how to mimic the pathological activities slightly differently.

This means that the accuracies reported can only be considered a reasonable initial estimate of the performance of ML algorithms on real patients. Even though this is the case, the signal might not completely represent the nuances of specific pathological gait patterns. In the real world each patient might have slightly different movement even within a disease group and there may be systematic differences between disease types. The data used in this chapter did however share the fundamental characteristics of low velocities and low accelerations. Clinicians did not record or review the signals because of the exploratory nature of the study.

Additionally, the algorithms predicting the broader activity types achieved better performance metrics compared to the algorithms predicting specific activity tasks. This might be because the behaviour of the accelerometer data did not differ excessively between a few individual activity tasks. For example, the acceleration signals for normal and fast walk shared similarities which makes it more difficult to classify activity tasks accurately. This was demonstrated in the confusion matrices, where there were more false-negatives and false-positives when classifying activity tasks than activity types. One of the reasons for that is because of the small amount of data collected. When larger amount of data is collected from more participants, the signals from each activity will be more distinct. This is because more data will enable better differentiation among similar activities.

Another potential limitation in this study is the potential for human error in labelling the activities. Even though there was a gold standard video, the activity labelling was completed manually and is potentially subject to human error. To minimise this risk, thorough steps were taken such as using slow-motion analysis, replaying analysis and triple counting each walk.

This study has several strengths. First, 30 healthy volunteers have been recruited to collect data, where these volunteers performed the activities under two conditions. This could equate to the collection of data from 60 volunteers overall. In comparison to datasets available online, the number of participants recruited for this study were much greater. For example, only WISDM and UCI datasets recruited 36 and 30 participants respectively. The other datasets, such as PAMAP2, MHEALTH, FoS, Opportunity and JSI recruited participants between four and ten participants. Second, this study has examined two potential locations for the activity monitors. The locations were used to answer two important questions; 1) whether the algorithm at the specific location can differentiate if the volunteer is performing the activities under normal condition or simulated-pathological condition. This is important because in the proposed

system for this thesis, the activity monitor will accomodate algorithms for both healthy and pathological populations. Therefore it is quite important to differentiate successfully the two conditions. And 2) whether the algorithm at the specific location can differentiate successfully the activities performed by the participants in each condition. Again, this is very important since the clinicians would like to get a correct perspective about the physical activity levels of their patients. The results suggested that the wrist location can be used to answer both questions with accuracy greater than 84.5%. The wrist location was the preferred choice for a group of people with RA, as shown in appendix A. Therefore, this study demonstrated that wrist location can be used for pathological populations since it is their preferred choice and it provides excellent results.

3.5 Summary

In this study, five machine learning algorithms were used to classify nine ADLs. Activities were performed by healthy volunteers in both normal and simulated-pathological conditions. The volunteers had activity monitors attached on their wrist and on their ankle. Two sets of classification were performed, condition and activity. The former was about the differentiation between the normal and the simulated-pathological conditions. The latter was about the classification of the different activities performed. The activities were classified into two groups, general activity type and specific activity task. This chapter answered the first three research questions posed in section 2.5.1. In terms of the first question, the results suggested that the wrist and ankle locations provided similar outcomes, therefore only wrist was used for further analysis. This is justified because not only did the wrist provide adequately good results, but it was also the only acceptable choice for the patients (see appendix A). Regarding the second question, an SVM classifier was used to identify whether the volunteer was performing the activities under normal or simulated-pathological conditions. Successfully, the algorithm was able to do that with excellent performance. The third question was related to the activity classification. The majority of the algorithms performed well when the training and test sets both came from the same population. Conversely, when the algorithms were trained with normal data and tested with simulated-pathological data, as would usually occur in the real-world with current consumer devices, the accuracy demonstrated was poor. When the ML algorithm was trained with a simulated-pathological dataset and subsequently tested using a simulated-pathological

dataset the accuracy improved to levels more comparable to the normal against normal conditions. It may therefore be possible to develop more accurate and clinically useful activity classification algorithms based on the accelerometer gait signals of individual people or specific sub-populations such as disease groups. Therefore, these algorithms could be further used by clinicians to evaluate the daily activity performance of chronic condition patients with walking impairments.

Chapter 4

Step count testing using algorithms from the literature

4.1 Introduction

In the previous chapter, activity-type and activity-task classifications were performed for the normal and simulated-pathological groups. *SVM* and *kNN* classifiers achieved the top performance for activity-type and activity-task classifications respectively. The *SVM* algorithm yielded accuracy scores of 98.4% and 96.7% for the normal and simulated-pathological groups respectively. The *kNN* algorithm achieved 94.3% and 84.5% accuracy scores for the normal and simulated-pathological groups respectively.

Step count is important since it is one of the useful available measures that indicates, overall, how physically active a person is (Sylvia et al. 2014). After reviewing several studies from the literature, four step count algorithms were implemented and compared in this chapter. These four algorithms were chosen because each algorithm used a different approach for step recognition, and also because they were the most descriptive to implement according to the information provided in the original paper. These algorithms were categorised into four groups based on the main methods that they used for step recognition. The categories were: (1) peak detection, (2) thresholding (frequency-domain), (3) thresholding (time-domain) and (4) template matching. One algorithm from each category was implemented and tested with the data previously collected in the pilot study. It is important to note that some algorithms tested use a combination of these methods.

Figure 4.3 below demonstrates in detail the structure of the two analyses performed in this chapter. The two analyses followed the same process to obtain the results, however they used different parts of the acceleration signal. Analysis A used all acceleration signals that had been manually labelled (true) by the researcher according to the video. This means that this analysis was done without any activity classification. Analysis B used the acceleration signal that had been automatically labelled (predicted) using the ML algorithms. For the activity-type classification an SVM algorithm was used, where the signal was classified as either static, transition, or dynamic activities. For the activity-task classification a kNN algorithm was used, where the individual tasks were identified (e.g. slow walk, normal walk). The reason for performing both analyses was to identify sources of potential errors. For example, if a greater error was calculated when using predicted activity labels, it would likely mean that the activity recognition step was partly responsible for that. On the other hand, if a greater error was calculated using true activity labels, it would likely mean that the step count algorithm had counted false positives and false negatives that matched with the original results. Lastly, if the results were similar for both true and predicted activity labels, it would mean that due to high accuracy of activity classification, the reconstructed signal was very similar to the original signal.

As shown in Figure 4.3, three different cases were tested. For case 1, all nine activities were included. For cases 2 and 3, the acceleration signals that did not corresponded to walking activities were filtered out using an activity type, and activity task approach described above. This process aimed to reduce the number of false positive steps. For the case 2, as walking is a dynamic activity, any signal corresponding to static or transition activities was *excluded*. Finally, for case 3, the individual dynamic tasks were identified (e.g. slow walk, normal walk), and any signals that corresponded to a list of pre-defined dynamic tasks were *included*. Activity-types and activity-tasks can be seen in (Figure 3.7). Figures 4.1 and 4.2 demonstrate the input acceleration signal of step count analysis A and step count analysis B.

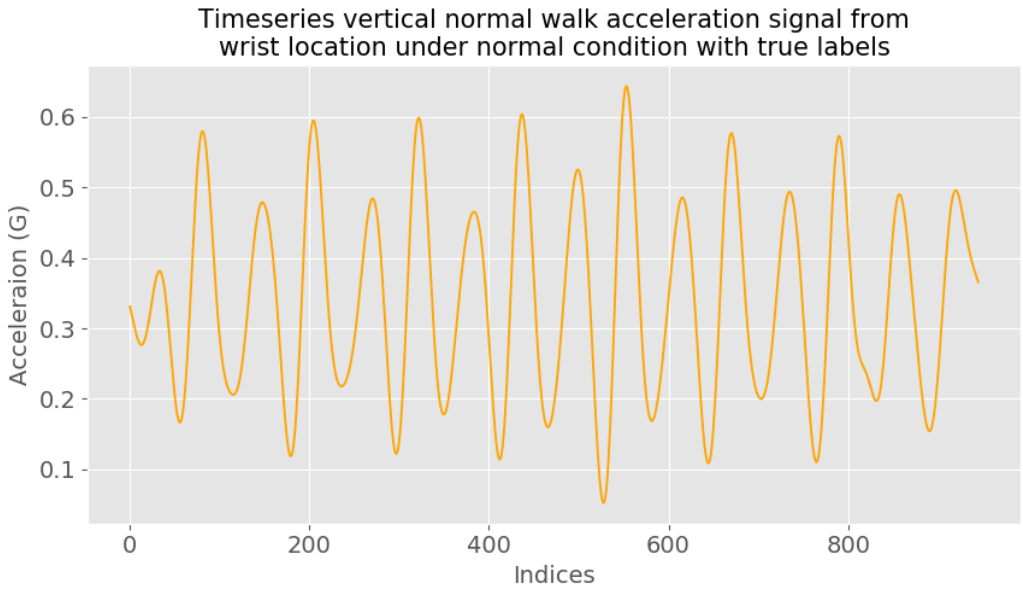


Figure 4.1: Normal walking acceleration signal used as input in the four step count algorithms under normal condition. The signal is based on true activity labels.

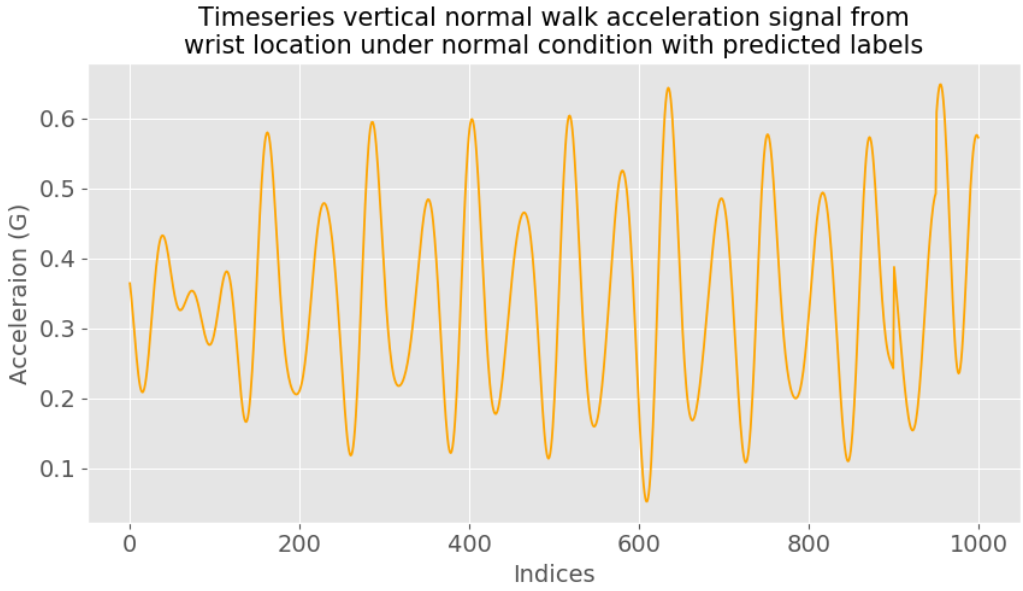


Figure 4.2: Normal walking acceleration signal used as input in the four step count algorithms under normal condition. The signal is based on predicted activity labels using machine learning algorithm.

The reason why these three approaches were tested was to see how the activity recognition step affects the results. As already shown in the results of the previous chapter, activity-type recognition produced better classification results than activity-task recognition. Therefore, this chapter also aimed to check how general and specific activity recognition might influence the results of the step count. Additionally, with the use of activity classification prior step count,

the need of having thresholds to identify each individual dynamic activity was avoided.

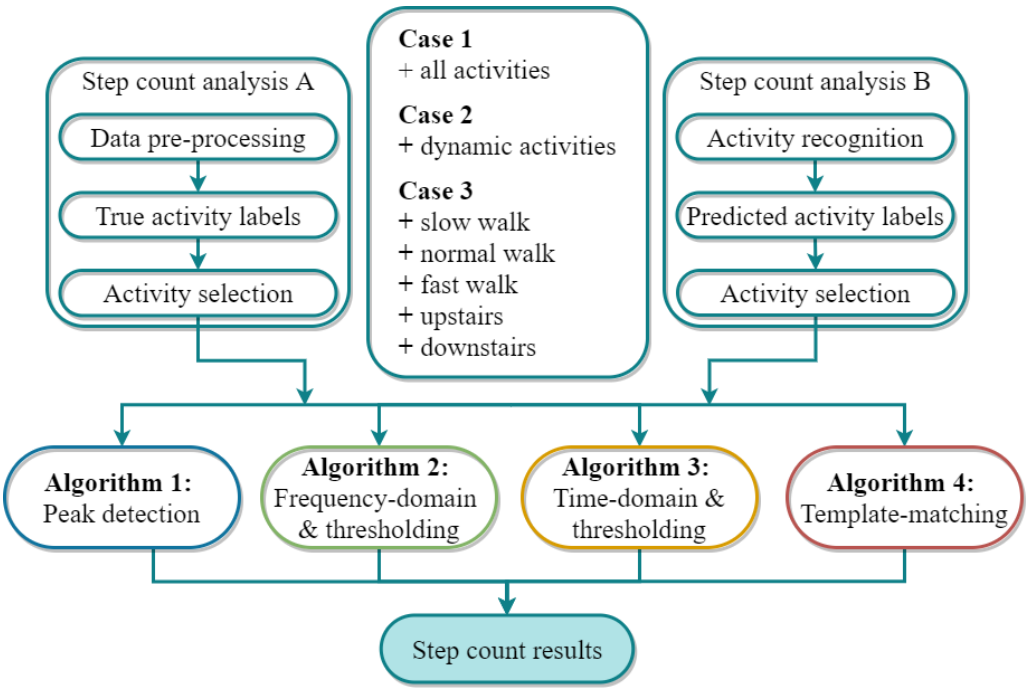


Figure 4.3: Diagram showing step count analysis using algorithms from literature

4.2 Methodology

As described in the literature review section 2.3.3, a peak in the acceleration signal is created when the foot touches the ground (Bui et al. 2018; Moe-Nilssen and Helbostad 2004; Pham et al. 2017; Sejdic et al. 2016), and walking is a repetitive process (Ao et al. 2018).

Four algorithms, corresponding to the four main categories of step-count algorithm, were selected from the literature. Each algorithm was re-implemented as faithfully as possible. For each of the participants in the pilot dataset the number of steps was estimated, and then compared to the actual number (according to video).

4.2.1 Algorithm 1: Peak detection

Peak detection is one of the most common algorithms (see Figure 4.4) that is used to identify peaks in a signal. In this instance, the peaks identified can be interpreted as the number of steps taken by the participant.

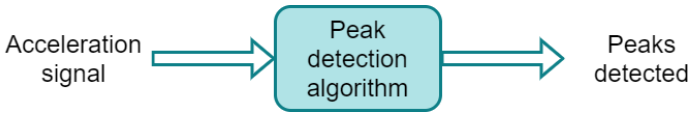


Figure 4.4: Input and output of peak detection algorithm

In the Palshikar study a peak detection algorithm was designed based on a peak function (S) that calculates the overall average difference in amplitude from a selected point (x_i) and its k preceding succedding neighbours (Palshikar 2009). This concept is demonstrated in Figure 4.5 for k=3.

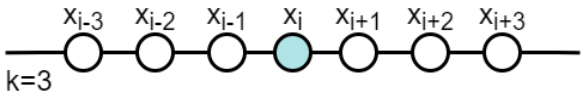


Figure 4.5: Right and left neighbours of a selected peak (x_i)

The peak function (S) is based on the moving average concept.

$$S(k,i,x,T) = \frac{\frac{x_i - x_{i-1} + x_i - x_{i-2} + \cdots + x_i - x_{i-k}}{k} + \frac{x_i - x_{i+1} + x_i - x_{i+2} + \cdots + x_i - x_{i+k}}{k}}{2} \tag{4.1}$$

where T is a univariate uniformly sampled time series; x_i is a given i^{th} acceleration point in T; k is the number of neighbours.

The peak function (S) can be used to find local peaks in T since it produces positive values for local peaks.

$$P = \{S(x_i) > 0\} \tag{4.2}$$

P contains all the local peaks identified using the S function. The next step was to find the significant peaks of the signal. This was done by removing local peaks which were "small" in global context.

$$P' = \{(S(x_i) > 0)\ \&\ (S(x_i) > (m + h \times std))\} \tag{4.3}$$

where h defines the significance of a peak detection ($1 < h \leq 3$); m is the mean of all positive values; std is the standard deviation of all positive values.

P' contains all the significant peaks. The final step was to retain only one peak within distance

k inside P'.

$$if\ |j-i| \leq k, remove\ the\ smaller\ value\ of \{x_i, x_j\} \tag{4.4}$$

4.2.2 Algorithm 2: Time-domain and thresholding

Thresholding algorithms use a threshold to identify the number of steps. The following algorithm developed by Thanh and colleagues used one threshold (Thanh et al. 2017). The first step of the algorithm was to identify peaks using a peak detection algorithm with minimum (index) distance of 11 discrete points between the peaks. The original authors used MATLAB function "findpeaks", however for this study "scipy.signal.find_peaks" python function was used. These two functions shared the same goal, and they used similar parameters to achieve this. The only possible difference might be the underlying algorithm used by the two functions, however they both use the same conceptual approach of comparing neighbouring values to locate local maxima. The threshold was set as $0.1g$, which is equal to 0.981 m/s^2 resting gravity acceleration. This threshold was used to ensure that the participant is moving and not resting in a static state. Another value was set, which was identified experimentally. The value was $0.01g$ and it was calculated based on the experimental testing performed by (Thanh et al. 2017). This value was used to avoid the effects of vibration in gravity acceleration in steady state. Figure 4.6 demonstrates the signal with the peaks identified and the threshold used for the algorithm. Additionally, Figure 4.7 demonstrates a flowchart describing the algorithm.

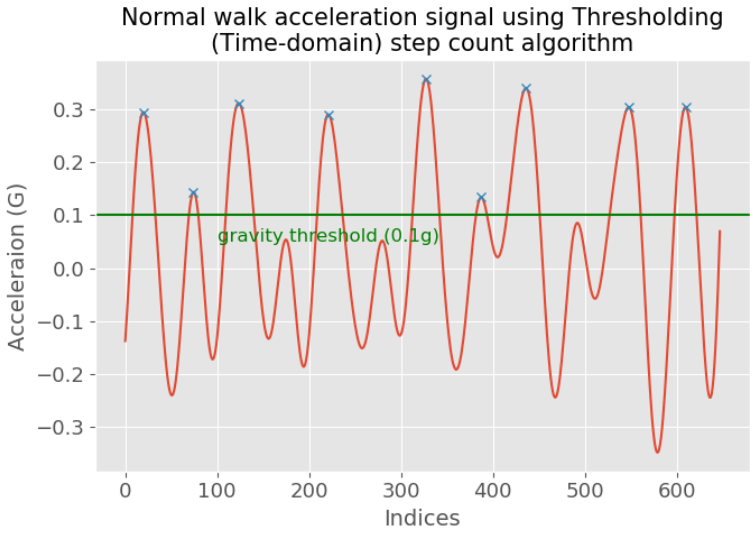


Figure 4.6: Acceleration normal walk signal using Thresholding (T-domain) algorithm

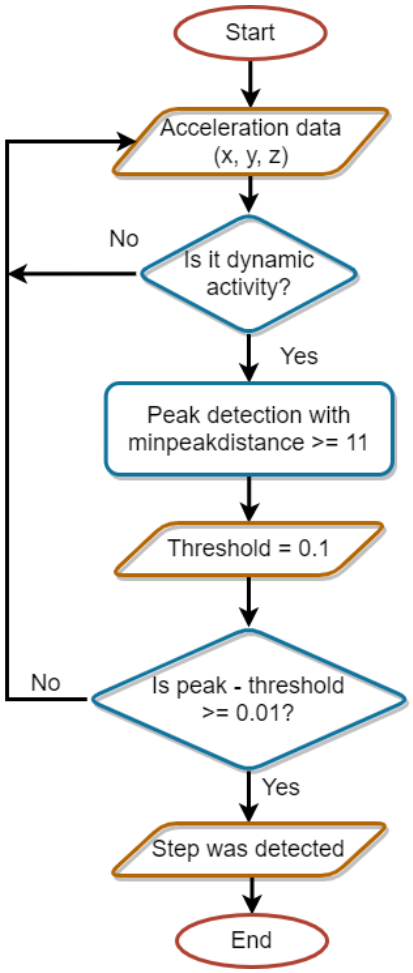


Figure 4.7: Flowchart of thresholding (time-domain) algorithm

4.2.3 Algorithm 3: Frequency-domain and thresholding

This algorithm used the frequency-domain instead of the time-domain. Generally, the Fourier transform can be used to decompose a signal into its constituent sinusoidal functions, as it is assumed that all signals are the sum of simple sinusoids (cosine and sine). In digital signal processing, signals are discrete-time signals for which both time and amplitude have discrete values. Acceleration signals are discrete, and thus fast Fourier Transform (FFT) algorithm was applied, which computed the discrete Fourier Transform (DFT) of a signal.

Additionally, the frequency of the acceleration signal changed over time, because the signal represented nine different activities. Each activity constituted from different frequencies. By applying the FFT over the entire data, the results do not reveal true transitions in the frequency domain. To overcome this problem, the acceleration data was split into several segments, called windows. The idea behind this method is that short windows can be considered stationary, and

thus reveal the spectral content of a single activity.

Sliding window technique was used to segment the data in the frequency domain. However, this method might produce some issues, for example spectral leakage. To reduce the spectral leakage, a Hamming window was used rather than a square window. The Hamming window has a sinusoidal shape and it is often used to cancel the nearest side lobe.

The step count algorithm was performed in the frequency-domain, using multiple thresholds which were defined from the authors (Dirican and Aksoy 2017). To use the FFT effectively, the signal is required to be periodic and in this case, the dynamic activities performed by the participants were indeed appropriately periodic. This produced the frequency content of the activities in each window. Each window should include enough data to show the periodicity of the acceleration signal. By including this conversion, the number of steps can be determined, since windows with no periodicity have no distinct peak. Figure 4.8 demonstrates the FFT transforms (real and imaginary parts) along with the thresholds used. As demonstrated in the flowchart (Figure 4.9), the thresholds of the imaginary parts were updated if the statement set was true. Since the FFT signal was separated into windows, each participant yielded multiple windows. Each window was used to calculate the number of steps performed by the participant in two seconds, which was done by counting the peaks in the area of interest (window).

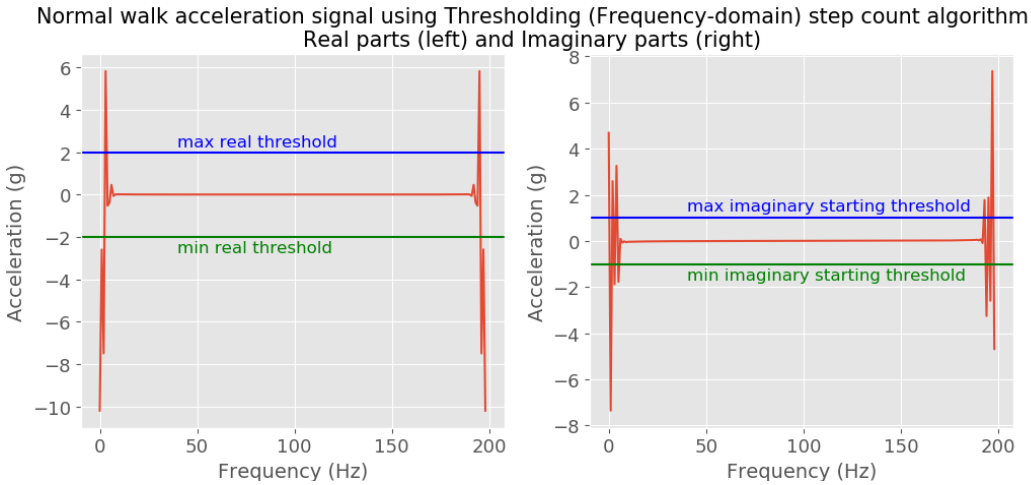


Figure 4.8: FFT transforms of the acceleration normal walk signal using Thresholding (F-domain) algorithm

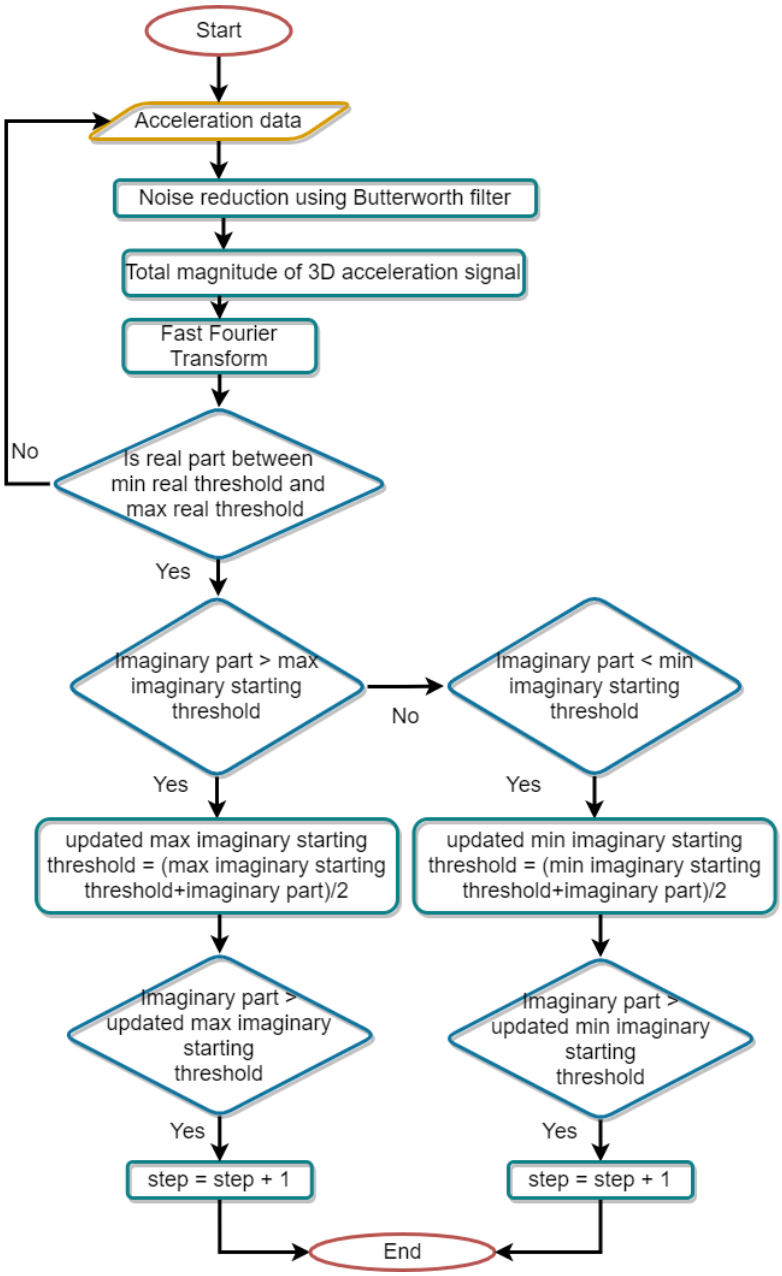


Figure 4.9: Flowchart of thresholding (frequency-domain) algorithm

4.2.4 Algorithm 4: Template-matching

This algorithm used a very different approach, ignoring any peaks, in comparison to the other algorithms implemented above. The concept behind this algorithm is based on the template-matching approach using correlation and dynamic time warping barycenter averaging (DBA) technique.

For algorithm 4, a template was developed for each participant based on his/her walking acceleration signal. Each template represented a step during walking activity. The aim was to

develop a measure that detects the template that best matches the reference acceleration signal. (Micó-Amigo et al. 2016) followed several steps as shown in Figure 4.10.

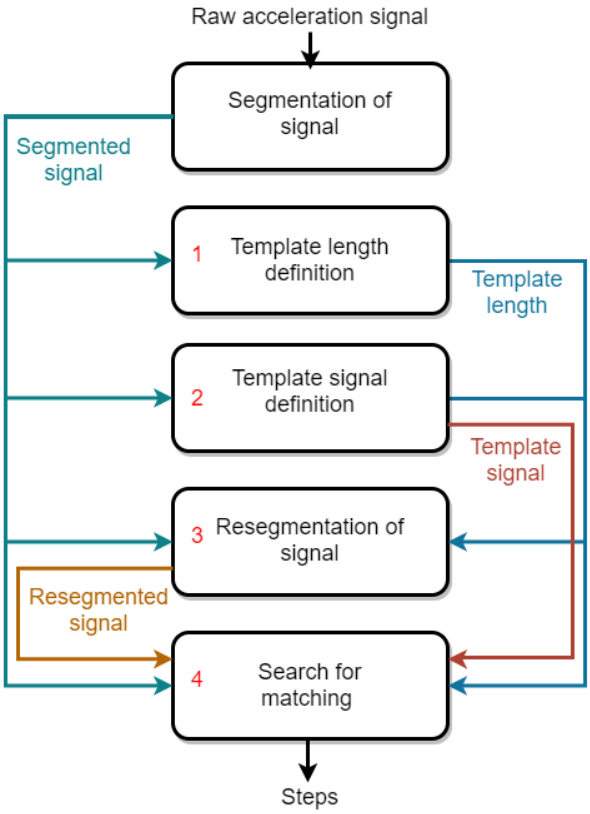


Figure 4.10: Flowchart of template-matching algorithm

4.2.4.1 Dynamic time warping (DTW)

DTW is an algorithm that is used to measure similarity between two sequences (signals) by identifying their optimal alignment. This can be done by identifying flexible similarities in time dimension (x-axis) by aligning the elements inside both sequences. This means that a non-linear alignment is produced, which allows similar shapes to match even though the sequences might be out of phase. Throughout this section let $Q =< q1, \ldots, qT >$ and $P =< p1, \ldots, pT >$, and let δ and D be a distance between elements of the sequences (Petitjean et al. 2011).

An n-by-m grid is formed by arranging the two sequences as shown in Figure 4.11. Each point on the grid is associated to an alignment between elements of the two sequences. The elements are aligned by a warping path in order to minimise the distance between them as demonstrated in Figure 4.12 (Bemdt and Clifford 1994). Several warping paths are defined, but once the best one is identified, a similarity score, which describes the fit between the two sequences, is calculated. The similarity score quantifies the fit degree by compressing or stretching the

sequences with respect to time.

The cost of the optimal alignment is computed recursively by:

$$D(Q_i, P_j) = \delta(q_i, p_j) + \min \begin{cases} D(Q_{i-1}, P_{j-1}) \\ D(Q_i, P_{j-1}) \\ D(Q_{i-1}, P_j) \end{cases} \tag{4.5}$$

The overall similarity score is calculated by:

$$D(Q_{|Q|}, P_{|P|}) = D(Q_T, P_T) \tag{4.6}$$

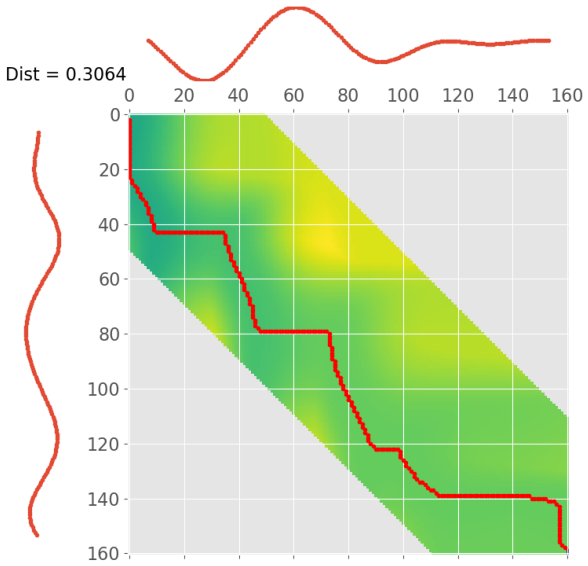


Figure 4.11: Warping path

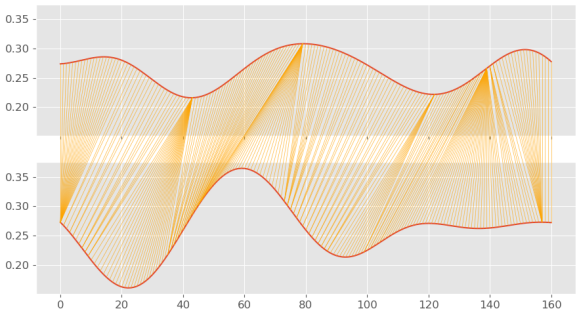


Figure 4.12: Dynamic time warping

4.2.4.2 DTW-barycentre averaging (DBA)

The DBA technique is a global averaging method developed by (Petitjean et al. 2011). It is an iterative algorithm where DTW is used to align the sequences to be averaged with an evolving average. Throughout this section let $S = S_1, \ldots, S_N$ be the set of N sequences from which we would like to compute an average sequence A, where $A =< a_1, \ldots, a_T >$. Additionally, let $A' =< a'_1, \ldots, a'_T >$ be the update of A since the average is updated.

The initial average sequence A is chosen randomly from the S dataset, which contains N sequences. After that, the DBA method becomes deterministic. In order to calculate the final

average sequence, the same process is repeated. For this process the DTW similarity is calculated between the temporary average sequence and each individual sequence S_i, and the path is saved. Using the saved paths, a new average A' is constructed. Each element of the A' sequence is updated as the barycentre of the elements associated to it during the previous step.

$$A'_t = barycentre(assoc(A_t)) \ where \ barycentre\{X_1, \ldots, X_\alpha\} = \frac{X_1 + \cdots + X_\alpha}{\alpha} \tag{4.7}$$

Where α is the total number of sequences used.

The assoc function can link each element of the average sequence to one or more elements of the sequences of S. The benefits of DBA method are the preservation of:

1. the shape of the sequences
2. the magnitude of the peaks/troughs on the y-axis
3. the timing of those peaks/troughs on the x-axis

Figure 4.13 demonstrates several sequences and their average, which was calculated using DBA method.

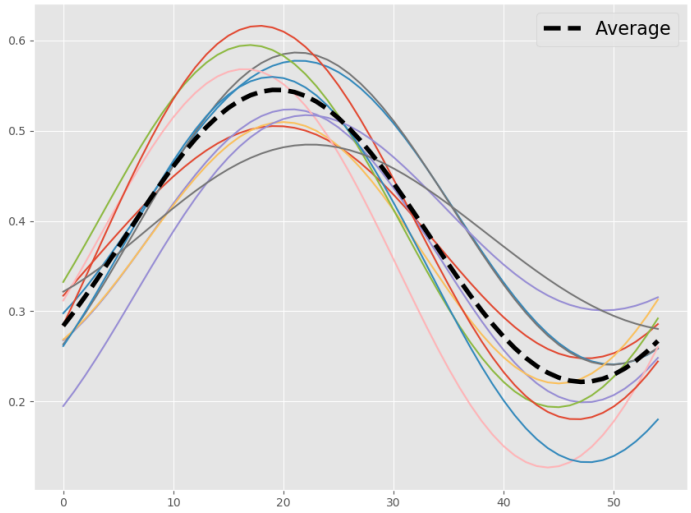


Figure 4.13: Averaging dynamic time warping

The first stage of this algorithm was to select the walking activities from the entire signal. Since activity classification has been performed, the activities of interest, and therefore sections of

accelerometer signal, were already selected. Again, the dynamic activities were selected, which included level-walking at slow, normal and fast speeds, as well as stair ascent and descent. This signal was defined as the *segmented signal*.

For the second stage, an autocorrelation function was used to calculate the *template length*. (Moe-Nilssen and Helbostad 2004) suggested that a cyclic signal produces autocorrelation coefficients with peak values that are equal to the signal's periodicity.

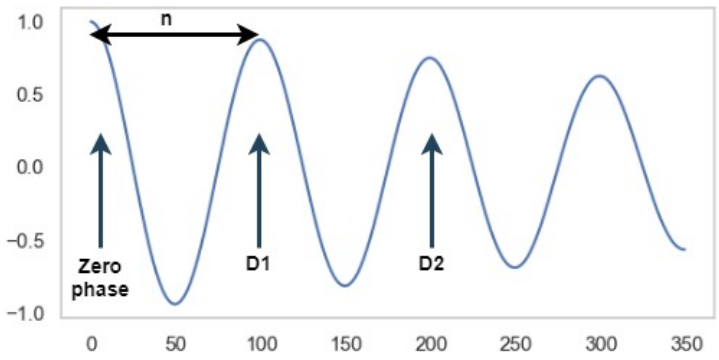


Figure 4.14: Unbiased autocorrelation plot of normal walk

Figure 4.14 demonstrates the autocorrelation coefficient sequence from the acceleration signal during walking. There are two dominant peaks, $D1$ and $D2$, which are equivalent to one step and one stride respectively. Additionally, n represents the samples per step, and in our case the length of the template. This is because the template represents one step. The signal was firstly normalised and since it is an autocorrelation function, the same signal is used for both normalisations. The reason for using two normalisations was to compress the signal in the y-axis, and the autocorrelation signal to be between -1 and 1.

$$normalisation_1 = \frac{signal - mean(signal)}{std(signal) \times length(signal)} \tag{4.8}$$

$$normalisation_2 = \frac{signal - mean(signal)}{std(signal)} \tag{4.9}$$

Then, the function "numpy.correlate" from the numpy library in python was used to calculate the autocorrelation coefficients. The autocorrelation coefficient sequence was produced similarly to Figure 4.14 above. Using the function "scipy.signal.find_peaks" from the scipy library in python, the first peak ($D1$) was identified. Based on that finding, n was found as well, which

corresponded to the *template length*.

The third step in the process was to develop the *template signal*. The original authors separated this step into further sub-steps. The first of these sub-steps was to identify a low limit and a high limit based on the *segmented signal* and the *template length*.

$$\textit{low limit} = 115\%\ \textit{template length samples (from the start of the segmented signal)} \tag{4.10}$$

$$\textit{high limit} = 115\%\ \textit{template length samples (before the end of the segmented signal)} \tag{4.11}$$

The "scipy.signal.find_peaks" function was then used to find peaks in the segmented signal between the low and high limits. The peaks needed to be at least 40% template length samples apart from each other. The next step was to define a start and an end section around the identified peaks.

$$\textit{start section} = 5\%\ \textit{template length samples (before the peak)} \tag{4.12}$$

$$\textit{end section} = 100\%\ \textit{template length samples (after the peak)} \tag{4.13}$$

These new sections were then used to develop a new average signal using dynamic time warping. In particular, "dtw_barycenter_averaging" function was used to average all the signals (Petitjean et al. 2011). This function is explained in detail in section 4.2.4.2. The basic idea is that the average of several signals was calculated using dynamic time warping.

The fourth step performed was associated with developing the *resegmented signal*. The *resegmented signal* was an extension of the *segmented signal* using the raw acceleration signal and the *template length*. The *segmented signal* was extended by template length samples and twice template length samples to the left and right respectively. This was done to ensure that no information was lost.

The final step was to match the *template signal* with the *resegmented signal* in order to detect the steps performed by the participants. To do this, the *resegmented signal* was separated into different windows of similar length to the *template length*. For example, <*resegmented*

$signal_{w1}, \ldots,$ *resegmented signal*$_{wn}$ >. Then, a signal representing the standard deviation of the difference between each resegmented window and the *template signal* was calculated and called *SD difference signal*.

$$SD\ difference\ signal = std(template\_signal\ -\ resegmented_{wi}) \tag{4.14}$$

Additionally, a signal with the correlation coefficients between each resegmented window and the template signal was also calculated.

$$tmp\ cor\ coef\ signal = corcoef(template\_signal,\ resegmented_{wi}) \tag{4.15}$$

This signal was multiplied with the ratio of ranges of the template signal and the resegmented signal to calculate *correlation signal*. The *coefficient signal* was calculated as the ratio between the normalised *correlation signal* and the normalised *SD difference signal*. The last step, counting the number of steps, was to find the peaks at the *coefficient signal* that are at least 60% template length samples apart from each other.

4.2.5 Data analysis

Root mean square error (RMSE) was calculated to measure the difference between the predicted and the true number of steps (recorded from video). RMSE is the standard deviation of the prediction errors, otherwise known as the residual. The residual measures how far the data points are from the regression line. A RMSE closer to zero indicates less error. This measure it is used because it can point out large errors, which in our case is important to identify such errors and avoid them. However, this means that it might be more sensitive to the presence of false data. Additionally, another downside of this method is that it cannot be used to identify either missed steps or over-counted steps. Therefore, in chapter 5 another measures are used to identify whether the algorithm missed or over-counted any steps.

4.3 Results

The following results firstly indicate whether algorithms already published in the literature work effectively. Secondly, it allows us to explore how the algorithms work, so that the most useful

aspects might be identified to help develop new, better performing, algorithms.

Additionally, we aimed to explore how the results of the step count algorithms were affected when activity classification algorithms were applied as a preprocessing step. In other words, the true activity labels were used to check the results of the step count algorithms without applying any ML activity classification algorithms. As the next step, the predicted activity labels were used to check the results of the step count algorithms after applying the ML activity classification algorithm.

In terms of the activities used, three cases were examined. The first case used the acceleration signal that includes all nine collected activities, the second case used the acceleration signal that includes all the dynamic activities, and the third case used the acceleration signal of each dynamic activity individually.

4.3.1 Normal group

4.3.1.1 True activity labels

Table 4.1: Results of step count algorithms for dynamic and individual activities under normal condition using true activity labels.

Activities	Algorithms			
	Peak detection	Thresholding (F-domain)	Thresholding (T-domain)	Template-matching
Case 1: All	**9.92** (9.45,10.36)	45.51 (44.59,46.41)	30.98 (30.23,31.70)	226.36 (222.21,230.44)
Case 2: Dynamic	30.31 (29.91,30.72)	**8.58** (8.15,8.99)	33.93 (33.44,34.41)	20.02 (18.08,21.78)
Case 3: Slow walk	**2.81** (2.7,2.92)	6.61 (6.34,6.86)	9.10 (8.76,9.43)	8.99 (8.34,9.59)
Case 3: Normal walk	4.03 (3.91,4.16)	4.09 (3.93,4.23)	6.75 (6.67,6.84)	**1.35** (1.28,1.42)
Case 3: Fast walk	5.07 (4.98,5.16)	5.85 (5.74,5.96)	5.82 (5.76,5.88)	**1.74** (1.63,1.84)
Case 3: Stair ascent	4.55 (4.41,4.69)	5.37 (5.21,5.52)	6.37 (6.31,6.42)	**1.90** (1.76,2.03)
Case 3: Stair descent	5.16 (4.98,5.32)	6.36 (6.23,6.48)	6.81 (6.73,6.89)	**1.11** (1.08,1.15)

Table 4.1 shows that the template-matching algorithm achieved superior results in comparison to the other three algorithms for normal walk, fast walk, upstairs and downstairs activities. It is worth mentioning that the template-matching algorithm had such a large error in case 1,

since the template created was not representative. This is because in case 1 the signal used includes all nine activities, each activity might has slightly different template. This means that the template created will not be representative for the nine activities, therefore this leads to the large error. Thresholding (frequency-domain) results were substantially better than the other algorithms for dynamic activities. Peak detection results were better than the other three algorithms for slow walk activity. The thresholding (time-domain) approach yielded the greatest error in almost all activities, except fast walk.

4.3.1.2 Predicted activity labels

Table 4.2: Results of step count algorithms for dynamic and individual activities under normal condition using predicted activity labels.

Activities	Algorithms			
	Peak detection	Thresholding (F-domain)	Thresholding (T-domain)	Template-matching
Case 1: All	**10.22** (9.69,10.72)	42.53 (41.60,43.44)	31.60 (30.87,32.32)	227.41 (221.43,233.24)
Case 2: Dynamic	31.99 (31.53,32.45)	**8.61** (8.18,9.02)	35.06 (34.54,35.57)	20.25 (17.67,22.54)
Case 3: Slow walk	**2.68** (2.58,2.79)	10.65 (10.15,11.12)	8.22 (7.89,8.54)	19.81 (18.19,21.31)
Case 3: Normal walk	4.23 (4.06,4.39)	3.95 (3.73,4.15)	6.05 (5.80,5.93)	**2.67** (2.33,2.56)
Case 3: Fast walk	5.50 (5.40,5.59)	5.71 (5.59,5.84)	5.87 (5.76,5.88)	**2.45** (1.63,1.84)
Case 3: Stair ascent	6.17 (6.01,6.32)	6.84 (6.64,7.03)	7.47 (7.36,7.58)	**3.63** (3.50,3.75)
Case 3: Stair descent	6.42 (6.25,6.58)	7.98 (7.83,8.13)	7.89 (7.80,7.97)	**3.19** (3.00,3.36)

The step count algorithms after activity classification followed a similar pattern of results to those obtained using true activity labels as shown in Table 4.2. However, the majority of the algorithms performed worse in comparison to the true activity label results.

4.3.2 Simulated-pathological group

4.3.2.1 True activity labels

Table 4.3: Results of step count algorithms for dynamic and individual activities under simulated-pathological condition using true activity labels.

Activities	Algorithms			
	Peak detection	Thresholding (F-domain)	Thresholding (T-domain)	Template-matching
Case 1: All	**50.13** (46.54,53.48)	202.37 (197.85,206.80)	102.39 (99.54,105.17)	486.62 (460.04,511.83)
Case 2: Dynamic	**42.93** (41.13,44.65)	135.91 (131.84,139.85)	90.72 (87.39,93.94)	296.83 (276.75,315.65)
Case 3: Slow walk	**20.80** (18.41,22.94)	27.68 (26.55,28.78)	32.87 (31.7,34.01)	84.07 (76.36,91.14)
Case 3: Normal walk	**15.94** (13.76,17.87)	23.13 (21.98,24.22)	25.94 (24.95,26.89)	50.87 (47.03,54.45)
Case 3: Fast walk	**12.64** (11.79,13.44)	15.81 (14.74,16.8)	15.81 (14.74,16.8)	43.02 (38.56,47.06)
Case 3: Stair ascent	10.01 (9.43,10.55)	38.06 (37.03,39.07)	**8.75** (8.44,9.06)	42.12 (39.18,44.88)
Case 3: Stair descent	**9.07** (8.27,9.80)	34.63 (33.62,35.61)	10.46 (9.99,10.91)	59.22 (56.15,62.13)

For the simulated-pathological group, the RMSE results differed from the results obtained in the normal group. Table 4.3 shows that the peak detection algorithm achieved better results compared to the other algorithms for all the activities except upstairs walking. The upstairs walking activity achieved the best outcome when using the thresholding (time-domain) algorithm. The template-matching algorithm showed the poorest performance in all activities, which is in contrast with the results from the normal group. This might result due to the shape of the acceleration signal. The signal collected for the simulated-pathological group was less periodic and with greater noise in comparison to the signal of the normal group. These parameters might have prevented the creation of a representative template that would enable the identification of a correct step for all the volunteers thus achieving very poor results.

4.3.2.2 Predicted activity labels

Table 4.4: Results of step count algorithms for dynamic and individual activities under simulated-pathological condition using predicted activity labels.

Activities	Algorithms			
	Peak detection	Thresholding (F-domain)	Thresholding (T-domain)	Template-matching
Case 1: All	**50.31** (46.65,53.73)	198.96 (197.45,203.36)	102.09 (99.24,104.86)	500.13 (473.90,525.06)
Case 2: Dynamic	**42.19** (40.45,43.86)	142.46 (138.46,146.35)	91.23 (87.93,94.41)	322.89 (298.15,345.86)
Case 3: Slow walk	**23.72** (21.75,25.55)	46.84 (45.06,48.55)	32.33 (31.17,33.44)	97.64 (92.64,102.39)
Case 3: Normal walk	**14.15** (12.61,15.54)	26.28 (24.76,27.71)	25.77 (24.8,26.71)	59.77 (56.43,62.93)
Case 3: Fast walk	**9.57** (9.17,9.95)	16.82 (15.84,17.74)	22.36 (21.52,23.17)	24.79 (22.89,26.56)
Case 3: Stair ascent	**7.60** (7.13,8.03)	33.77 (32.92,34.6)	8.20 (7.94,8.45)	43.87 (40.87,46.67)
Case 3: Stair descent	**6.24** (5.67,6.77)	31.27 (30.39,32.12)	10.41 (10.0,10.8)	47.57 (45.57,49.48)

In this instance, the peak detection algorithm outperformed all the other algorithms for all the activities. Similarly to the normal group, the majority of the algorithms performed worse in comparison to the true activity label results.

4.4 Discussion

This chapter presents the step count analysis of data collected from wrist-worn accelerometers in volunteers performing activities under normal and simulated-pathological conditions using four existing literature algorithms. The results presented here answer the fourth research question posed in section 2.5.1; Can we accurately measure step count in healthy participants under normal and simulated-pathological gaits?

Earlier studies have attempted step count using several algorithms including those used here. One of the main problems identified in the literature has been that wearables provide less accurate results at slower walking speeds. This has consistently been reported for different wearable locations as well as different devices. Commercial and research-based devices attached on the wrist, hip, waist and ankle, improve in accuracy at faster speeds (Cho et al. 2016; Chow et al. 2017; Feng et al. 2017; Klassen et al. 2016; Motl et al. 2011; Stansfield et al. 2015). The

results of this thesis showed similar results to the literature. However, there are few studies that examine how comparator step count algorithms work with different walking speeds and stairs. This thesis showed that for each activity creating person-specific and activity-specific algorithms achieves better results especially in simulated pathological situations. Additionally, the same behaviour has been observed for IMUs attached on the wrist and ankle, and smartphones placed inside trouser pocket, jacket pocket, bags or held in the user's hand(s) (Mikov et al. 2013; Pham et al. 2018; Rhudy and Mahoney 2018). A similar pattern of results was obtained here. The results of most of the step count algorithms in both groups, normal and simulated-pathological, showed worse performance when volunteers walked at a slower speed. This was observed in all algorithms for the simulated-pathological group, and all of the algorithms for slow walking except peak detection in the normal group.

Earlier studies have also shown that the step count accuracy differed in different clinical groups. For example, studies comparing the same step count algorithms in a healthy population and in patient population (neurological, orthopaedic, multiple sclerosis, impaired mobility), have demonstrated that some algorithms had greater accuracy for the healthy group in comparison to the patient group (Marschollek et al. 2008; Motl et al. 2011; Oudre et al. 2018). Overall, these findings are in accordance with findings reported in this chapter in which less error was observed in the normal group in comparison to the simulated-pathological group. As mentioned previously in chapter 3, the main difference between the conclusions of this chapter and to the studies in the state of the art, is the use of the simulated-pathological gait.

Previous studies have also shown that different algorithms result in different outcomes when using different datasets and activities. Various authors developed their own algorithms for step count and compared their results with other state of the art algorithms (Ao et al. 2018; Cho et al. 2016; Godfrey et al. 2016; Gu et al. 2017; Marschollek et al. 2008; Rodríguez et al. 2018). Other studies also developed their own algorithm and compared their results with the results from wearables and/or smartphones (Cho et al. 2016; Gu et al. 2017; Rodríguez et al. 2018). A similar conclusion was reached here, where the four algorithms demonstrated different results, especially in the simulated-pathological group.

Based on our results, all step count algorithms demonstrated high levels of RMSE (8.58 - 500.13) for case 1 (all activities) and case 2 (dynamic activities) in both normal and simulated-pathological groups. The amplitude and time period of the acceleration data differ among the

different tasks. Therefore, the thresholds set in the algorithms for the dynamic activity were not representative for all the individual tasks, which in turn led to the large RMSE in both groups. Additionally, since case 1 included all the performed activities, more false positives were counted, hence there was a large RMSE in both groups. Due to this increased RMSE, case 1 and case 2 will be excluded from further comparisons. For any future studies that it might be required to use the algorithms in cases 1 and 2, it is suggested to use a window technique in order to identify smaller periods of activities that might include just one activity in each window. Depending on the content of each window, the appropriate thresholds of the step count algorithm could be used.

Regarding the normal group, the template-matching algorithm yielded the smallest RMSE, followed by peak detection, thresholding (frequency-domain), and lastly thresholding (time-domain) (as demonstrated in Table 4.1. The template-matching algorithm used autocorrelation function to identify step length. The function produces a cyclic signal with autocorrelation coefficients that are peaks equal to the periodicity of the signal. The signal in that case represents a walking activity, which produces a cyclic signal. Hence, the results from the autocorrelation function suggest that the horizontal length up to the first dominant peak represent one step (Moe-Nilssen and Helbostad 2004). This is demonstrated in Figure 4.14. Periods of maximum match between a patient-specific averaged template and corresponding acceleration signal were searched to count the number of steps. Since the acceleration signal for walking was periodic, the algorithm produced results with a high level of accuracy.

Upper limb movement occurs naturally during gait. The lower limb swing and upper arm swing appear to move alternatively (Cola et al. 2016; Koo and Lee 2016). The accelerometer was attached on the wrist and while the arm was swinging in a repetitive process, a peak was created, therefore the peak detection algorithm identified most of the walking peaks correctly, and had the smallest (2.81) RMSE for slow walking activity.

The thresholding (frequency-domain) algorithm had similar results compared to the peak detection algorithm. This algorithm used the acceleration signal in the frequency-domain rather than the time-domain. A possible reason for the higher RMSE created for this algorithm might be the thresholds that were used. Some of those thresholds were based on the dataset used to develop this algorithm and hence they might not work well for our dataset. The reason for that is because their threshold was obtained based on a few participants, which might not result in

representative thresholds for the general healthy population. In general, for all four tested step count algorithms for all the activities, the thresholds used were similar to the thresholds used in the original paper so as to re-implement the same algorithm described in the original paper. However, in some cases the thresholds from the original papers would not work for the data being analysed in this thesis, and in these cases the algorithms were adapted in order to work with the collected data. Additionally, the algorithm was created for normal walking, however this chapter studied different walking speeds, as well as stair climbing.

The worst algorithm in all activities was the thresholding (time-domain). Similarly to the thresholding (frequency-domain) algorithm, the peak detection parameters and thresholds used might not be representative of the dataset used in this study and therefore a larger error occurred.

Exploring the performance of those algorithms using the data from the simulated-pathological group, it is important to note again that the recruited participants attempted to replicate the compromised motion, hence we cannot be certain that their movements accurately reflect the movements in real-world pathological gaits. They were instructed to generally move slower in comparison to the normal condition and this also meant that they had reduced arm swing. Since the accelerometer was attached on the wrist, this affected the amplitude, time period and periodicity of the signal. All the algorithms produced very poor results under these conditions. Peak detection produced the best results for all activities, except stairs ascent. Thresholding (time-domain) produced the least error for stairs ascent activity. Contrary to the results from the normal group, the worst performance for the simulated-pathological group was obtained from the template-matching algorithm. This might be because the template generated for each participant of the simulated-pathological group was not representative enough. Hence, it might be beneficial to build algorithms to generate better walking step templates.

In all four algorithms, except thresholding (frequency-domain), stair related activities showed the least error for the simulated-pathological group. The reason for this may be due to the systematic movement associated with walking up and down stairs. Additionally, stair climbing might produce more distinct peaks than level-walking, hence better step detection.

The four tested algorithms were not suitable for a group of people with walking impairments and reduced arm swing. Additionally, the acceleration signal was not periodic in each participant, although this might be because participants were simulating the pathology and their simulation

may have varied across each step.

Based on the outcomes from the normal group, the majority of the state of the art algorithms might be suitable for the healthy population, however there is a need for step count algorithms that perform well in patient populations with impaired walking. The following chapter will investigate an approach that provides better results for counting the number of steps in both healthy and patient populations.

The results of the step count algorithms, when true activity labels and predicted activity labels were used, followed a similar pattern for almost all algorithmic approaches. As expected, the results from the predicted labels showed slightly worse performance than the results from the true labels, even though the activity classification results achieved similar accuracies.

This study had a few limitations. First, the tested algorithms were developed based on the description given in the literature and while they were built to the best of our knowledge, there may have been minor variations not described in the underpinning papers. Additionally, some were built using different software so the functions used might slightly differ for each software package, therefore affecting the results.

There are many different algorithms that can be used to calculate step count, however in this study only four were used and there is a possibility that if different algorithms were used, the results might be better or worse. Even though four algorithms were used, these algorithms covered most of the key step count algorithmic categories.

The approach utilised suffered from the limitation that the data collected was in a laboratory setting which maximised the amount of steps that each participant was able to undertake in a straight line. A similar study should be replicated in free-living settings while conducting other activities of daily living.

The results presented in this study were calculated based on four existing literature algorithms. The majority of the algorithms showed good results for the normal group, however the results from the simulated-pathological group were not as good. This creates an opportunity for the development of a better step count algorithm that could provide exceptional results for both normal and simulated-pathological groups.

4.5 Summary

In this study, we used four state-of-the-art step count algorithms to measure the number of steps undertaken by the participants. Walking activities were performed under normal and simulated-pathological conditions. The results of this chapter are used to answer the fourth research question posed in section 2.5.1. The majority of the algorithms performed well for the normal condition, probably due to the periodicity of the acceleration signal. On the other hand, all the algorithms performed poorly in counting specific steps for the simulated-pathological condition. Therefore, this confirmed the need to develop more accurate and clinically useful step count algorithms for the population with impaired gait. Additionally, since the results from the step count algorithms were affected from the activity classification results, it is essential to develop population-specific or patient-specific activity classification algorithms.

Chapter 5

Development and validation of a new step count algorithm

5.1 Introduction

In chapter four, four step count algorithms from literature were implemented and tested using the acceleration data from the pilot study. The main outcome from the chapter was that step count algorithms that might work well in a healthy population might not be suitable for a pathological population. Hence, there is a need to develop population-specific algorithms that will result in better outcomes for their associated population.

This chapter describes the development of a new step count algorithm based on template-matching using DTW. The algorithm was inspired by the step count algorithms described in the previous chapter. The reason for selecting these two methods was because they both use the acceleration signal, not just its peaks. Additionally, DTW is a similarity algorithm that could be used to calculate the similarity of two signals that are out of phase. This is very useful for the walking activities because humans might walk with a similar pattern but with different speeds. This algorithm enables calculation of the number of steps even if the template is not identical to the recorded acceleration signal. The aim of developing a new algorithm was to enable more accurate step count predictions mainly in people with walking impairments. The algorithm developed in this chapter is for the wrist location.

5.2 Methodology

5.2.1 Algorithm development

The template-matching using DTW algorithm was developed to automatically detect the number of steps undertaken, solely from wrist acceleration signals.

Studies in the literature used different methods to calculate a template representing one step or one gait cycle. Manually annotated templates was one of the methods, where the researchers have annotated the acceleration signals manually to represent a single step or one gait cycle (Mantilla et al. 2017; Oudre et al. 2018). Another technique used is identifying the peaks of the acceleration signal using local maximum and local minimum between a certain distance (index). Based on the identification of the peaks, a template is generated by averaging the sequences identified (Ailisto et al. 2005). Lastly, another technique used for the template generation is clustering acceleration windows using K-means to identify the most representative length for the template. Then, for each of the clusters a reference signal and its length are calculated. All the acceleration windows are temporally realigned using DTW to have similar length to the reference signal. The final step is to average all the realigned signals to create the template (Mantilla et al. 2017). The average can be calculated using different methods such as Euclidean distance, cross-correlation and DTW (Xu et al. 2017).

The algorithm was inspired by several authors (Kaptein et al. 2014; Micó-Amigo et al. 2016; Xu et al. 2017), but mainly by (Micó-Amigo et al. 2016), who created a personalised template of a walking period. The new template-matching using DTW devised here differs in two respects; 1) peaks and troughs are used to calculate the template length of a single step instead of just peaks, 2) DTW is used to identify the number of steps instead of correlation measures. The algorithm which was created by (Micó-Amigo et al. 2016) is referred in the text as "template-matching" algorithm. All computation was done using python and the source code is available in a github repository `https://github.com/ValeriaF22/Thesis-Project`. Additionally, the DTW algorithm used was applied using the dtw function from the tslearn library.

The main idea of the algorithm is to create a template for a single step using the acceleration (vertical component only), and then this template will be used against the acceleration signal. If the template matches with a section of the acceleration signal, it will be assumed that this section is a step. The final outcome is the total number of steps.

The flowchart, shown in Figure 5.1 below, demonstrates the overall algorithm.

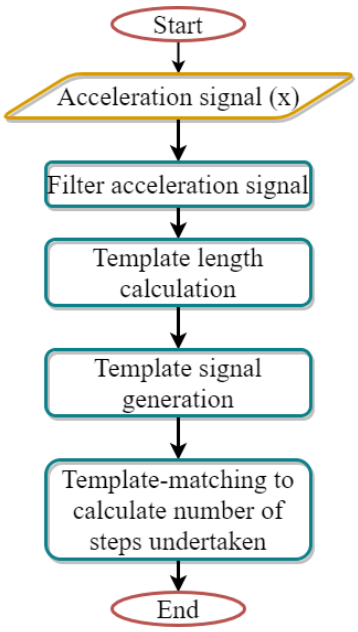


Figure 5.1: Flowchart of template-matching using dynamic time warping algorithm

Unlike many standard step count algorithms, which only consider the number of peaks in a cyclical signal, the main idea behind the template-matching using DTW algorithm is to use a template that models the overall shape of the acceleration signal.

5.2.1.1 Template length calculation

Initially, the acceleration signal was filtered once again to reduce any other noise with a 4th order Butterworth filter and a 2 Hz cut-off frequency. After this, the following steps are performed:

1. Obtain the unbiased autocorrelation signal of the input signal as shown in Figure 5.2.

$$A_{unbiased} = \frac{1}{N - |m|} \sum_{i=1}^{N-|m|} x_i x_{i+m} \tag{5.1}$$

   Where x_i is a time series, N is the total number of time series used and m is a lag parameter.

2. Find the peaks of the autocorrelation signal.

3. Calculate the index (distance) difference between the first two peaks of the autocorrelation signal.

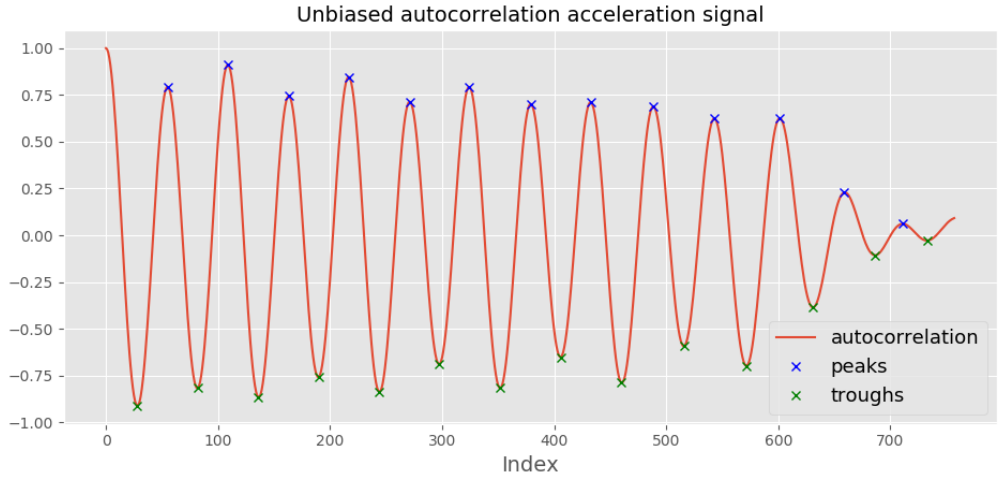


Figure 5.2: Unbiased autocorrelation signal of normal walking with peaks and troughs

4. Calculate an adaptive threshold that will be used for the minimal horizontal index (distance) in samples between neighbouring peaks. The adaptive threshold depends on: the index difference calculated in step 3 and a constant identified experimentally using the data collected. The constant depends on the activity performed and its associated condition (normal or simulated-pathological). Table 5.1 below shows the different values of the constant, which were computed via trial and error. These thresholds might need to be updated when larger dataset is used. However, the ideal scenario is to have personalised step count algorithms, which means that every person would have his/her own thresholds.

Table 5.1: Constant used to calculate adaptive threshold for each individual dynamic activity.

	Normal	Simulated-pathological
Slow walk	0.65	0.95
Normal walk	0.10	0.90
Fast walk	0.10	0.75
Upstairs	0.10	1.30
Downstairs	0.10	1.30

5. Find the peaks of the original acceleration signal using the adaptive threshold calculated in step 4 which is used for the minimum distance between neighbouring peaks.

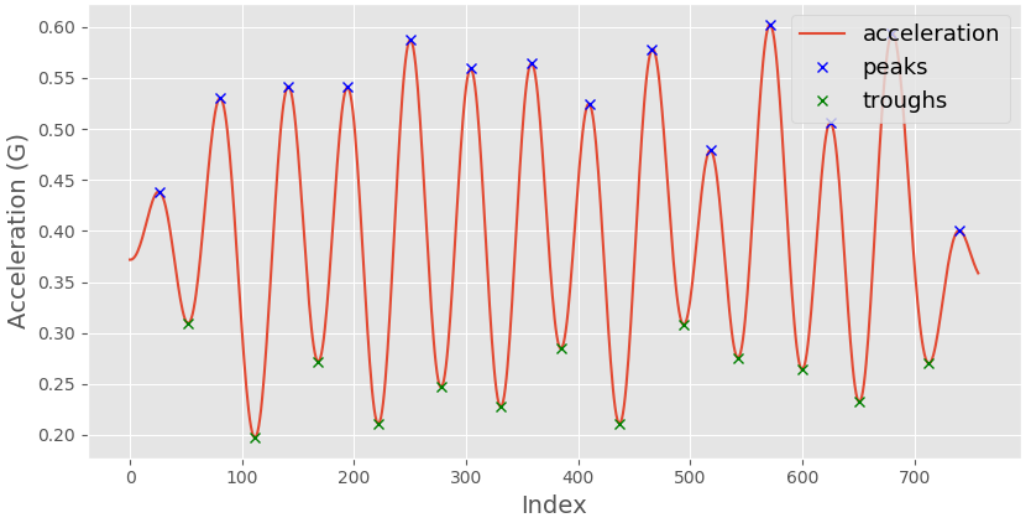


Figure 5.3: Acceleration signal of normal walking with peaks and troughs

6. Calculate the index difference between each consecutive pair of peaks identified in step 5.

$$peak\ index\ difference = peak\_index_{i+1} - peak\_index_i \tag{5.2}$$

7. Calculate the mean index difference of all the pairs of peaks calculated in step 6.

$$mean\ peak\ index\ difference = mean(peak\ index\ difference) \tag{5.3}$$

8. Repeat steps 3-7 for troughs.

9. Calculate the template length as the mean of the mean index difference of peaks and the mean index different of troughs. The index difference of peaks and troughs for each participant is represented by two arrays. Subsequently, the mean of each array is calculated and the mean of the two means is calculated as a final result.

$$template\ length = mean(mean\ peak\ index\ difference + mean\ trough\ index\ difference) \tag{5.4}$$

5.2.1.2 Template signal generation

Original acceleration signal and template length are the inputs in this calculation and the following steps were followed:

1. Segment the acceleration signal, using a sliding window technique without overlapping, into windows of exactly the same length to the template length.
2. Generate a template (average) signal representing one step using the windows created in the previous step. The template signal is generated using DBA method as demonstrated in Figure 4.13

5.2.1.3 Template-matching to calculate number of steps undertaken

Template signal and acceleration windows are the inputs in this calculation. The DTW technique was used to calculate the similarity between the template signal and each segmented window for each participant. This technique was developed to solve any difficulties identified when analysing pattern similarity for time-series data. It is used to evaluate the similarity between two time-series data that might vary in non-linear time-series data and time frames (Lee 2019).

1. Calculate a similarity score between template length and each segmented acceleration window using DTW.
2. Calculate an adaptive threshold that will be used as a threshold for the similarity score. The adaptive threshold depends on the maximum and minimum similarity scores identified for each participant, and a constant identified experimentally using the data collected. The constant depends on the activity performed and its associated condition. Table 5.2 below shows the different values of the constant.

Table 5.2: Constant used to calculate adaptive DTW threshold for each individual dynamic activity.

	Normal	Simulated-pathological
Slow walk	0.90	0.90
Normal walk	0.90	0.40
Fast walk	0.90	0.40
Upstairs	0.90	0.00
Downstairs	0.90	0.00

$$mid\_range\_similarity\_score = \frac{\max(similarity\_score) + \min(similarity\_score)}{2} \quad (5.5)$$

$$DTW\_threshold = mid\_range\_similarity\_score + (mid\_range\_similarity\_score \times constant) \quad (5.6)$$

3. A step is counted if the similarity score is below the $DTW\_threshold$

5.2.2 Data analysis

The analysis was conducted to compare the template-matching using DTW algorithm with multiple different methods from the literature to check which method performs best (refer to chapter 4). The comparisons of these methods were made for both normal and simulated-pathological walking activities. The true number of steps was measured from the gold standard video. This was compared with the predicted number of steps generated from the step count algorithms mentioned above. Both the predicted and true number of steps were calculated for each activity separately. The same steps are followed with the analysis of chapter 4, however in this chapter the analysis is more advanced.

The percentage error between the predicted and true number of steps was first visualised as a box plot. A box plot displays the data distribution as the first and third quartiles, minimum, maximum and median. Also, a box plot shows any outliers of the data, how data is grouped and skewed. This method was used to enable the visualisation of a simple measure, the percentage error. The box plot was used because it could show the step differences for each algorithm, and thus understand which step-count algorithms produce consistent errors or not.

RMSE was calculated to measure the difference between predicted and true number of steps per activity. It is the standard deviation of the prediction errors, which represent the vertical distance between the regression line and the data point. The residuals measure how far the data points are from the regression line. An RMSE closer to zero indicates minimum error. The main reason for choosing this particular measure to evaluate the accuracy of the step count algorithms was because it penalises large errors. This is important in this case because we would like to avoid large errors.

The results were also assessed using a modified version of Bland-Altman analysis. This first involved creating the plot to visually inspect the agreement among the predicted and true number of steps. The Bland-Altman plot is a scatter plot, in which Y axis demonstrates the difference between the two paired measurements and the X axis shows the average of these measures (Giavarina 2015). In our case, the X axis shows the true number of steps instead of the average of the two methods (Giavarina 2015; Sasko et al. 2018). Additionally, in terms of limits of agreement (LOA), V-shaped 95% LOA were used because the bias was not proportional to the number of steps (Hans 2015). The LOA represent the range where most differences between the measurements of the two methods, gold standard and algorithm, will lie. They are calculated based on the mean and standard deviation of the differences between the measurements (Bland and Altman 1999). This type of plot is useful to illustrate the agreement between two methods.

5.3 Results

5.3.1 Normal condition

The performance of the step count algorithms was examined on a set of healthy volunteers, while performing the activities under normal conditions.

5.3.1.1 Root mean square error

The template-matching using DTW algorithm yielded better results than the other algorithms for both slow and normal walk activities. However, the results of the template-matching using DTW algorithm were close to the results of the template-matching without DTW algorithm for all the activities (normal walk, fast walk, ascent and descent stairs) except the slow walk. The results of the other three algorithms, peak detection, thresholding (frequency-domain) and thresholding (time-domain) were worse than the template-matching using DTW algorithm. In terms of activities, normal walk yielded the least error in almost all the step count algorithms. Table 5.3 demonstrates the RMSE between the true and predicted number of steps along with the lower and upper confidence intervals. It is important to note that the lower the RMSE the better the model fit. In other words, the algorithm had predicted well the number of steps.

Table 5.3: Results of step count algorithms for individual dynamic activities under normal condition using true activity labels.

Activities	Algorithms				
	Peak detection	Thresholding (F-domain)	Thresholding (T-domain)	Template-matching	Template-matching using DTW
Slow walk	2.81 (2.70, 2.92)	6.61 (6.34,6.86)	9.10 (8.76,9.43)	8.99 (8.34,9.59)	**1.65** (1.31,1.71)
Normal walk	4.03 (3.91,4.16)	4.09 (3.93,4.23)	6.75 (6.67,6.84)	1.35 (1.28,1.42)	**1.31** (0.88,1.01)
Fast walk	5.07 (4.98,5.16)	5.85 (5.74,5.96)	5.82 (5.76,5.88)	**1.74** (1.63,1.84)	2.33 (2.25,2.45)
Stair ascent	4.55 (4.41,4.69)	5.37 (5.21,5.52)	6.37 (6.31,6.42)	**1.90** (1.76,2.03)	2.24 (1.87,2.1)
Stair descent	5.16 (4.98,5.32)	6.36 (6.23,6.48)	6.81 (6.73,6.89)	**1.11** (1.08,1.15)	2.69 (2.40,2.74)
Average	4.32	5.66	6.97	3.02	1.88

5.3.1.2 Performance by activity

Slow walk Figure 5.4 shows that, for slow walking, the most accurate algorithm was the new algorithm developed in this chapter, template-matching using DTW. Despite using similar methods, the performance of template-matching without DTW was highly variable in comparison to the template-matching using DTW algorithm.

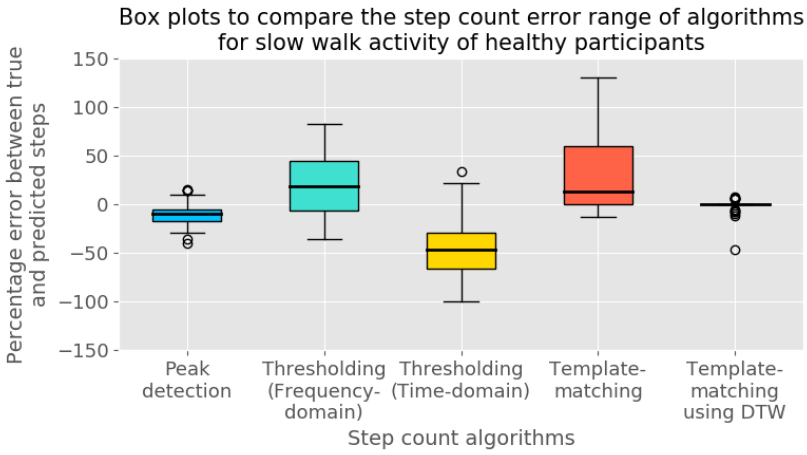


Figure 5.4: Percentage error between true and predicted number of steps using box plots for slow walk activity under normal condition

For the template-matching using DTW algorithm, the bias estimated was -0.31, which is close to zero as shown in Figure 5.5. No change in bias was observed with increasing number of

steps. The 95% LOA in this range were -3.23 and 2.61. This compares favourably with the other algorithms, in which the minimum LOA ranged from -6.16 to 3.20 and the maximum LOA ranged from -9.15 to 19.57.

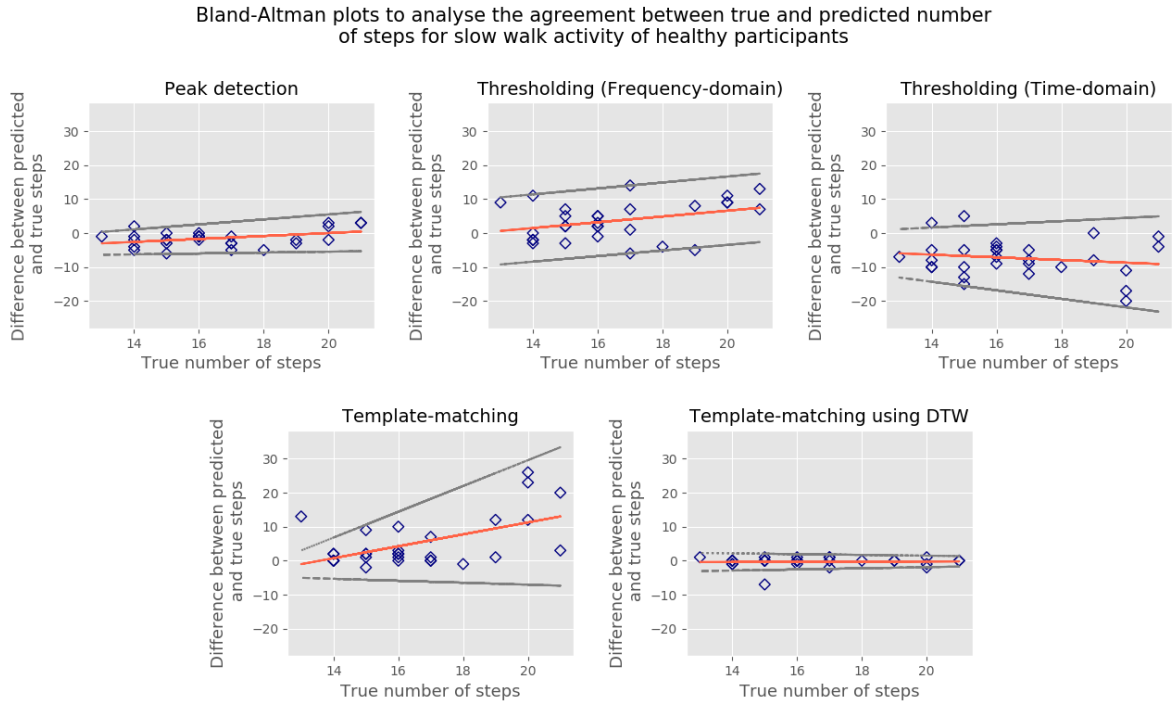


Figure 5.5: Difference between true and predicted number of steps using modified Bland-Altman plots for slow walk activity under normal condition

Normal walk Figure 5.6 demonstrates that, for normal walking, the two most accurate algorithms were the template-matching using DTW and template-matching. Template-matching using the DTW algorithm performed better than the template-matching without using DTW. For the template-matching using DTW algorithm, the majority of the participants had zero error between the predicted and true number of steps. On the other hand, the template-matching without DTW algorithm overestimated the number of steps in the majority of the participants.

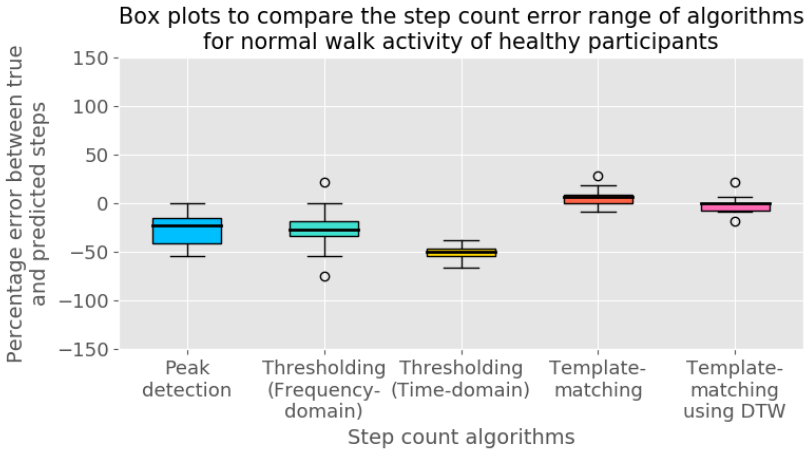


Figure 5.6: Percentage error between true and predicted number of steps using box plots for normal walk activity under normal condition

For normal walking, the template-matching using DTW and the template-matching algorithms, showed similar performance (Figure 5.7). However, template-matching using DTW performed better since the bias was smaller, and the range of LOA was also smaller. The bias estimated was -0.28, which is close to zero. The LOA in this range were -2.05 and 1.50. Regarding the template-matching algorithm, the bias estimated was 0.66 and the LOA ranged between -1.66 and 2.97. All the other algorithms showed greater error, where the minimum LOA ranged from 8.91 to -4.40 and the maximum LOA ranged from -8.12 to 1.63.

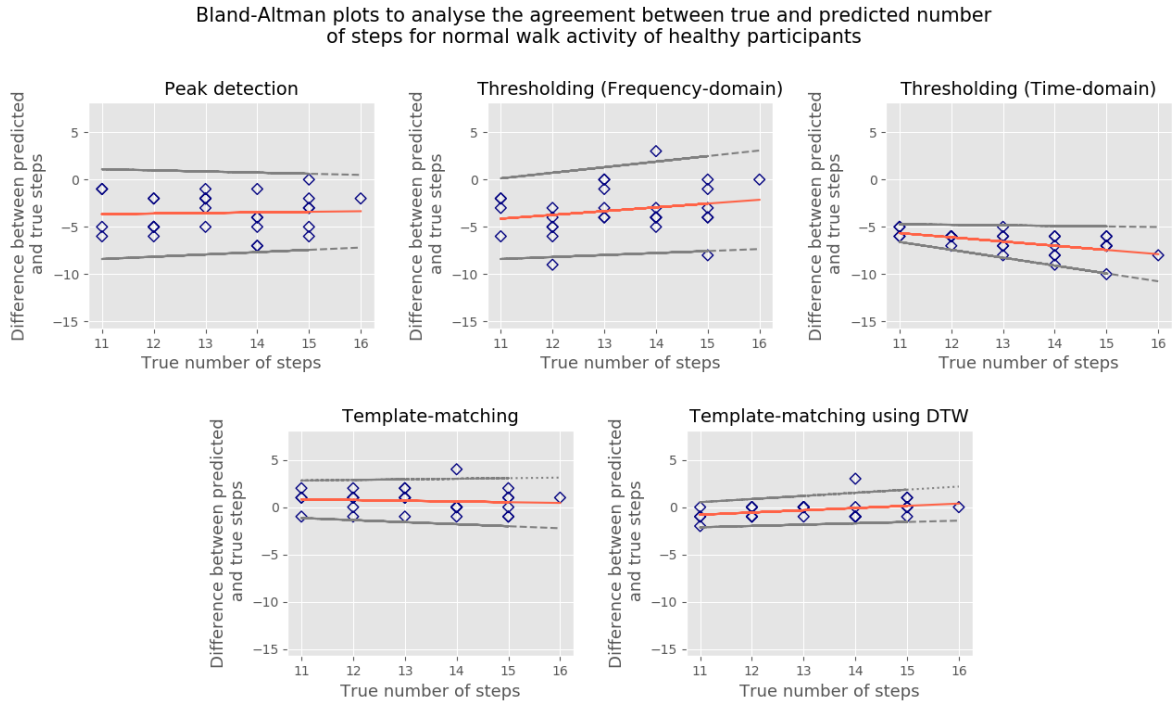


Figure 5.7: Difference between true and predicted number of steps using modified Bland-Altman plots for normal walk activity under normal condition

Fast walk Figure 5.8 shows that for fast walking, the most accurate algorithm was the template-matching. Even though utilising similar methods, the performance of the template-matching using DTW algorithm was more variable than the template-matching without DTW algorithm.

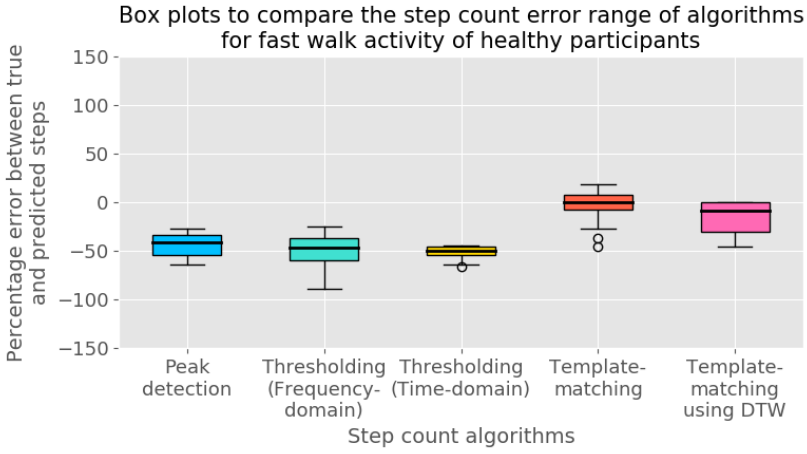


Figure 5.8: Percentage error between true and predicted number of steps using box plots for fast walk activity under normal condition

For fast walking, the template-matching algorithm had the smallest bias, which was -0.34. For the template-matching using DTW algorithm, the bias estimated was -1.72. Considering the

LOA, the two template-matching algorithms and the thresholding (frequency-domain) algorithm had similar and greatest ranges of LOA. Among the three, template-matching using DTW had the smallest range of LOA, which ranged from -4.85 to 1.40. The other two algorithms had smaller range of LOA. The LOA of peak detection ranged from -7.49 to -2.31 and the LOA of thresholding (time-domain) ranged from -7.44 to -4.08 as demonstrated in Figure 5.9.

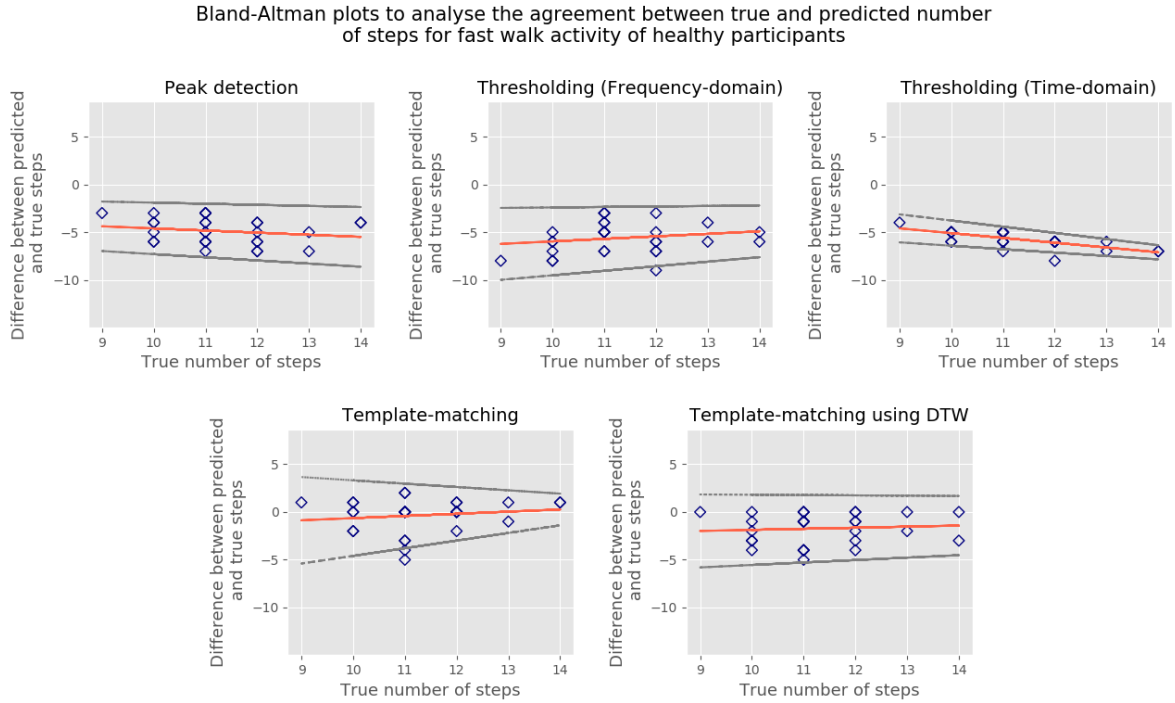


Figure 5.9: Difference between true and predicted number of steps using modified Bland-Altman plots for fast walk activity under normal condition

Ascending stairs Figure 5.10 demonstrates that, for stair ascent, the two most accurate algorithms were the template-matching using DTW and template-matching. The former algorithm performed better than the latter algorithm. The results are similar to the normal walk, where the majority of the participants for the template-matching using DTW had zero error between the predicted and true number of steps. On the other hand, the latter algorithm overestimated the number of steps in the majority of the participants. Despite employing similar techniques, the template-matching using DTW has slightly greater percentage error variance than the template-matching algorithm.

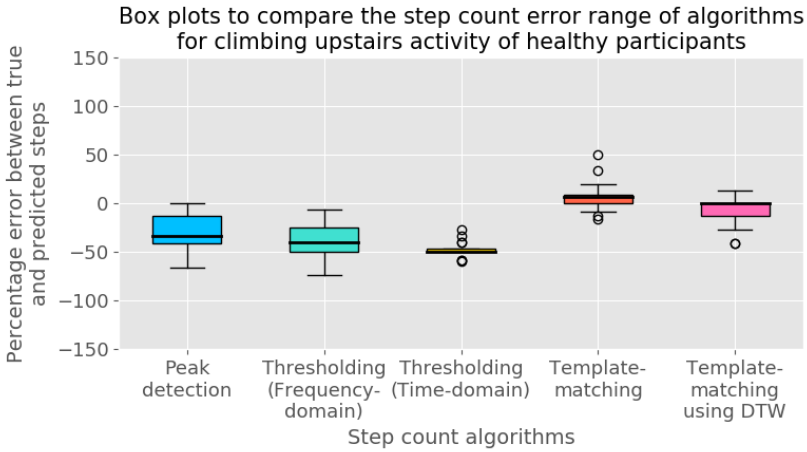


Figure 5.10: Percentage error between true and predicted number of steps using box plots for ascending stairs activity under normal condition

For stair ascent, the template-matching using DTW and the template-matching algorithms, showed similar performance. For the former, the bias estimated was -0.93 and the LOA ranged from -4.38 to 2.52. For the latter, the bias estimated was 0.72 and the LOA ranged from -2.72 to 4.17. The LOA of the thresholding (time-domain) had the smallest range among all the algorithms, which were between -7.95 and -4.6. The other two algorithms had larger range of LOA as shown in Figure 5.11. The LOA of peak detection ranged from -8.66 to 1.00 and the LOA of thresholding (frequency-domain) ranged from -9.31 to -0.41.

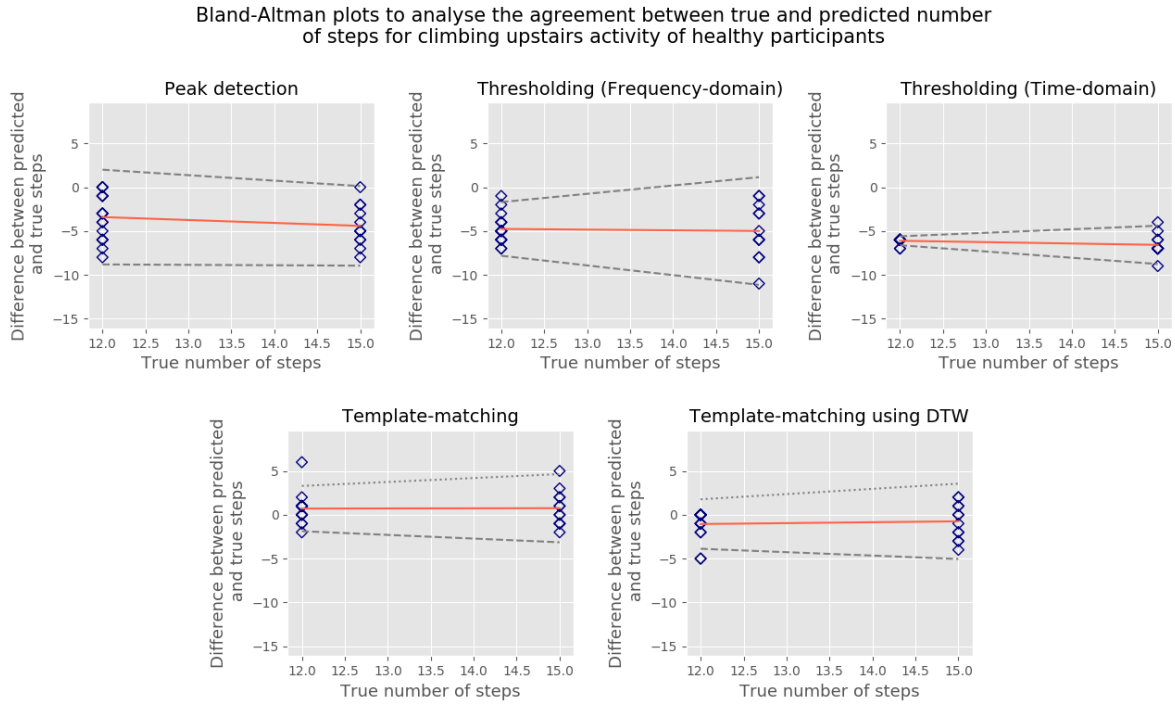


Figure 5.11: Difference between true and predicted number of steps using modified Bland-Altman plots for ascending stairs activity under normal condition

Descending stairs Similarly to fast walking, Figure 5.12 demonstrates that, for stair descent, template-matching was the most accurate algorithm. Even though utilising similar methods, the performance of the template-matching using DTW algorithm was more variable the template-matching algorithm.

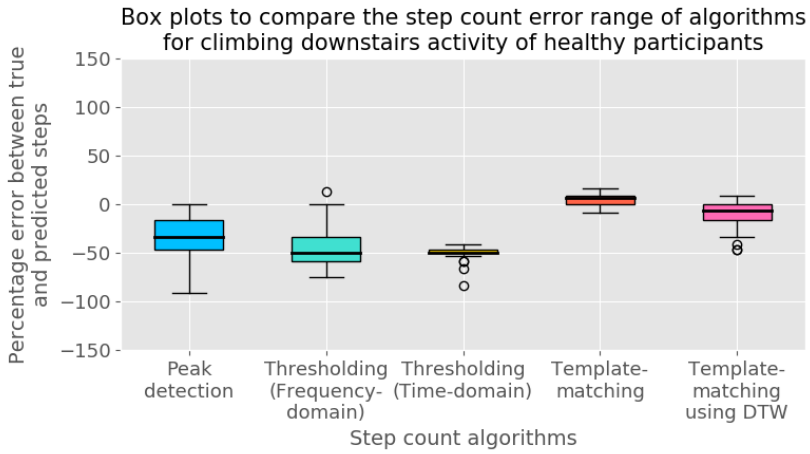


Figure 5.12: Percentage error between true and predicted number of steps using box plots for descending stairs activity under normal condition

For stair descent, the template-matching algorithm had the smallest bias, which was 0.62, and

the smallest range of LOA, which were between -1.19 and 2.43. For the template-matching using DTW algorithm, the bias estimated was -1.45, which was the second smallest. However, this algorithm had the third largest range of LOA among the other algorithms. The LOA ranged from -5.62 to 2.72. Among the other three algorithms, the LOA of thresholding (time-domain) ranged from -8.84 to -4.61. This was the second smallest range among all the five algorithms. The other two algorithms had larger range of LOA. The LOA of peak detection ranged from -9.56 to 0.66 and the LOA of thresholding (frequency-domain) ranged from -10.82 to -0.84 as demonstrated in Figure 5.13.

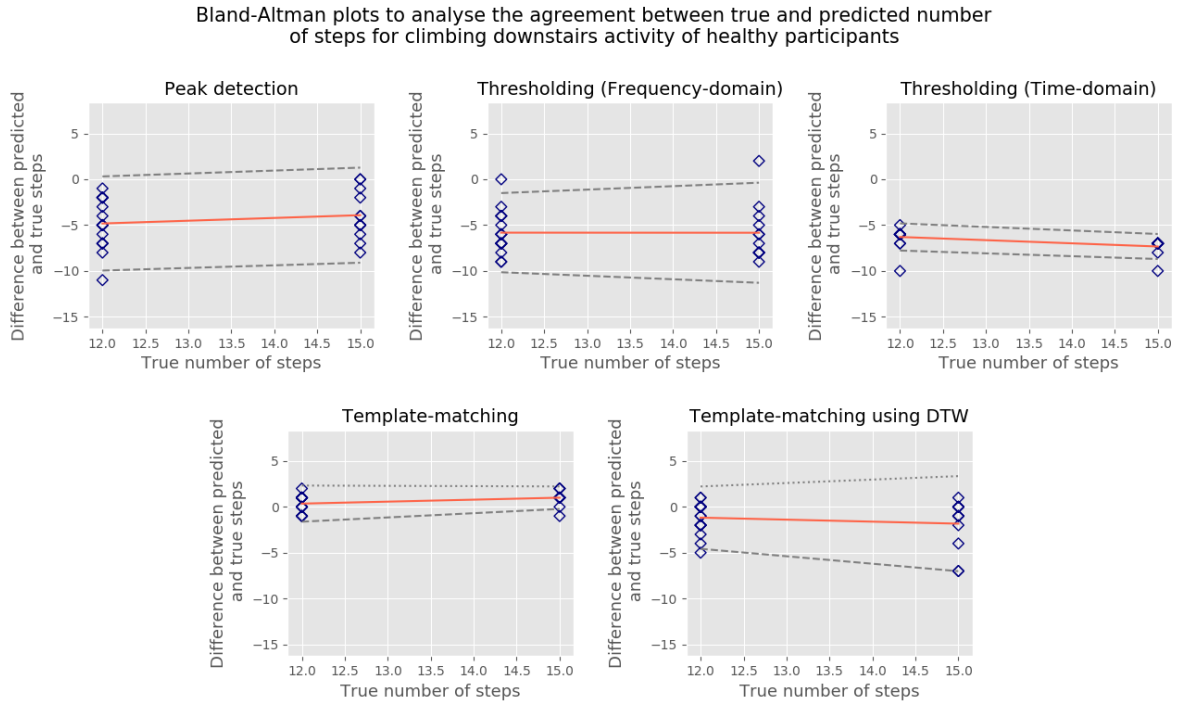


Figure 5.13: Difference between true and predicted number of steps using modified Bland-Altman plots for descending stairs activity under normal condition

5.3.2 Simulated-pathological condition

After examining the algorithms on the healthy volunteers under normal conditions, the performance of the algorithms was examined on a set of healthy volunteers while performing the activities under simulated-pathological conditions.

5.3.2.1 Root mean square error

The template-matching using DTW algorithm had better results than all the other algorithms in all the activities. However, for the slow walking results, the template-matching using DTW

algorithm had similar RMSE with peak detection algorithm. The results of the thresholding (frequency-domain) and template-matching algorithms had great difference from the results of the template-matching using DTW algorithm. Table 5.4 demonstrates the RMSE between the true and predicted number of steps along with the lower and upper confidence intervals. Contrarily to the results of the healthy volunteers under normal condition, the RMSE was greater in this case. This means that the algorithm did not predict the number of steps as well as the normal condition.

Table 5.4: Results of step count algorithms for individual dynamic activities under simulated-pathological condition using true activity labels.

Activities	Algorithms				
	Peak detection	Thresholding (F-domain)	Thresholding (T-domain)	Template-matching	Template-matching using DTW
Slow walk	20.80 (18.41,22.94)	27.68 (26.55,28.78)	32.87 (31.7,34.01)	84.07 (76.36,91.14)	**20.67** (19.08,22.15)
Normal walk	15.94 (13.76,17.87)	23.13 (21.98,24.22)	25.94 (24.95,26.89)	50.87 (47.03,54.45)	**13.96** (13.23,14.66)
Fast walk	12.64 (11.79,13.44)	15.81 (14.74,16.8)	15.81 (14.74,16.8)	43.02 (38.56,47.06)	**9.25** (8.73,9.74)
stair ascent	10.01 (9.43,10.55)	38.06 (37.03,39.07)	**8.75** (8.44,9.06)	42.12 (39.18,44.88)	8.89 (5.53,6.42)
stair descent	9.07 (8.27,9.80)	34.63 (33.62,35.61)	10.46 (9.99,10.91)	59.22 (56.15,62.13)	**7.74** (7.17,8.28)
Average	13.69	27.86	34.82	55.86	11.52

5.3.2.2 Performance by activity

In the following graphs, the range of error detected was much larger for the simulated-pathological group in comparison to the healthy group under normal conditions.

Slow walk Figure 5.14 demonstrates that, for slow walking, peak detection algorithm had the greatest accuracy. The second most accurate algorithm is the template-matching using DTW. Even though the two template-matching algorithms use similar methods to calculate the number of steps, the template-matching had greater variability than the template-matching using DTW. Additionally, it can be seen that there is one outlier with very large error for the template-matching algorithm.

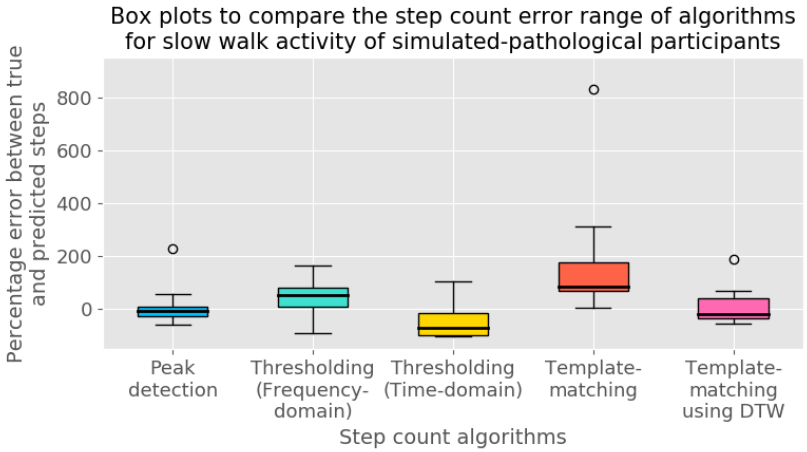


Figure 5.14: Percentage error between true and predicted number of steps using box plots for slow walk activity under simulated-pathological condition

The template-matching using DTW and peak detection algorithms showed similar performance as demonstrated in Figure 5.15. For the former, the bias estimated was 1.79 and the LOA ranged from -38.57 to 42.16. For the latter, the bias estimated was 1.28 and the LOA ranged from -39.42 to 41.97. The LOA of the thresholding (time-domain) had the smallest range among all the algorithms, which were between -70.21 and 27.10. The template-matching algorithm showed the worse performance at slow walking. The LOA of this method ranged from -64.90 to 178.35. The bias of template-matching was 56.72, and it increases as the number of walking steps increases.

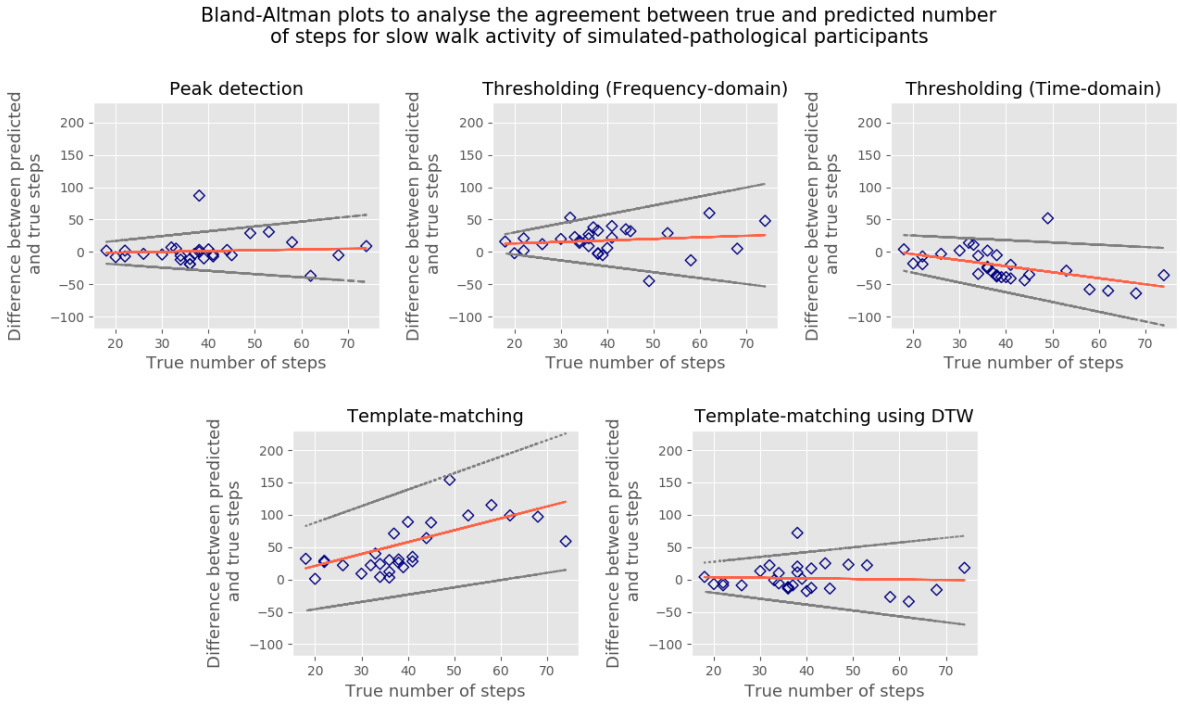


Figure 5.15: Difference between true and predicted number of steps using modified Bland-Altman plots for slow walk activity under simulated-pathological condition

Normal walk Similarly to slow walking, the same pattern of results was observed for normal walking as well (Figure 5.16). The peak detection algorithm had the least percentage error. The second most accurate algorithm was the template-matching using DTW. Once again, despite the fact that the two template-matching algorithms employed similar methods to calculate the number of steps, the template-matching demonstrated greater error than the template-matching using DTW. Additionally, it can be seen that there was one outlier with very large error for the template-matching algorithm.

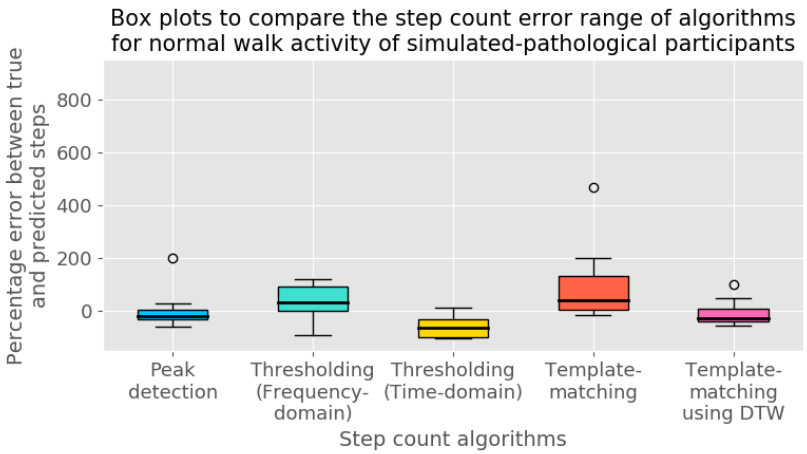


Figure 5.16: Percentage error between true and predicted number of steps using box plots for normal walk activity under simulated-pathological condition

For normal walking, the template-matching using DTW algorithm had the second smallest bias, which was -3.90, and the smallest range for the LOA, which was between -30.17 to 22.38. For the peak detection algorithm, the bias estimated was -1.45, which was the smallest among all the algorithms as shown in Figure 5.17. Additionally, this algorithm had the third smallest range of LOA among the other algorithms. The LOA ranged from -32.87 to 29.22. Among the other three algorithms, the LOA of thresholding (time-domain) ranged from -51.36 to 9.98. This was the second smallest range among all the five algorithms. The other two algorithms had larger range of LOA. The LOA of thresholding (frequency-domain) ranged from -22.09 to 50.09 and the LOA of template-matching ranged from -46.40 to 109.71.

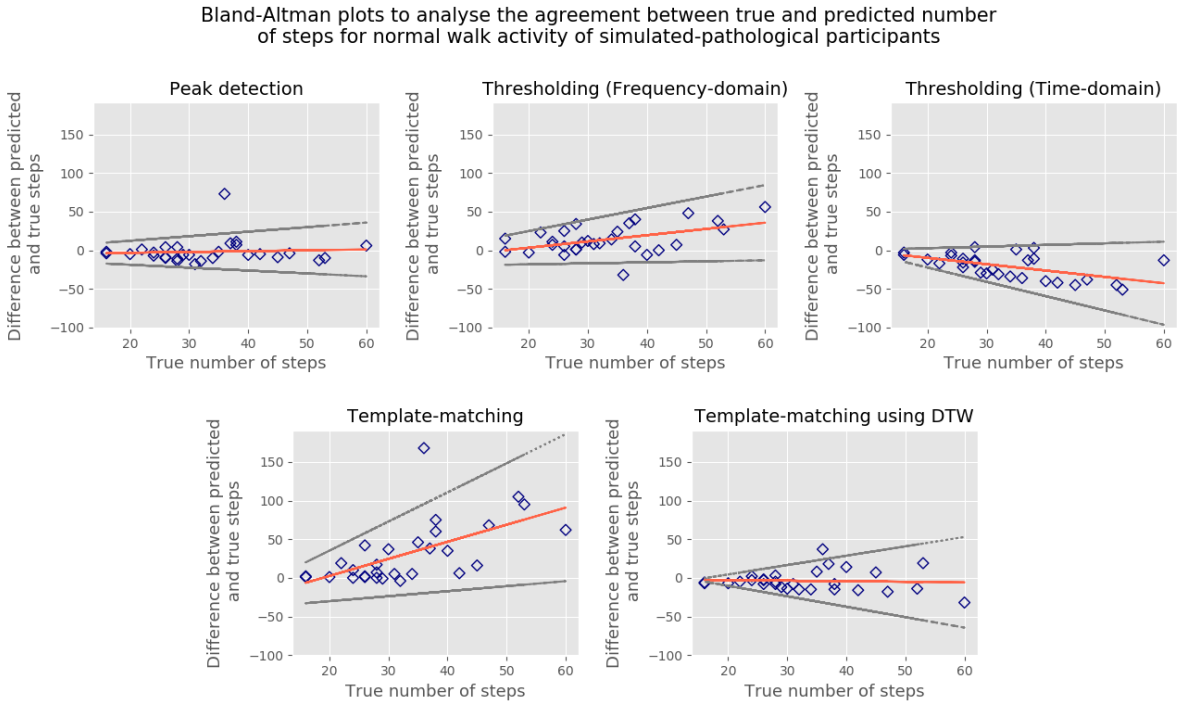


Figure 5.17: Difference between true and predicted number of steps using modified Bland-Altman plots for normal walk activity under simulated-pathological condition

Fast walk In the case of fast walking, the template-matching using DTW algorithm had the smallest percentage error as demonstrated in Figure 5.18. Despite using similar methods, the performance of template-matching was very variable in comparison to the template-matching using DTW algorithm.

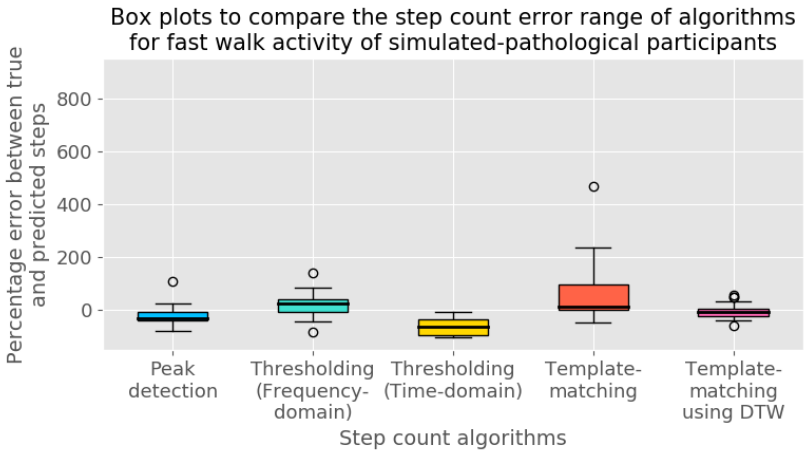


Figure 5.18: Percentage error between true and predicted number of steps using box plots for fast walk activity under simulated-pathological condition

For the template-matching using DTW algorithm, the bias estimated was -0.14, which is close

to zero. No change in bias was observed with increasing number of steps. However, the LOA ranged from -18.27 to 17.99. Even though, large LOA were observed for the template-matching using DTW algorithm (Figure 5.19), this compares favourably with all of the other algorithms, in which the minimum LOA ranged from -27.79 to 17.30 and the maximum LOA ranged from -51.20 to 94.51.

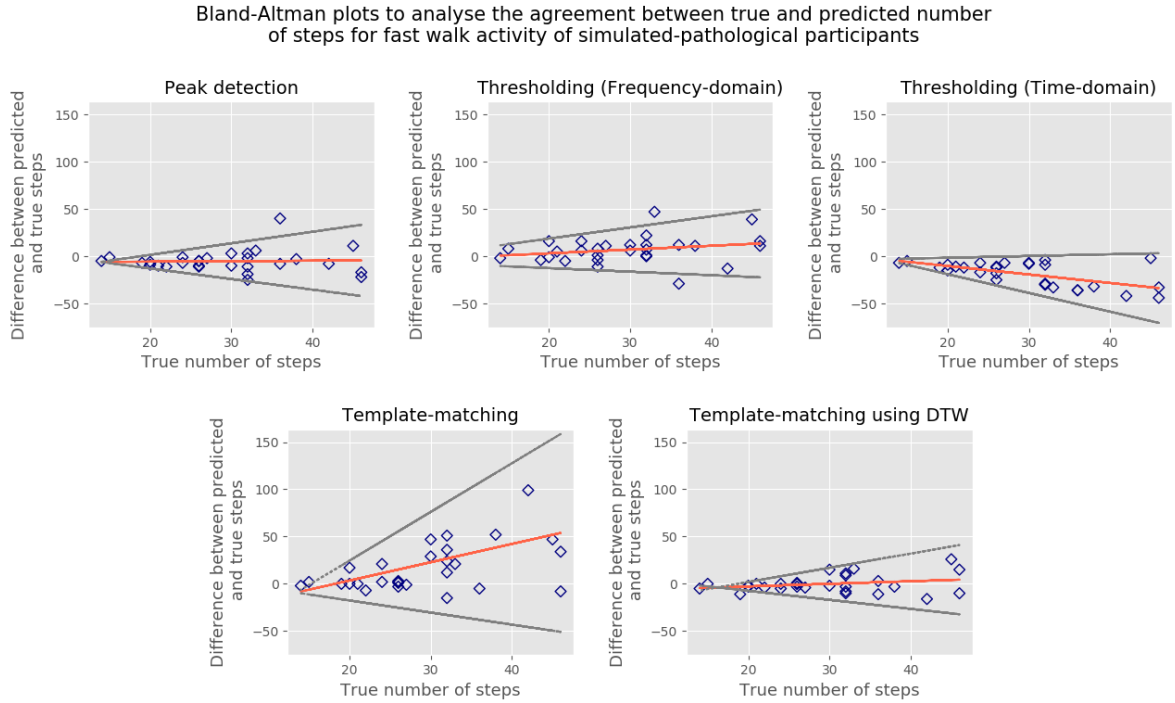


Figure 5.19: Difference between true and predicted number of steps using modified Bland-Altman plots for fast walk activity under simulated-pathological condition

Ascending stairs Similar to fast walking, Figure 5.20 shows that for stair ascent, the greatest accuracy was achieved by the template-matching using DTW algorithm.

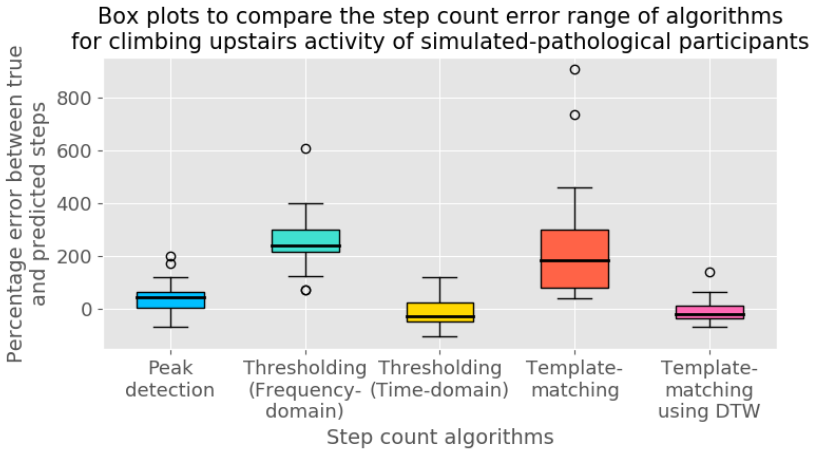


Figure 5.20: Percentage error between true and predicted number of steps using box plots for ascending stairs activity under simulated-pathological condition

Again, the template-matching using DTW algorithm showed the best overall performance as shown in Figure 5.21. The bias estimated was -0.97 and the LOA ranged from -12.56 to 10.63. The template-matching algorithm showed the poorest performance in terms of the LOA. Similarly to all the aforementioned activities, the LOA of this method was large and it ranged from -23.81 to 86.50.

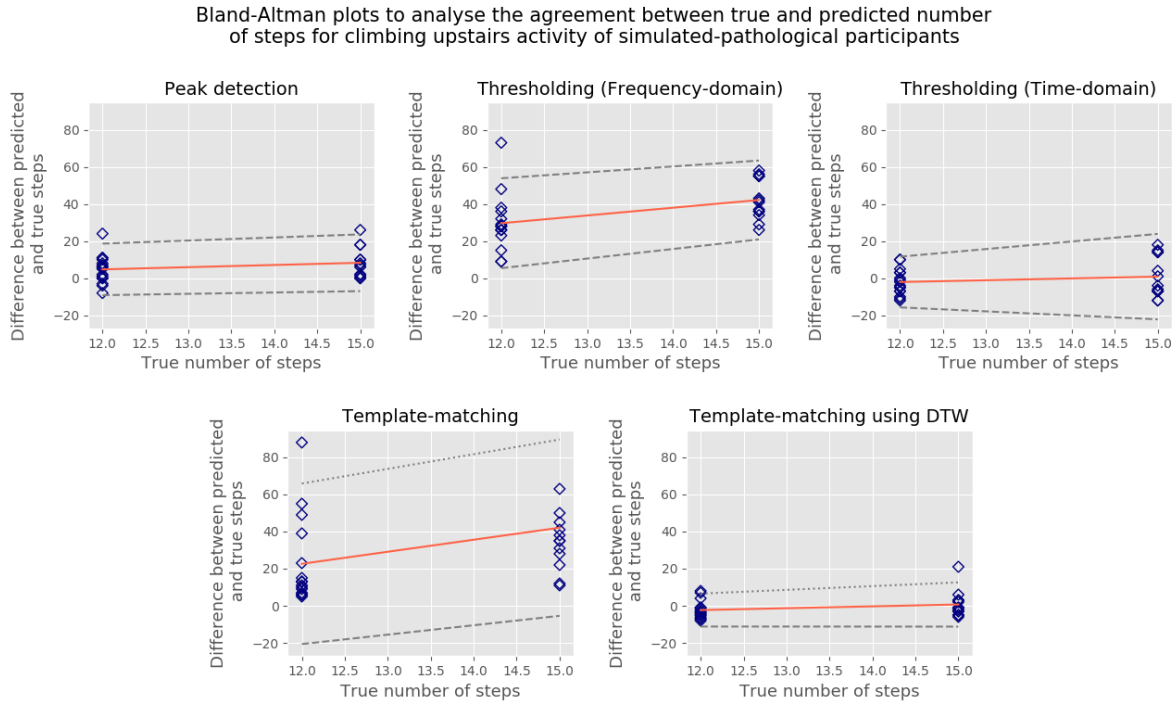


Figure 5.21: Difference between true and predicted number of steps using modified Bland-Altman plots for ascending stairs activity under simulated-pathological condition

Descending stairs Figure 5.22 demonstrates that, for stair descent, the two most accurate algorithms were the peak detection and template-matching using DTW. Again, the template-matching algorithm showed greater variability than template-matching using DTW.

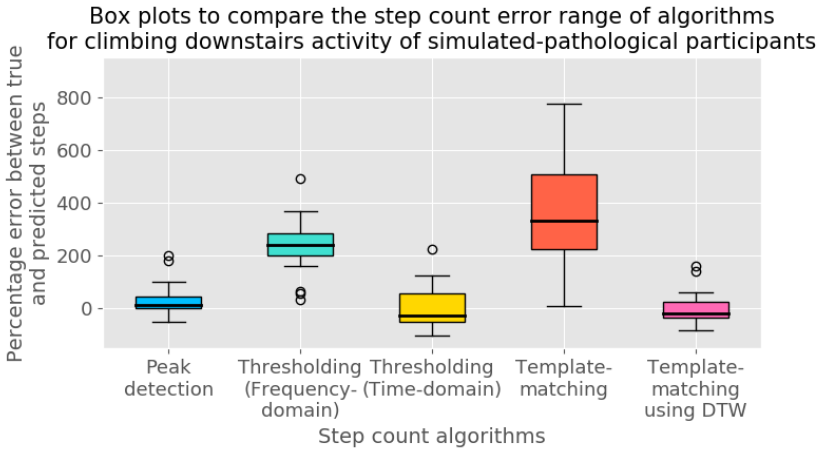


Figure 5.22: Percentage error between true and predicted number of steps using box plots for descending stairs activity under simulated-pathological condition

The template-matching using DTW algorithm showed the best overall performance. The bias estimated was 0.28 and the LOA ranged from -14.89 to 15.44. The range of LOA of the peak detection was similar to the template-matching using DTW algorithm, which was between -10.21 and 19.87. The template-matching algorithm showed the worse performance in terms of both bias and LOA as demonstrated in Figure 5.23. The estimated bias was 49.93 and the LOA ranged from -12.46 to 112.32.

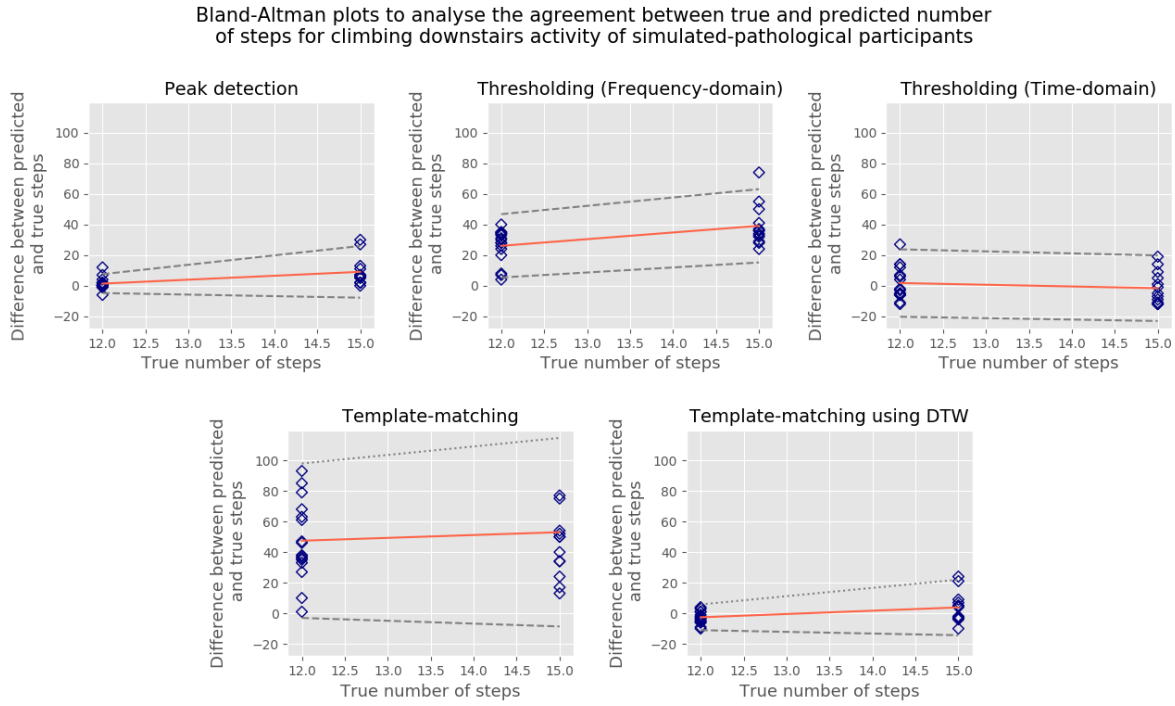


Figure 5.23: Difference between true and predicted number of steps using modified Bland-Altman plots for descending stairs activity under simulated-pathological condition

5.3.3 Testing the Template-matching DTW algorithm in an external public dataset

In order to further validate the ability of the template-matching using DTW algorithm, it was tested using the Oxford-Step-Counter external dataset [1]. This dataset includes only two subjects, for each subject the raw acceleration signal is included along with the true number of steps performed from each volunteer. The reason for using this dataset, even though only two subjects were available, is the fact that it was the only dataset that included the true number of steps. The dataset is available at: `https://github.com/Oxford-step-counter/DataSet/tree/master/validation`. This data set was collected using a Samsung S6 smartphone in six different positions: 1) hand, 2) front pocket, 3) back pocket, 4) neck pouch, 5) bag, and 6) armband. However, for the purpose of this thesis, only the data obtained from the hand and armband locations were used. (Pham et al. 2018) have also previously tested their algorithm on the Oxford-Step-Counter dataset and comparative results are presented below in Table 5.5. Pham et al developed a step count algorithm that used peak detection along with four features: minimal peak distance, minimal peak prominence, dynamic thresholding, and vibration elim-

[1] https://github.com/Oxford-step-counter/DataSet/tree/master/validation

ination. The template-matching using DTW algorithm mainly overestimated the number of steps and the algorithm developed by (Pham et al. 2018) underestimated the number of steps. Using the template-matching using DTW algorithm, the greatest difference between true and predicted number of steps was found to be four steps. Conversely, eight steps was the greatest difference between true and predicted number of steps while using the algorithm developed by Pham et al (Pham et al. 2018).

Table 5.5: Number of steps calculated using template-matching DTW and (Pham et al. 2018) algorithms using the Oxford-step-counter dataset.

Positions	User	True steps	Template-matching using DTW	Peak detection & four features
Hand	1	326	328	323
Hand	2	340	344	332
Armband	1	335	338	335
Armband	2	343	341	335

5.4 Discussion

This chapter presents the step count analysis of data collected from wrist-worn accelerometers in volunteers performing walking activities under normal and simulated-pathological conditions. The results presented here answer the fourth research question posed in section 2.5.1; can we accurately measure step count in healthy participants under normal and simulated-pathological gaits?

Previous researchers used a variety of approaches and methods to study the accuracy of counting the number of steps (Rhudy and Mahoney 2018; Bui et al. 2018; Ao et al. 2018; Bunn et al. 2019). The main differences among these studies were: (a) type of device; (b) location of device; (c) population examined; (d) activities performed; (e) variation in walking speed; and (f) algorithms used for analysis. Three main type of devices were used throughout the literature; smartphone, IMUs and wearables. These devices were attached on different locations throughout the body in both healthy and patient populations. However, most of the literature focuses on the healthy population rather than patients. The most frequently used locations were: wrist, thigh, ankle/foot, chest, waist and/or pockets. The main finding from the literature was the need for better step count algorithms to be employed in (a) people with slow walking activity; (b) functionally compomised patient populations; and (c) when devices were worn on the wrist.

The template-matching using DTW algorithm was tested using data representing both normal

and simulated-pathological conditions. The results agreed with the literature as the accuracy of the step count algorithm was higher (RMSE: 1.88) in the normal data, in comparison to data from the simulated-pathological gait (RMSE: 11.52). This could be seen from the average RMSE results calculated for the template-matching using DTW algorithm. The average RMSE for the normal and the simulated-pathological error was 1.88 and 11.52 respectively. The template-matching using DTW algorithm was compared with four step count algorithms from the literature and these four algorithms represented a wide variety of possible techniques. In both the normal and simulated-pathological states, the new template-matching using DTW algorithm achieved the smallest average RMSE in comparison to the four existing algorithms reported in the literature.

This difference might be due to the fact that some of the existing algorithms were developed for devices in different locations, and they were targeted at different populations. Additionally, their initial goal was calculating the step length, and in order to do that, it was essential to calculate the number of steps first although this might not have been their first main aim.

The template-matching using DTW algorithm was tested for slow, normal, and fast walking activities, as well as for stair ascent and descent, the latter two activities were not well reported in the literature. When participants walked normally, the template-matching using DTW algorithm showed low RMSE (< 2.69) in all the tested activities, including the slow walk, stairs ascent, and stairs descent activities. Similarly, in the results from the simulated-pathological condition, the template-matching using DTW algorithm demonstrated better results for all the activities. Conversely, this algorithm produced a larger RMSE in the slow and normal walking activities. Even though in most of the activities the template-matching using DTW algorithm achieved the smallest RMSE when compared with the four literature algorithms, the error was still greater than the errors calculated for the normal condition group. The difference of the results between the two groups might be due to several reasons. In general, the acceleration signal recorded from the normal condition group is periodic since walking process is a rhythmic movement (Menz et al. 2003; Yan et al. 2020). On the other hand, the acceleration signal recorded from the pathological condition group (in this case simulated-pathological) is less regular and contains greater noise. It becomes harder to calculate the number of steps using an irregular signal rather than a periodic signal. Additionally, the variability of the walking signal has an important role especially for the simulated-pathological condition. Healthy participants

performed activities under their normal routine and also under a pathological condition. In their attempt to perform the activities under pathological conditions, the variability of their gait was much greater. Bland-Altman plots showed the range of LOA for the normal condition was much smaller than the range of LOA of the simulated-pathological condition. This suggested that the acceleration signal was irregular in most simulated-pathological cases, therefore the autocorrelation method did not perform as well as under the normal condition.

In terms of the activities associated with the normal condition group, the template-matching using DTW algorithm and most of the algorithms developed and tested in the literature performed exceptionally well. Under the simulated-pathological conditions though, the results from literature and the template-matching using DTW algorithm were not as accurate as the results from the normal condition. Even though this was the case, the results of the template-matching using DTW algorithm demonstrated better performance than the other four algorithms generally.

The template-matching using DTW algorithm showed excellent performance using the data from the normal condition, and it also showed improved performance over existing algorithms when using the data from the simulated-pathological condition. This is very encouraging as it enables more accurate step count data to be provided to clinicians for their patients who suffer with chronic diseases.

Finally, the template-matching using DTW algorithm was also tested with the Oxford Step Counter dataset. The reason for that was to validate the algorithm using an unseen external data set. For the purposes of this study, only data from hand and arm were used to validate the proposed algorithm since it was developed according to the wrist location. This location is one of the most popular locations used by the users. The percentage error between the true number of steps and the predicted number of steps was $\pm 1\%$. This suggested that the template-matching using DTW algorithm showed excellent performance using unseen data representing a normal walking condition.

There are some potential limitations associated with the algorithms derived from the literature. The four existing algorithms were recreated by the researcher using python. While these were created as faithfully as possible by the researcher, most of the algorithms were originally developed in different programming languages. Since each language has its own functions, the functions used in python might differ from the ones used for the original algorithm, which might affect the overall performance of the algorithm. Additionally, the existing algorithms were de-

veloped for different device locations, populations, and activities. For example, Thanh et al and Dirican et al developed their algorithm for normal walking and running in a healthy population, and the device was attached on the waist of the participants (Thanh et al. 2017; Dirican and Aksoy 2017). Also, Mico-Amigo and colleagues developed an algorithm to mainly detect steps from devices attached on the lower back and on the heel for the healthy elderly population (Micó-Amigo et al. 2016). Lastly, Palshikar et al developed a simple peak detection algorithm that it was used to detect peaks in different types of signals (Palshikar 2009).

Another limitation was associated with the simulated data which was collected to represent the pathological condition. The variation of the acceleration signal is reflected better among different volunteers (inter-variability) than among different trials of the same volunteer (intra-variability) (Racic and Pavic 2010b; Ponce et al. 2016). However, in this study for some cases, especially under simulated-pathological conditions, there is large variability in the walking activity in each participant. For example, each step undertaken to complete the slow walk activity might differ among each other in terms of step length and step width even though the participant was the same (Whittle 2007).

One more limitation is associated with the performance metrics used to evaluate the existing step count algorithms in the literature. There is no consistent metric used to measure the accuracy of such algorithms, hence each research team selected the most appropriate metric for their study. However, it becomes difficult to compare the results from different studies since several metrics have been used.

Lastly, the template-matching using DTW step count algorithm used a constant value, which was derived experimentally, when calculating the distance between the peaks. Depending on the activity performed and the group, an appropriate value was selected. It would be good in the future to improve this feature of the algorithm. Instead of using the activity to select the value, data from the specific participant should be used instead. This means that the step count algorithm will automatically be participant-specific.

This study has some strengths, including the use of a range of step count algorithms that enabled us to explore in depth the advantages and disadvantages of each algorithm. Another strength of this study is the number of participants recruited. As reported in chapter 3, the majority of the studies that tested step count algorithms previously recruited fewer participants than the current study. This suggested that the results for this study are more generalisable.

A third strength for this study was the combination of two powerful techniques, DTW and DBA. These techniques have not been used before for creating a step count algorithm. The performance of the novel algorithm developed for this study proved much better for both normal and simulated-pathological conditions in comparison to the four existing literature algorithms. Finally, the algorithm developed for this study is suitable not only for normal speed walking, but for various walking speeds and for climbing stairs. The majority of the existing studies in the literature tested the step count algorithms for normal walking, and some studies tested various walking speeds. However, only a few studies tested a step count algorithm in more challenging activities such as climbing stairs.

5.5 Summary

In this study, template-matching using DTW step count algorithm was compared with four existing algorithms found in the literature. The four algorithms were selected because they covered a wide range of different step count category algorithms. The data used to test the algorithms was collected from the pilot study presented in chapter 3. The participants performed several activities under normal and simulated-pathological conditions. Again, the results of this chapter are used to answer the fourth research question posed in section 2.5.1. The template-matching using DTW algorithm demonstrated excellent accuracy when using the data from the normal condition group for slow and normal walking speeds. Regarding fast walk, stair ascent and stair descent, the results were good enough and close to the results of the template-matching algorithm. Although results of the template-matching using DTW algorithm in the simulated-pathological group were not as good as seen in the normal condition group, the new algorithm still showed the best overall performance in the simulated-pathological condition group. These results suggest that it may be possible to develop better step count algorithms for more compromised patient populations, which in turn would mean that clinicians will get more accurate and representative results.

Chapter 6

A mathematical model to generate synthetic acceleration signals

6.1 Introduction

The original plan had been to collect accelerometer data from patients, however due to COVID-19 an alternative approach was developed and this chapter is the result. In this chapter the development of a mathematical model is described. The model was inspired by a dynamic model for electrocardiogram (ECG) signal introduced by (McSharry et al. 2003). The original model is based on ECG morphology and is capable of generating realistic synthetic signals that describe the rhythm and electrical activity of the heart. For the current study a similar dynamic model for generating acceleration walking signals was developed. Development of such a model will enable researchers to generate multiple synthetic signals that simulate acceleration signals from activity monitors from healthy and impaired gaits. This should be useful for developing new gait accelerometer analysis strategies, as collection of real data from patients with impaired gait is laborious in terms of time and effort. A synthetic data approach would also facilitate data sharing since there will be no information privacy concerns, making it easier to access data (Wang et al. 2019). The models developed represented the acceleration signal of walking with normal speed for both normal and simulated-pathological conditions. The generated synthetic signals were then used as a test dataset for condition classification, while the training dataset was based on the original data collected in the pilot study described in chapter 3.

6.2 Methodology

6.2.1 Mathematical description

In 2003, McSharry and colleagues proposed a dynamic model for generating realistic synthetic ECG signal in order to assess different biomedical signal processing techniques (McSharry et al. 2003). These techniques were used to compute statistical variables from the ECG signal. The model was developed using a set of state-space equations that generates a 3D trajectory in a 3D state-space with coordinates (x, y, z) as shown in Figure 6.1. The ECG signal is described by the z-direction since it has one dimension, and the other two directions are used to control the period of the ECG signal. The foundation of this model was used to develop our dynamic model for generating realistic synthetic walking acceleration signals. The generated signal represented the vertical direction of the acceleration against time. The reason for generating the vertical direction is because it is the most informative direction in terms of walking. This is due to the up and down movement of the person while walking as shown in Figure 2.6 in the literature review.

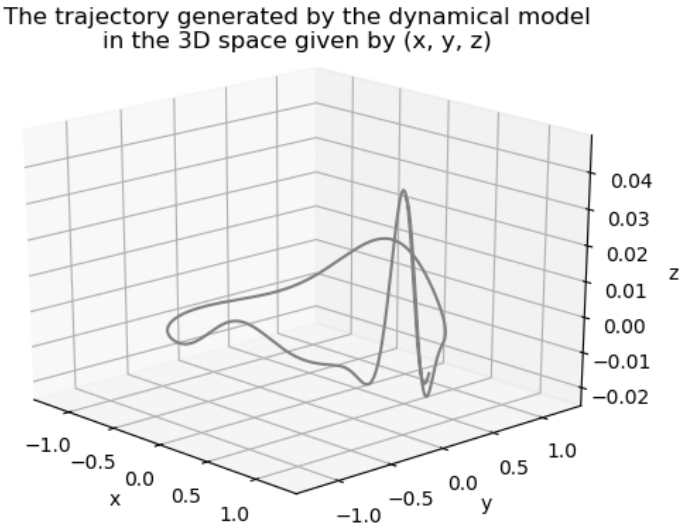


Figure 6.1: 3D trajectory in 3D state using a set of state equations

The model consists of a circular limit cycle of unit radius in the (x, y) plane around which the trajectory is pushed up and down as it approaches the four distinct areas in the acceleration as shown in Figure 6.1. Quasi-periodicity, the property of a system that displays irregular period-

icity, of the acceleration is reflected by the movement of the trajectory around the attracting limit cycle. The dynamical equations of motion are given by a set of three ordinary differential equations in Cartesian coordinates:

$$\dot{x} = \alpha x - \omega y \tag{6.1}$$

$$\dot{y} = \alpha y + \omega x \tag{6.2}$$

$$\dot{z} = -\sum_{i=1}^{n} a_i \times \Delta\theta_i \times exp(-\frac{\Delta\theta_i^2}{2c_i^2}) - (z) \tag{6.3}$$

Definitions:

$\alpha = 1 - \sqrt{x^2 + y^2}$

$\Delta\theta_i = (\theta - \theta_i) mod 2\pi$

$\theta = tan^{-1}(y, x)\ with - \pi \geq tan^{-1}(y, x) \geq \pi$

$n =$ *number of areas of interest*

ω *is the angular velocity of the trajectory as it moves around the limit cycle*

$a =$ *amplitude*

c *is the standard deviation that controls the width of a Gaussian distribution curve*

Considering the system equations in turn, equations (6.1) and (6.2) describe a periodic oscillating (i.e. repeating) signal. In this scenario, the oscillator is used to generate the circular motion of the unit circle (Stefanovska et al. 2001).

Equation (6.3) represents a sum of Gaussian functions. The reason for using Gaussian function is due to its symmetric "bell curve" shape at the centre, where half of the values are to the left and the other half are to the right. Each curve represents the area of interest for the acceleration signal (Clifford 2006). Walking acceleration signals are represented by consecutive, broadly symmetrical peaks and troughs, hence the Gaussian function was used. The Gaussian function includes three parameters that can be used to describe the curves individually; 1) the

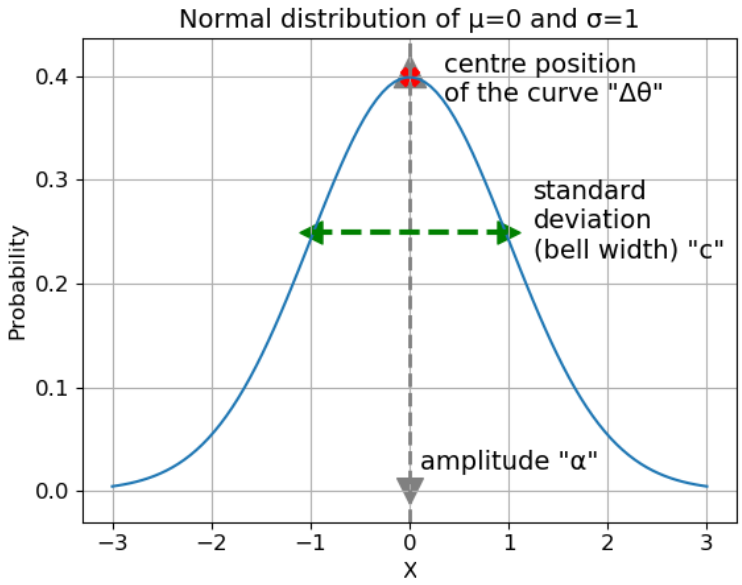


Figure 6.2: Characteristics of normal distribution

amplitude of the curve, 2) the centre position of the curve and 3) the standard deviation that controls the width of the "bell" as demonstrated in Figure 6.2.

These parameters can also be seen in equation (6.3) as a, $\Delta\theta$ and c respectively. As discussed in section 2.3.3 of the literature review, there are four areas of interest, therefore four Gaussian functions were used.

A Gaussian function has the form:

$$f(x) = \alpha e^{\frac{(x-\Delta\theta)^2}{2c^2}} \tag{6.4}$$

To model variation between gait cycles, white Gaussian noise is added to the amplitude of the signal by generating random samples from a normal distribution with zero mean and standard deviation. It is important to mention that the white noise was added only to the simulated-pathological gaits because the gait cycles had greater variability among them. Hence, Table 6.1 demonstrates the parameters used for the generation of white Gaussian noise.

Table 6.1: Gaussian noise parameters used for normal and simulated-pathological conditions.

Parameters	Normal	Simulated-pathological
Mean	0	0
Standard deviation	0	0.05-0.30

6.2.2 Model parameter estimation

An algorithm was developed in python to generate a synthetic acceleration walking signal with similar characteristics to a real signal. Figure 6.3 demonstrates the steps that were followed to produce the synthetic signal.

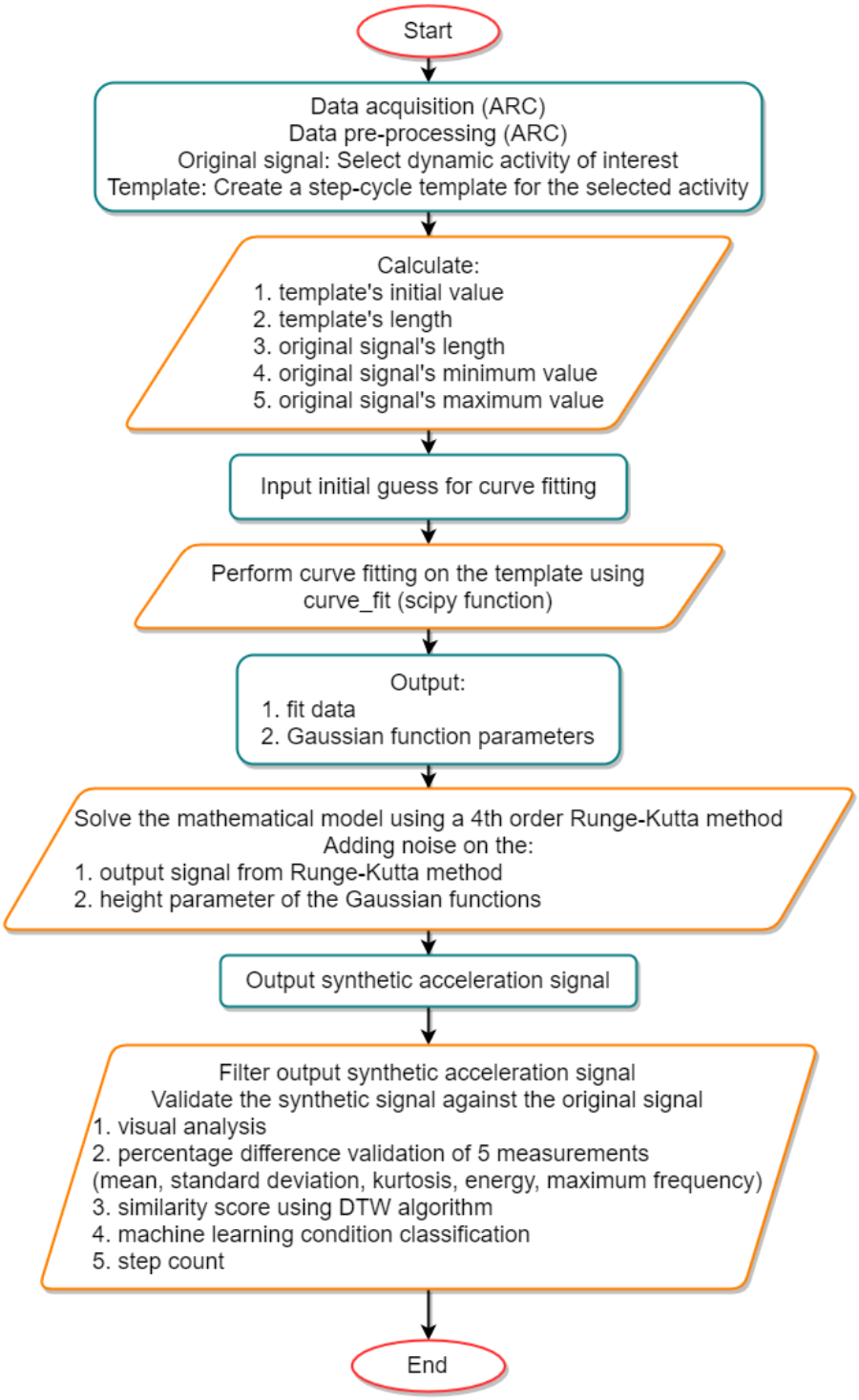


Figure 6.3: Flowchart describing the process of generating a synthetic acceleration signal

Initially, a subset of the data from chapter 5 that only contained periods of dynamic activities was created, in which participants were walking normally and simulated-pathologically with a normal walking speed. Based on the original acceleration signal, three parameters were calculated: 1) the entire length of the walking signal, 2) the maximum acceleration value of the signal, and 3) the minimum acceleration value of the signal.

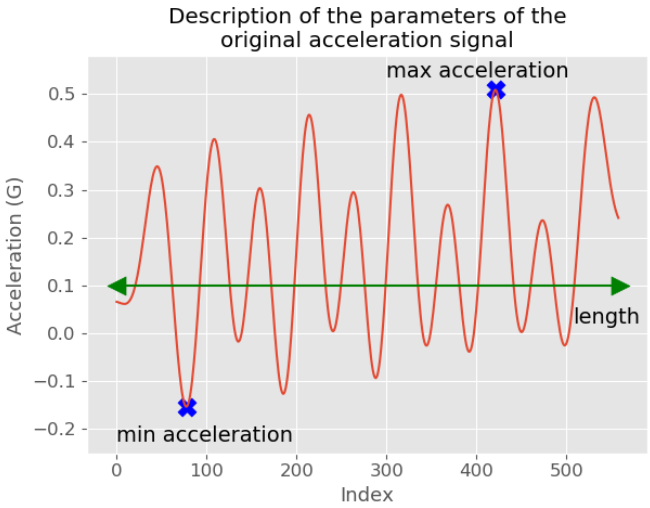


Figure 6.4: Characteristics of the parameters measured in the original acceleration signal

After this, a template was produced representing one gait cycle using the dynamic time warping barycentre averaging method proposed in chapter 5. Another two parameters were calculated, which are associated with the template; 1) initial acceleration value and 2) the entire length of the template, were also calculated as demonstrated in Figure 6.5.

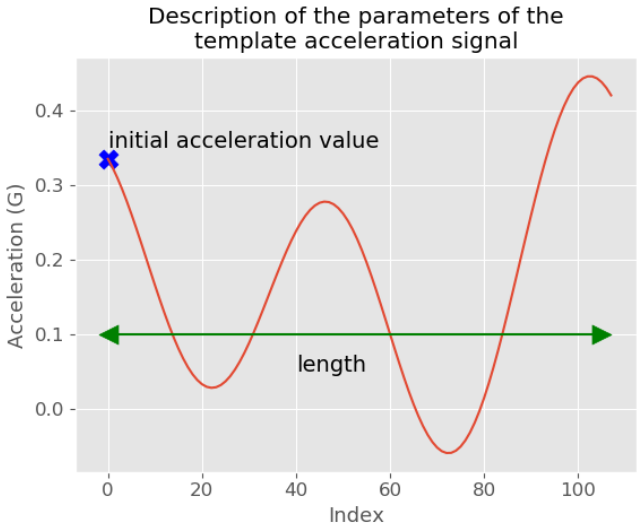


Figure 6.5: Characteristics of the parameters measured in the single gait cycle template acceleration signal

Next, curve fitting was performed on the template signal in order to calculate the three essential parameters that are used in equation (6.3). These parameters are: 1) the amplitude, 2) the position of the centre of the peak, and 3) the standard deviation that controls the width of the peak. In order to calculate these parameters, the first step is to set appropriate initial conditions to perform the curve fitting. If no initial conditions are set, the parameter optimisation settles in local minima, leading to a poor template match. Hence, to avoid this problem, the initial estimate should place each Gaussian at roughly the right location and then the optimisation can find the exact location. The initial conditions were estimated visually based on the amplitude and position of the peaks of the gait cycle template. For the curve fitting, equations (6.1)-(6.3) were used to fit the gait cycle template model.

When the parameters, amplitude, centre position of the peak and standard deviation that controls the width of the peak were calculated, they were used to numerically integrate equations (6.1)-(6.3) using a 4th order Runge-Kutta method. In order to create a similar acceleration signal to the original, some of the parameters associated with the template and original acceleration signals were also used. For example, the initial amplitude acceleration value of the template was used as initial condition, which is essential to enable the integration. Additionally, the three parameters associated with the original acceleration signal were also used in order to generate a synthetic signal that is similar with the original signal in terms of maximum and minimum acceleration value, as well as the length of the signal.

Table 6.2 demonstrates the parameters calculated from the Gaussian function.

Table 6.2: Gaussian function parameters for normal and simulated-pathological conditions.

	Index(i)	Peak 1	Peak 2	Peak 3	Peak 4
Healthy					
	α_i	-8.1085	-6.0593	-11.1617	-3.4723
Participant A	θ_i	36.2864	-118.9687	87.3911	183.1252
	c_i	0.5303	1.0831	0.5909	0.8357
	α_i	1.0227	-6.8367	3.5758	4.6952
Participant B	θ_i	32.22	4.6672	96.9272	114.0162
	c_i	3.3918	0.7658	0.1592	-0.1266
	α_i	8.2461	10.7214	5.1627	4.9345
Participant C	θ_i	21.2639	67.3084	109.5181	246.471
	c_i	0.4658	-0.4913	0.3502	-0.6197
Simulated-pathological					
	α_i	0.4236	0.2936	1.2551	0.4934
Participant A	θ_i	-41.1673	180.194	98.2208	191.8672
	c_i	0.6617	2.77	-0.0007	0.6781
	α_i	4.1021	1.3508	0.9692	4.2263
Participant B	θ_i	-49.8088	106.1262	150.0181	110.2147
	c_i	0.5865	-0.0004	1.7954	0.3665
	α_i	0.1100	0.2127	0.1792	0.4171
Participant C	θ_i	-62.186	-3.214	174.6702	192.1822
	c_i	1.3449	0.5569	2.1935	0.3793

The output signal, after the integration, was then smoothed using Savitzky–Golay filter to increase the data precision by keeping the tendency of the signal.

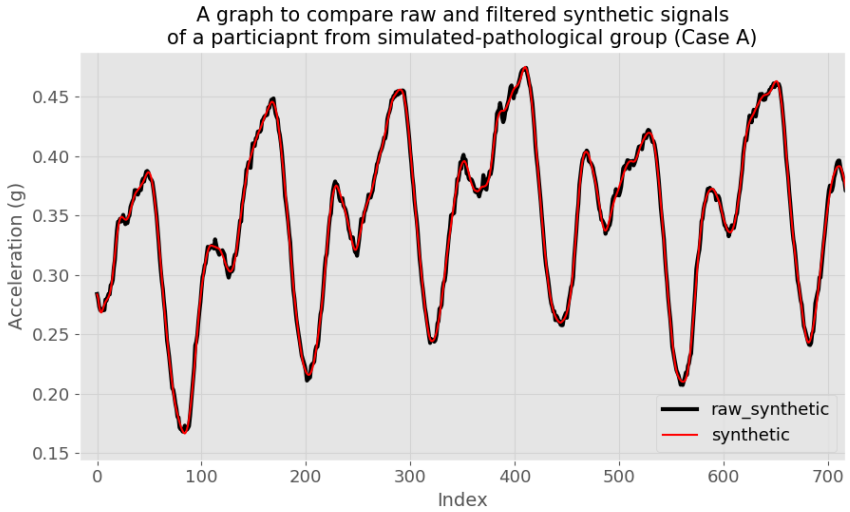


Figure 6.6: Graphs demonstrating the difference between a raw and a filtered signal

The last step was the validation of the synthetic signal against the original signal. The particular step was described in the following section.

6.2.3 Validation of the models

The synthetic signals generated using the mathematical system were validated against the original acceleration signals. Different validation types were performed to ensure the validity of the synthetic signals; 1) visual, 2) percentage difference of signal metrics 3) DTW similarity, 4) machine learning classification and 5) step count.

6.2.3.1 Visual

For visual inspection, the original signal and the generated synthetic signal were overlaid in a graph to compare them visually as shown in Figure 6.8 for normal and Figure 6.11 for simulated-pathological conditions. The visual comparison helped to check whether the synthetic signal had similar morphology to the original signal.

6.2.3.2 Performance metrics

Five signal metrics: 1) mean, 2) standard deviation, 3) kurtosis, 4) energy and 5) dominant frequency were calculated for both synthetic and original signals. The percentage error for each metric was then calculated. These five metrics were calculated because they provide a range of signal characteristics that can be used to compare the synthetic signal with the original one.

Table 6.3: Signal metrics used to compare the original and synthetic acceleration signals.

Features	Definition
$Time-domain$	
Mean	The average value of an entire signal
Standard deviation	The variability of a signal from the mean
Kurtosis	The distribution shape of a signal relative to Gaussian distribution
$Frequency-domain$	
Energy	The strength of a signal
Dominant frequency	The highest magnitude of the sinusoidal component

6.2.3.3 DTW similarity

DTW was used to quantify the similarity between the original and synthetic signals. This method enabled checking of the similarity of the signals even though they might be out of

phase, but following the same underlying pattern. In terms of the DTW similarity score, the closer it is to zero, the more the signals were considered to match. Additionally, a range of DTW similarity scores were also demonstrated among different signals in order to have a reference point.

6.2.3.4 Machine learning condition classification

An Activity-Recognition-Chain (ARC) process was followed to complete this section. The data used in for the condition classification was the acceleration signal in x-direction, which represented the vertical axis of the accelerometer. The labels "0" and "1" represented normal and simulated-pathological conditions respectively. This validation method was performed to check whether the classification performance is similar on the synthetic and real data.

Following the ARC process, the acceleration signals were segmented into different windows. For each window, 14 features (see Table 6.4) were calculated and then scaled. The next step was to perform PCA for feature reduction, in order to generalise the model.

Table 6.4: Features used for the machine learning classification for both original and synthetic acceleration signals.

Time-domain		Frequency-domain
Mean (x)	Standard deviation (x)	Energy (x)
Median (x)	Skewness (x)	Max frequency 1(x)
Kurtosis (x)	Interquartile range (x)	Max frequency 2(x)
Root mean square (x)	Median absolute (x)	Mean frequency (x)
Mean power spectral density (x)		Entropy (x)

For the last step, which was the actual classification, a Support Vector Machine algorithm was used. The training data was based on the original acceleration signals. The test data was represented by the synthetic acceleration signals and the scenario conducted for the condition classification was:

- algorithm trained on original combined normal and simulated-pathological datasets; tested on synthetic combined normal and simulated-pathological datasets

6.2.3.5 Step count

For the step count validation, synthetic signals of similar length to their associated original signal were generated. The calculated number of steps based on the original dataset using the

algorithm in chapter 5 is known. Therefore, the algorithm was used to also calculate the number of steps measured using the synthetic signals. The calculated number of steps from the synthetic dataset was compared to the calculated number of steps from the original dataset by calculating their difference in terms of number of steps. If the difference was low, it suggested that the signals were almost identical. Hence, calculating a similar value for the predicted number of steps.

6.3 Results

The particular participants analysed were selected since they represent most of the data of normal walking activity in the normal and the simulated-pathological conditions.

6.3.1 Validation

6.3.1.1 Normal condition

Figure 6.7 demonstrated the results from the dynamic model parameter optimisation. The data used was the averaged gait cycle calculated for each participant. The fitted line was essential to calculate the parameters for the areas of interest. For participant C, the fitting had matched the data with great precision. For both participants A and B, the fitting was of high precision, however at two points the fitting was not as smooth as the gait cycle data. The similarity between the template and the fitting was calculated using the DTW technique. The results were: 0.051, 0.077 and 0.028 for participants A, B and C respectively. The closer the result is to zero, the highest the matching is between the template and the fitting.

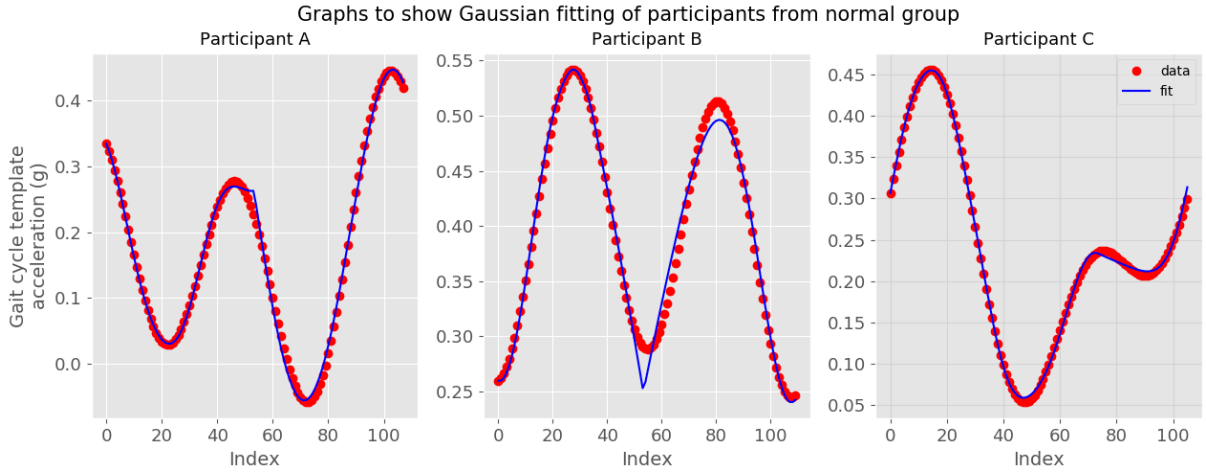


Figure 6.7: Example of Gaussian fitting for a gait cycle during normal walking under normal condition (Participants A, B and C)

Regarding the visual validation, all three synthetic signals seemed to follow the underlying pattern of their original signal as demonstrated in Figure 6.8. However, the synthetic signals were slightly out of phase with the original signal. DTW was used to examine the similarity of the compared signals even though they might be out of phase. The calculated DTW score for participants A, B and C were 1.031, 0.830 and 1.414 respectively.

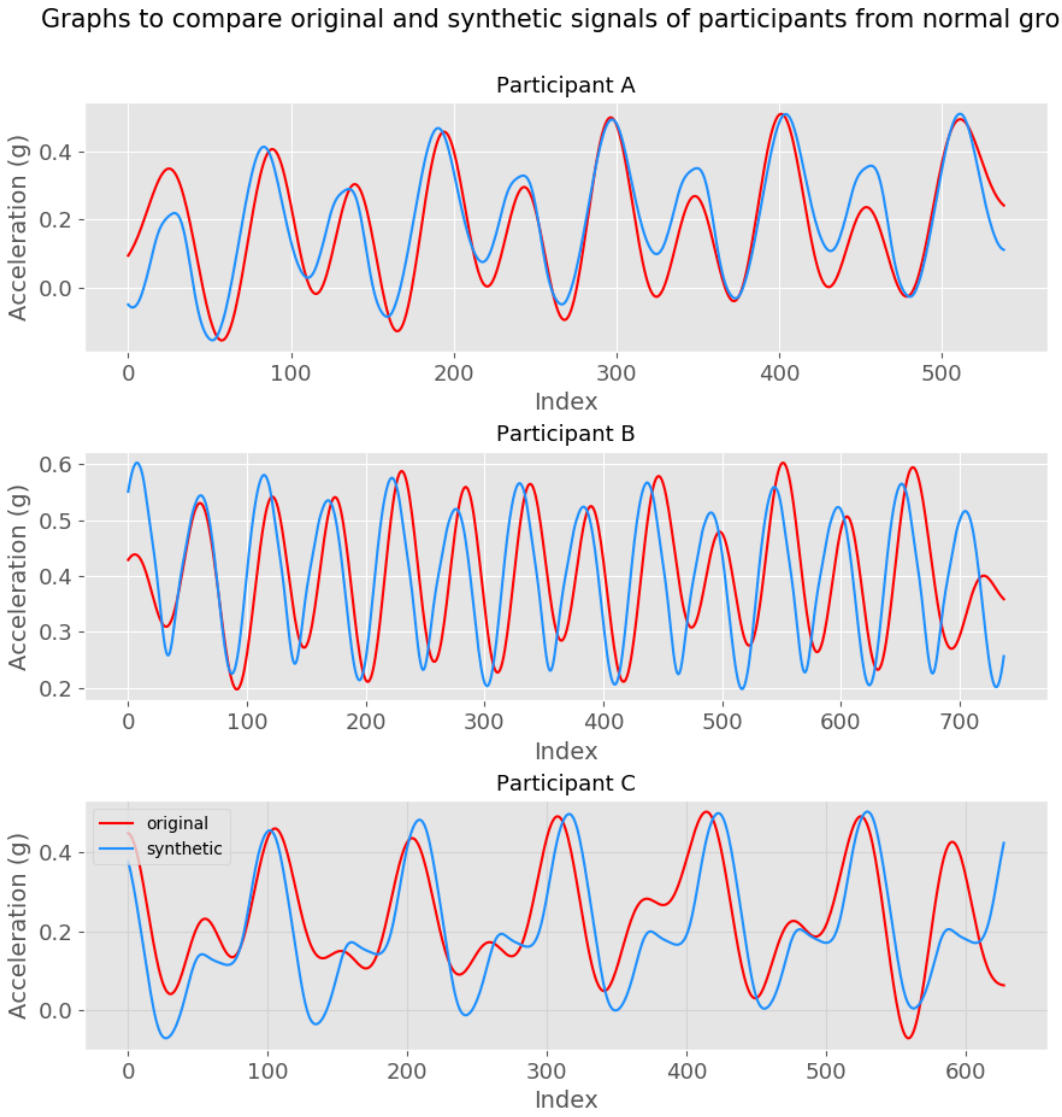


Figure 6.8: Comparison of normal walking original and synthetic acceleration signals under normal condition (Participants A, B and C)

In order to check whether the DTW similarity scores were acceptable, the original and synthetic data from the other two participants was compared and a range of DTW similarity score was therefore developed. This range contained poor similarity scores, since the signals tested were from a different participant, and hence did not match well. Figure 6.9 demonstrate the results

for the comparisons and table 6.5 shows the DTW similarity scores.

Table 6.5: Results of DTW similarity between original and synthetic acceleration signals under normal condition.

Participants		DTW similarity
Original	Synthetic	
A	B	4.254
A	C	1.752
B	C	3.350
B	A	3.952
C	A	1.838
C	B	2.904

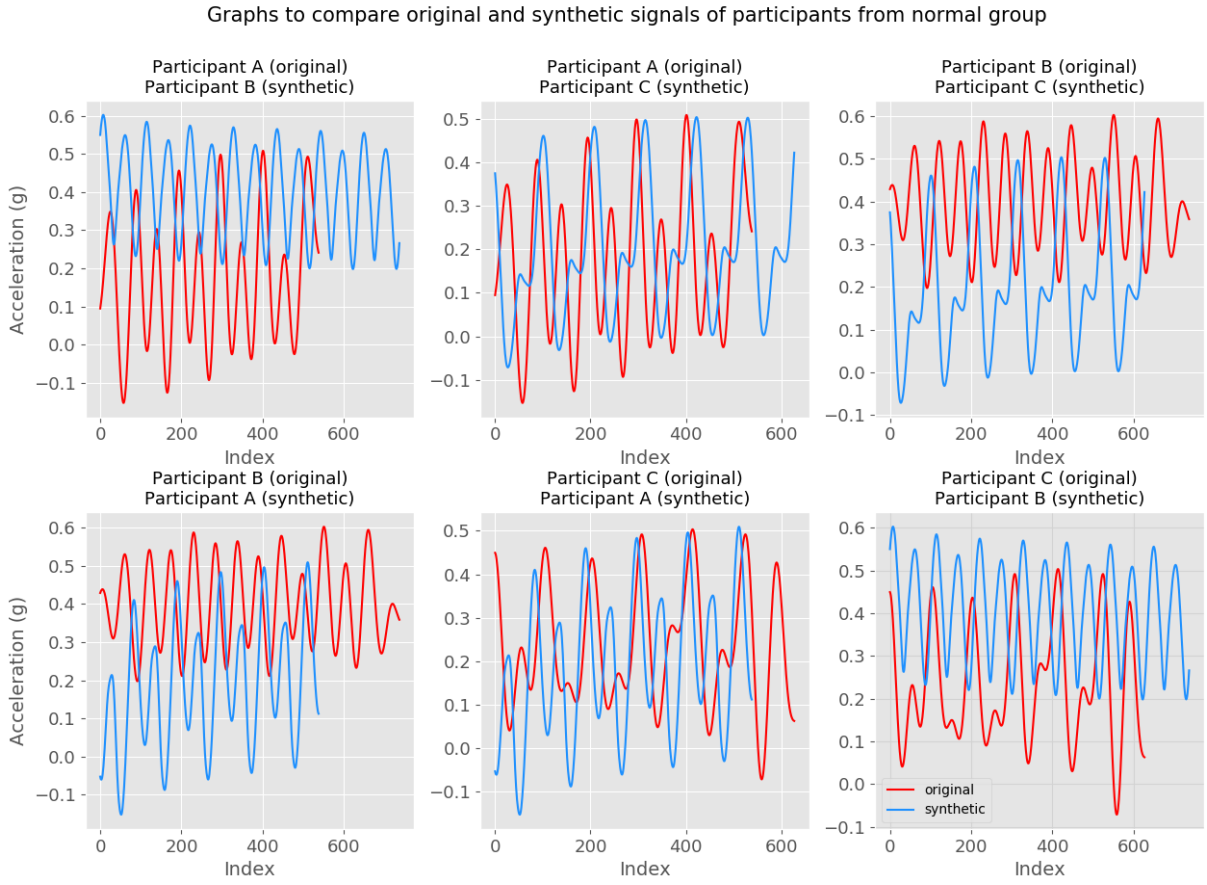


Figure 6.9: Comparison of normal walking original and synthetic acceleration signals under normal condition between two participants

The DTW similarity scores were smaller when the original and synthetic signals were from the same participant (< 1.414) rather than when comparing signals from different participants (> 1.838), meaning that the signals match quite well.

The following results represented the percentage difference between the synthetic and original

signals. As the DTW score also suggested, the synthetic signal based on participant B had the lowest percentage difference for almost all measurements among the three participants as demonstrated in table 6.6.

Table 6.6: Percentage difference of signal metrics to compare original and synthetic acceleration signals under normal condition.

Percentage difference (%)	Participant A	Participant B	Participant C
Mean	8	2	14
Standard deviation	3	4	9
Kurtosis	-5	-14	-15
Energy	12	8	20
Dominant frequency	19	4	15

6.3.1.2 Simulated-pathological condition

Similarly to the normal condition, curve fitting was performed at the averaged gait cycle for the simulated-pathological condition. From Figure 6.10 below, it was demonstrated that the averaged gait cycle of the simulated-pathological condition exhibited greater complexity than the results from the normal condition. The similarity between the template and the fitting was calculated using the DTW technique. The results were: 0.036, 0.233 and 0.002 for participants A, B and C respectively.

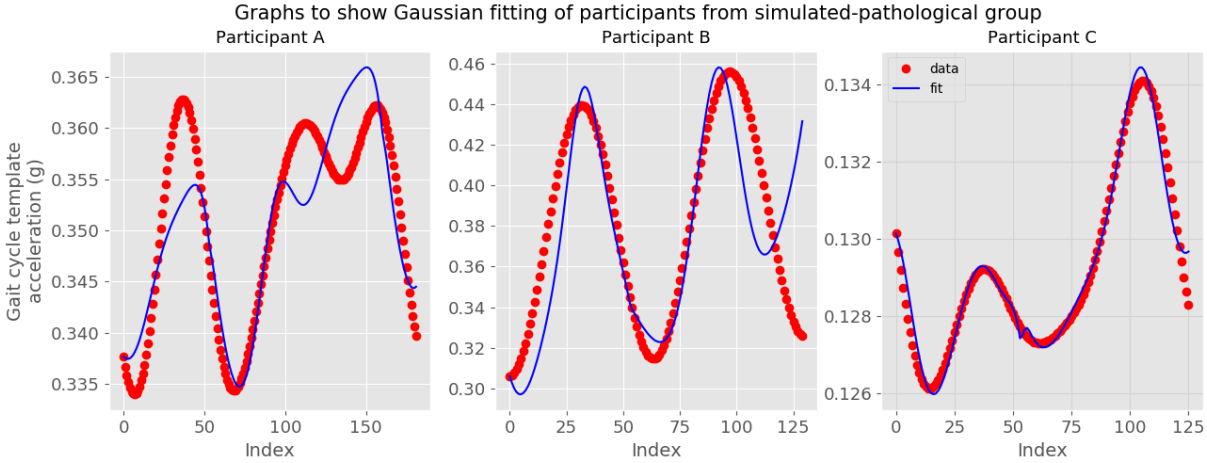


Figure 6.10: Example of Gaussian fitting for a gait cycle during normal walking under simulated-pathological condition (Participants A, B and C)

The synthetic signals of the simulated-pathological condition followed the underlying pattern of their original signal. The underlying pattern was based on the template created. The acceleration signals that fall in the simulated-pathological condition were more complex in

terms of morphology than the signals of the normal condition. Additionally, they did not have very consistent pattern between steps. However, it could be seen in Figure 6.11 that they match at some occasions. The signals were also out of phase with DTW scores 1.615, 1.222 and 0.794 for participants A, B and C respectively.

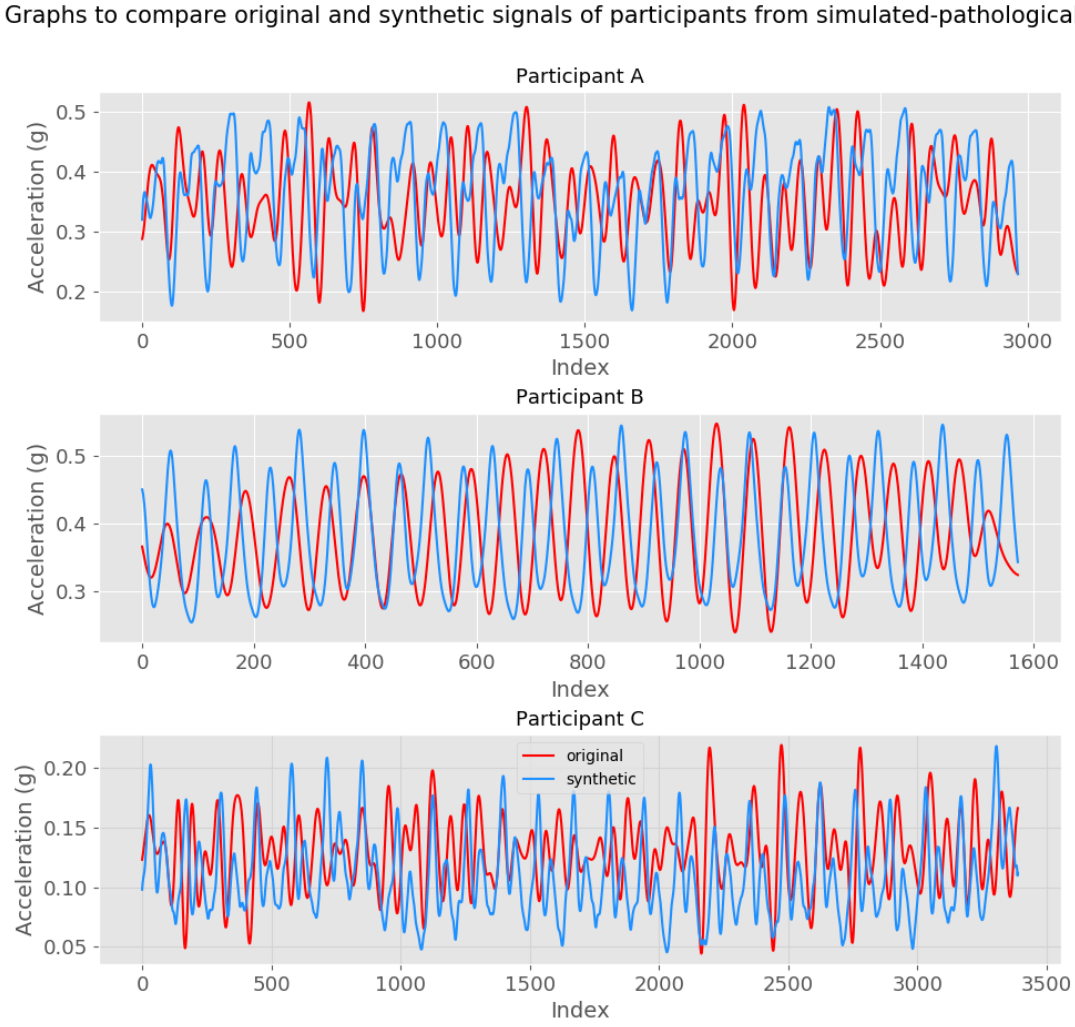


Figure 6.11: Comparison of normal walking original and synthetic acceleration signals under simulated-pathological condition (Participants A, B and C)

Similar to what was done with the data from the normal group, in order to check whether the DTW similarity scores were acceptable for the simulated-pathological group, the original and synthetic data from the other two participants was compared. Figure 6.12 demonstrates the results for the comparisons and Table 6.7 shows the DTW similarity scores.

Table 6.7: Results of DTW similarity between original and synthetic acceleration signals under simulated-pathological condition.

Participants		DTW similarity
Original	Synthetic	
A	B	2.037
A	C	9.522
B	C	10.677
B	A	1.662
C	A	10.269
C	B	10.599

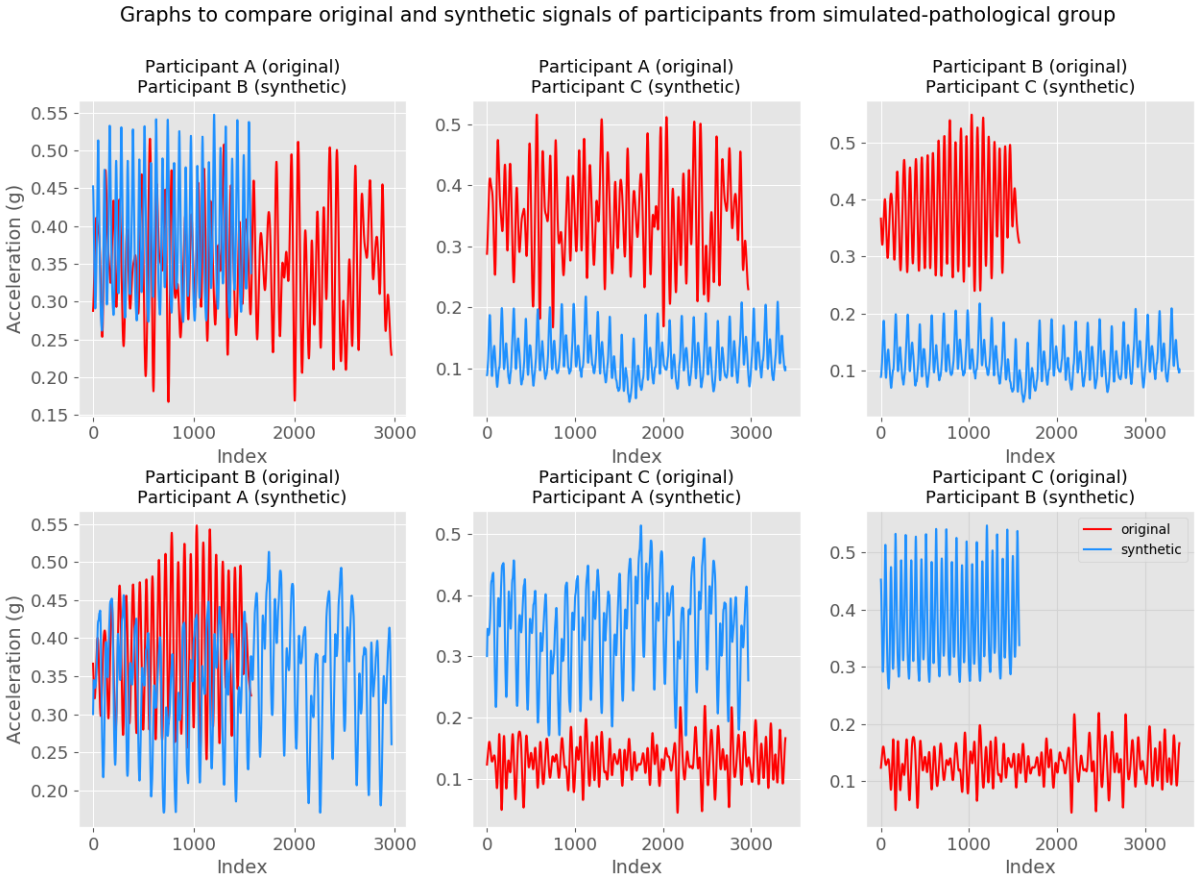


Figure 6.12: Comparison of normal walking original and synthetic acceleration signals under simulated-pathological condition between two participants

Similar to the normal group, the results of the DTW similarity scores seem to match well for participants A, B and C.

From the visual comparison and the DTW scores, participant C demonstrated greater matching between the original and synthetic signals in comparison to participants A and B. However, based solely on the percentage differences for each participant, participant B had the best

performance. Participants A and C had large percentage difference for energy and kurtosis respectively.

Table 6.8: Percentage difference of signal metrics to compare original and synthetic acceleration signals under simulated-pathological condition.

Percentage difference (%)	Participant A	Participant B	Participant C
Mean	5	1	10
Standard deviation	9	4	10
Kurtosis	-24	-6	129
Energy	32	26	26
Dominant frequency	9	4	12

6.3.1.3 Condition classification

An SVM classifier was used to classify the two conditions. The algorithm was trained from the original (real) data collected for the pilot study which includes both normal and simulated-pathological conditions, and the test data used was the synthetic (unseen) signals generated for both normal and simulated-pathological conditions. The model was trained using the features shown in Table 6.4. This classification was performed to check whether the data from each condition can be differentiated. High performance scores were achieved for this classification, with the precision score to be the highest as demonstrated in table 6.9.

Table 6.9: Performance metrics of condition classification to classify normal and simulated-pathological conditions accurately.

Performance metrics	Support Vector Machine
Accuracy	0.704
F1-score	0.740
Precision	0.802
Recall	0.704

6.3.1.4 Step count

Another way to test the generated acceleration signals was to use the new step count algorithm developed in chapter 5. If the generated synthetic signals are similar to the original signals, the algorithm should provide similar answers for the predicted number of steps of the original and synthetic datasets. The results for the normal group agree with all the validation results of this condition. On the other hand, the results for the simulated-pathological group exhibited greater variation in the results as demonstrated in table 6.10. This might be because the

generated signals of the simulated-pathological group did not match completely with the signals of the original dataset. Hence, a difference of 5 to 20 steps was observed between the original (predicted) signal and the synthetic (predicted) signal.

Table 6.10: Results for counting the number of steps using Template-matching using DTW algorithm.

	Normal			Simulated-pathological		
	Case A	Case B	Case C	Case A	Case B	Case C
Original (predicted) signal	10	13	12	31	24	51
Synthetic (predicted) signal	10	14	12	44	19	31

6.4 Discussion

This chapter presents the results exploring the generation of synthetic acceleration data representing normal and pathological gaits. The results presented here answer the final research question posed in section 2.5.1; Can we accurately generate synthetic acceleration data that represent normal and atypical walking patterns?

Earlier studies used three different approaches to generate synthetic data. The three approaches were: (1) mathematical model using coupled equations (McSharry et al. 2003; Santaniello et al. 2006; Almasi et al. 2011; Racic and Morin 2014), (2) mathematical model using Lagrange approach (Al-zu et al. 2012; Agarana and Akinlabi 2018) and (3) Generative Adversarial Networks (GANs) (Alzantot et al. 2017; Hassouni et al. 2018). From these approaches, only the third approach was used to generate acceleration signal that represented daily activities, such as walking, sitting, standing, etc. For this chapter, the first method was used to generate the synthetic acceleration signals due to the fact that the second method is based on kinematics and hence did not provide acceleration information, and the third method required a large amount of data to enable the appropriate generation of synthetic signals. Additionally, the first method has been used to generate different types of signals, for example ECG, phonocardiogram, jumping signals. This means that the method is flexible enough to be applied to several situations and generate the required signal. Also, until the current study this method has not been used to generate acceleration signals from walking. The majority of studies noted above have used only a visual comparison to check the match of the synthetic signals with the original signals. Hassouni and colleagues used the GANs approach to generate acceleration signal of walking, jogging, standing, sitting, stair ascent and stair descent and an accuracy score of 97.33% was reported for

classifying these activities (Hassouni et al. 2018). The second approach, Lagrange approach, was used to generate synthetic data to model the arm movement, however, the synthetic data was angular velocity and angular displacement (Al-zu et al. 2012; Agarana and Akinlabi 2018). For the current study, the acceleration signal was of interest rather than angular velocity and angular displacement.

There was no definitive way to test for validity of the model developed, as there was no gold standard. The following steps were taken however: (1) visual assessment and comparison with real-world data, and (2) checking whether the simulated gaits (normal and pathological) were correctly classified by the classifier trained on real data.

The data demonstrated that the signals associated with the normal population achieved better performance scores in terms of matching with the original data in comparison to the signals associated with the simulated-pathological population. This might be due to the fact that simulated-pathological signals had greater noise than the signals from the normal population. In general, it is simpler to generate periodic signals than signals with random peaks due to greater noise. Based on the results, DTW score showed that the original signals were similar to the synthetic data in terms of their overall pattern. The synthetic signals might followed a similar but out of phase pattern with the original signals. This is mostly supported from the five performance metrics which have a percentage difference up to 20% and 129% for the normal and simulated-pathological groups respectively. The synthetic signals were also tested using condition classification, and a similar scenario was tested as the one mentioned in chapter 3. The condition classification achieved scores around 74%, and the reason might be that the selected normal and simulated-pathological signals had some differences among them, which means that the classifier could not differentiate the signals of the two conditions with excellent accuracy. The signals were also tested using the step count algorithm. The predicted number of steps using the original and the synthetic signals should be the same if the two signals match because the same step count algorithm is used for both. If the predicted number of steps of the original and synthetic signals did not match, this means that the two signals were different. For the normal group, participant B had one step difference from the results from the predicted original and synthetic signals. Additionally, participants A and C found the exact number of steps in both signals. Unfortunately, that was not the situation for the simulated-pathological group. Cases A, B and C had -13, 5 and 20 steps difference from the results of the two signals.

This suggested that the approach for generating the synthetic signal worked well for the normal group, but it needs some refinement for the simulated-pathological group.

A limitation observed was that the model might be basic since the overall shape of the signal was taken into consideration in order to create the template for a single gait cycle. This means that some irregularities of the template signal, especially for the simulated-pathological group, were not generated since they were lost, for example where an average of all the walking cycles for the normal walking was calculated. The model could generate acceleration signals for the normal group well, but was not able to generate the acceleration signals for the simulated-pathological group with similar precision. This might be because the signals of the simulated-pathological group had greater variability. Hence, a more complex model might be essential to be developed to ensure better pathological synthetic signals are developed. This model will consider greater variation of parameters that will enable the generation of signals with more irregularities.

This chapter has some strengths as well. The acceleration walking signals have been generated because there is a need for more acceleration data. There is not enough publicly available data from wrist accelerometers, and especially from pathological populations. This method enables us to generate synthetic acceleration signals from activity monitors for both healthy and pathological populations. Additionally, it is a quite simple but powerful technique, which can be applied to different dynamic activities, such as slow and fast walk, and climbing stairs where only the variables used will change, while the underlying mathematical approach will be the same. Lastly, this method was new for the particular sector, however the results were promising for the generation of acceleration signals more widely.

6.5 Summary

In this chapter, several synthetic acceleration walking signals were generated using three coupled equations. This is beneficial for the research community since there is limited availability of public datasets that includes accelerometer data collected from the wrist. The results of this chapter were used to answer the fifth question posed in section 2.5.1. This method can be used to generate walking signals for people who walk normally but is not so accurate for people with walking impairments. It is based on the morphology of the signal, therefore several walking speeds can be generated as well. The results are positive and promising for the accurate generation of signals for normal gaits, however, the algorithm needs refinement in order to generate

representative signals for simulated-pathological gaits as well. The source code of the algorithm is available in a github repository `https://github.com/ValeriaF22/Thesis-Project`.

Chapter 7

Discussion & Conclusion

This work aimed to create and develop novel algorithms that could be integrated into wearable devices with the ambition of successfully monitoring the activity of people with chronic conditions. After reviewing the literature, it was noted that there was insufficient research for activity classification and step count algorithms which target the patient population with different pathologies, especially walking impairments.

Specifically, this thesis aimed to: 1) develop tuneable algorithms to more accurately measure physical activity in people with walking impairments and 2) generate synthetic acceleration walking signals that represent normal and atypical gait patterns, to be used as a potential dataset.

To achieve these objectives, a pilot study was initially developed using healthy individuals to investigate whether, under normal and abnormal conditions, (1) the proposed data collection method would be suitable for real patients; (2) machine learning methods could identify individual activities; (3) machine learning methods could be used to classify normal and abnormal walking; (4) accelerometer signals could be processed to accurately estimate number of steps. In this study, 30 healthy volunteers were recruited and asked to perform nine different activities in a laboratory setting while wearing a wrist- and an ankle-worn accelerometer. They performed the activities as normal, and then repeated them while emulating a pathological gait (simulated-pathological).

In this study two accelerometers were attached on each participant, one on the wrist and another one on the ankle. These two locations are very popular since 1) the wrist location

can be used for many applications, and they can provide vital information quickly without developing any extra burden to the user (Al-Eidan et al. 2018), 2) the ankle location is often used to solve problems related to gait. To answer the first objective of the study which was related to the wear site, both quantitative and qualitative approaches were performed. The qualitative approach involved interviewing stakeholders, for example clinicians and patients (see Appendix A). It is important to understand the needs of the stakeholders since they will be the end users (clinicians and patients). It is essential to develop algorithms that are useful and helpful to them. The quantitative approach was a preliminary study using machine learning algorithms to study a few cases for both condition and activity classifications at both wear sites. Condition classification is used to distinguish between different groups of people, such as normal and simulated-pathological. Activity-type and -task classifications are used to distinguish three general types of physical activity and nine specific tasks of physical activity respectively. The outcomes of the "Patient and Public Involvement discussions" suggested that patients are willing to wear an activity monitor only on their wrist, and they did not want to wear the device on their ankle. Hence, the quantitative study was performed to check objectively which wear site is the best for the classifications. The results suggested that the wrist mounted location can be used to differentiate the user's condition i.e. healthy or (simulated) pathological condition. After identifying the condition of the user, the most appropriate algorithms could be used depending on the condition to ensure best performance. Additionally, the results related to activity recognition showed acceptable performance for both normal and simulated-pathological conditions and this was observed for the wrist and ankle locations. Due to these outcomes and also the results of the patient interviews from the stakeholder analysis, the wrist location was used for subsequent detailed analysis.

The detailed analysis of the wrist location to inform condition classification was used to answer the second objective of this study, which was related to the performance of machine learning algorithms in identifying whether a patient is moving normally. The outcomes suggested that the two conditions, normal and simulated-pathological, could be classified with high accuracy. This might be because of the differences in the morphology of the accelerometer signals for each condition. For example, the participants tended to walk slower under the simulated-pathological condition in comparison to the normal condition. Therefore, the signals will differ in terms of their amplitude and period, and this might influence the values of the features calculated. Two distinct clusters of features might be created and the two conditions would be discriminated

with high accuracy (Mannini et al. 2016). This might be a step towards personalised healthcare, since it will enable the development of algorithms that could target solely the group of interest.

The detailed analysis of the wrist location about activity classification was used to answer the third objective of this study, which was related to the performance of machine learning algorithms in identifying different activities under normal and simulated-pathological conditions. In general, the outcomes suggested that machine learning methods for activity classification were more accurate for the healthy group than for the simulated-pathological group. For the activity-type classification, the difference between the two results of the normal and simulated-pathological conditions was low (1.7%). This suggested that activity-type classification can be used to distinguish different states of physical activity, such as static and dynamic. According to the clinicians stakeholders distinguishing between sedentary and active states is of great importance, therefore this will be used as a feature of the desired system. Additionally, these results confirmed that it is not only possible to develop algorithms targeting a specific group of people, but also better classifications can be achieved. As demonstrated in chapter 3, the results were improved when the algorithm was trained and tested with similar data. Based on the activity-type classification results, it is assumed that clinicians will be able to distinguish objectively whether their patients were mainly active or sedentary throughout a long period of time. This will enable the clinicians to understand how much effective their treatment was, how to continue the treatment plan and advise better the patients on what to do regarding their physical activity.

For the activity-task classification, the difference between the two results of the normal and simulated-pathological conditions was higher (9.8%) than the results of activity-type classification (1.7%). This is because it becomes harder to correctly classify each individual activity. For example, there are three walking activities (slow, normal, fast) that share similar signal shape but differ in some signal characteristics. This might become even harder when the person-to-person variability is considered. The activity-task classification of the normal condition achieved higher (0.943%) performance than the classification of the simulated-pathological condition (0.845). As demonstrated in chapter 4 , to achieve better step count results is important to be able to classify each individual activity. Therefore, the step count results will be negatively influenced if the activity-task classification is not accurate enough. This suggests there is a need for improvement for activity-task classification, especially for the pathological

conditions. Some of the improvements could be to collect greater amount of data since the machine learning algorithms are dependent on the amount of input data. Additionally, a more complex machine learning algorithm will be developed in order to achieve better outcomes. For example, if a Neural Network algorithm is used, instead of having just one layer more layers would be used. In this study, only one layer was used because of the amount of data.

In machine learning a key to high performance is data. For example, the nature of data, the amount of data and the calculation of the appropriate features to maximise the discrimination of those features. From the above examples it was demonstrated that 1) the normal group achieved higher performance than the simulated-pathological group in terms of activity-type and -task classifications, 2) in both groups activity-type classification achieved better scores than the activity-task classification. One of the reasons, for the normal group to achieve better outcomes than the simulated-pathological group, might be the combination of the acceleration signal with the calculation of the features. For example, the features representing the normal group might had greater differences among them, hence better performance was achieved. On the other hand, the features of the simulated-pathological group were closer together in some cases, and this might have confused the algorithm to make the wrong prediction. To address that, greater amount of data is essential. A similar reason might be responsible for the better performance of activity-type classification in comparison to activity-task classification. As mentioned previously, the three walking activities will have features very similar to each other since the shape of the signal is similar. Additionally, the variability of each person was considered since supervised machine learning algorithms were used. In other words, the slow walk of one person might be similar to the normal walk of another person. The supervised machine learning algorithm uses the labels set to each chunk of data to calculate the accuracy and other performance metrics. Therefore, the value of the feature might be similar, however the label might be different. This might cause a confusion to the algorithm, and hence reduced performance is achieved.

Of the five machine learning algorithms used for the activity classification, three were markedly better. In all cases, the top three classifiers were Support Vector Machine, k-Nearest Neighbour and Neural Network. Gaussian Naïve Bayes always had the worst performance, and Random Forest always had the second worse performance. This was true when the algorithms were trained on healthy data and tested on healthy data and when they were trained on simulated-

pathological data and tested on simulated-pathological data. Regarding the Gaussian Naïve Bayes, the reason for the poorer performance of this classifier might be due to the nature of the dataset. Gaussian Naïve Bayes is named after the Gaussian distributions that represent the dataset in the training dataset. Therefore, if our dataset does not follow the Gaussian distributions, this means that the classifier will not perform at its best performance. The Gaussian Naive Bayes algorithm was used as a baseline algorithm since no alterations were made in its parameters. In general, based on how each algorithm works and the type of data used, some machine learning algorithms are more suitable than the others. Therefore, this study suggested that k-Nearest Neighbour, Support Vector Machine and Neural Network are suitable for condition and activity classifications using accelerometer data.

The classification of specific tasks was useful in the context of step count, which is also considered to be an important physical activity metric (Bassett et al. 2017). This is because steps are objective, and they can easily be translated from scientific results into simple outcomes that lay audience could understand. Additionally, steps can also be used to distinguish whether someone is active or not, since being active might mean greater amount of steps (Bassett et al. 2017). In this context, activity classification enables to filter out irrelevant data before trying to calculate steps. This was taken into consideration, and hence tested in chapter 4. In this chapter, the signal processing and machine learning methods showed that they could be used to estimate step count for normal and simulated-abnormal gait at a range of walking speeds, but that step count estimates were poor for abnormal gait.

Multiple existing approaches were investigated to estimate step count from accelerometer data. Many algorithms work by firstly identifying gait period using a threshold. Following this, methods such as peak detection or template-matching are often applied to count the number of steps. Four standard approaches from the literature were compared against a new step count algorithm (Palshikar 2009; Thanh et al. 2017; Dirican and Aksoy 2017; Micó-Amigo et al. 2016). The results from the four standard approaches suggested that they achieved better performance for the normal condition rather than the simulated-pathological condition. This has to do with the nature of the acceleration signals. For example, the signals for both conditions are dynamic, however the signals of the healthy group are simple and mainly periodic. The signals of the simulated-pathological group are more complex and mainly periodic, however some participants performed more sporadic movements. This was demonstrated by the results of the confusion

matrices in chapter 3 where there were larger errors in the simulated-pathological condition in comparison to the normal condition. Additionally, it was also demonstrated in chapter 5 because the template representing a single step was more representative of the signal in normal condition in comparison to the simulated-pathological condition. The template was created based on the periodicity of the acceleration signal since autocorrelation was used. The shape and characteristics of the signal enables the calculation of step count results with higher accuracy for the normal condition. It becomes harder to discriminate the correct number of steps when the acceleration signal is irregular. Therefore, it might be beneficial to develop algorithms that are tailored to specific group of people, hence the algorithm can be developed with greater complexity and with specific parameters. Then, people with walking impairments can count the steps undertaken with higher accuracy, which enables clinicians to take better decisions about their patients. Additionally, although the results in general were acceptable for the normal condition, most of the algorithms did not do so well for the slow walking activity.

Based on these findings, chapter 5 introduced a new step count method, template-matching using Dynamic Time Warping. This outperformed the other four existing algorithms in most scenarios, including normal and simulated-pathological gait. This might be because the idea behind the new algorithm is to calculate the number of steps based on the shape of the acceleration signal rather than solely from the peaks of the signal. Another reason for the improved results might be that the new algorithm took into consideration not only the peaks but the combination of both peaks and troughs of the signal. This was decided because it became apparent when the signals were visually inspected that in most of the cases the peaks and troughs were almost identical. Therefore the combination of peaks and troughs provided better results in comparison to the use of only the peaks. Hence, it was decided to consider both peaks and troughs in order to have a more accurate average outcome regarding the length of the step. The new algorithm is a step forward towards the development of an improved algorithm that could be potentially used to count the number of steps of people with walking impairments.

The template-matching using Dynamic Time Warping did particularly well for slow walking (normal condition), where most algorithms performed poorly. One of the reasons for the good performance of the proposed algorithm for slow walking might be that an adaptive threshold was used to take into account the estimated walking speed. This is of great importance since the amplitude of the signal plays a key role in creating a representative template for a walking

step. The value of the amplitude varies among different activities, therefore a threshold that can be adapted based on each activity could possibly benefit the outcome of the algorithm. This feature enables to get results of high performance, not only for people with the average walking speed, but also for people who might walk slower than the average. Additionally, the walking speed of each individual might vary, therefore if someone tends to walk with varied speed, the results could still be accurate.

The initial plan for this thesis involved validating the algorithms developed in chapters 3, 4 and 5 for a new cohort of data from patients with pathological gait. However, it became apparent that collecting real-world data from patients was going to be problematic because of the challenges in conducting research on patients, especially during the COVID-19 pandemic.

Instead, methods for producing synthetic data were investigated. Systems for generating synthetic data offer many advantages. Firstly, it is often time-saving and cost-effective to generate synthetic data rather than collect real data, especially in the healthcare sector. This is because the health data is sensitive and hence requires a time-consuming process until the collection of data. Another benefit of generating synthetic data is associated with the privacy of the data (Wang et al. 2019). When using synthetic data there is no issue of disclosing private data since no confidential data is exposed publicly (Park et al. 2013), as only the essential statistical information are represented. This means that the data can be used by many researchers in order to conduct their own research.

A mathematical model was developed to generate synthetic acceleration walking signals. The model used three coupled differential equations to represent a three-dimensional space around a circle of unit radius in a two-dimensional plane. Using this approach a one-dimensional signal was derived in order to mimic the walking acceleration signals collected from the pilot study. The shape and size of the one-dimensional signal were represented by the sum of Gaussian exponentials. Put simply, each Gaussian distribution has a single peak, therefore four Gaussian distributions were used to emulate the walking pattern. The alternative approaches identified from the literature were not applicable for the scope of this thesis. For example, the pendulum system required external information for every participant and also the ultimate result was not acceleration. Therefore this was not a sufficient solution. Additionally, the Generative adversarial networks approach seemed an effective method because the existing data is used to create a greater amount of that data with some differences. This approach was not used

because it is required to have a large amount of data in order to successfully develop realistic synthetic data.

The periodic signals from normal gait were replicated relatively successfully. Real-world simulated-pathological gait tended to be much less regular, with variability in the amplitude of the peaks for each gait cycle. This variation was not well accounted for in the model, so the resulting generated signals were qualitatively different to the original signal from which the morphology was based. This means that we might need to increase the complexity of our mathematical model to ensure that signals with higher complexity could be generated successfully as well. This can be achieved by including a feature that could add a degree of irregularity to match the irregularity of the real world signals.

Even though synthetic data offers many benefits, it has some limitations. While it can mimic many properties of the original data, the current model was not able to emulate real world data completely accurately. This is because often the models identify the general pattern (average) in the original data, hence sometimes the authenticity of the data might be lost. This can be solved by identifying second order patterns, where a better average representative could be developed. In general, the quality of the synthetic data is based on the quality of the original data, therefore it is important to be considered for any future work. However, a range of parameters will be provided in this thesis to generate the synthetic signals without the need of developing the whole model. Finally, a general drawback of the synthetic data is the acceptability to the user because of perceived concerns over validity or representation of the real-world signal. The best way to ease concerns is by validating the model with a large number of cases. Additionally, the potential users of the synthetic dataset could be identified and then asked different questions regarding the acceptability of the model. The researcher would try to develop a model that covers the needs of the potential users. In terms of the technical performance of the simulated-pathological signal, the developed synthetic pathological signals should be validated with real pathological data.

7.1 Study limitations

Most of the limitations for each chapter have been discussed in the above section, however there are some limitations that influence the results of all the chapters in general. These are associated mainly with the data collected. The data was collected from 30 healthy volunteers.

Even though this sample might be enough for a technical pilot study (Julious 2005), it is not large enough to be considered representative of the real-world. For example, in UK more than 400000 people suffer from Rheumatoid Arthritis. Based on the study conducted by (Israel 1992), the results showed that for more than a 100000 size of population, the sample size should be 1111, 400, 204 and 100 for precision of ±3%, ±5%, ±7% and ±10%. Therefore, to collect data from a sufficient number of participants, more than 100 participants would ideally be recruited.

Additionally, each volunteer performed the nine activities twice, once under the normal condition and once under the simulated-pathological condition. This means that it was not possible to explore repeatability of data from the same volunteer for a specific condition. For example, if each volunteer had performed the activities two or three times for each condition, the intra-subject variability of the results could have been tested. Understanding intra-subject variability would be useful for developing refined personalised algorithms to have a more general overview of how a volunteer performs the activities.

Another limitation is that volunteers performed the activities in a structured way in a laboratory setting which limited the possibility of performing the activities as in real-life. Again, the structured laboratory setting and the structured method are not ideal for performing the activities, in which a single activity was performed at a time, obviously does not accurately reflect real-life.

Lastly, and probably most importantly, healthy volunteers were asked to emulate patients with walking impairments. Even though the data collected under simulated-pathological conditions differs from the data under normal conditions, it is not truly representative of the real patient population. However, the data were collected under controlled conditions and had the essential characteristics to develop the desired algorithms. For example, participants performed the activities with a slow shuffling gait. This data can be the basis for future refinement, since this PhD was about providing an engineering solution, not to provide clinical data. The engineering solution in this case can be considered as a proof of concept of a potential future engineering solution to a known clinical problem. This might be another limitation, however for future plan a prototype would be built to test the models developed. . The prototype would include components such as ideal signal capture, training data, field test data and recessing approaches (e.g. on board or online). The prototype was not part of this thesis.

7.2 Generalisability & implications for further research

As mentioned previously, the data collection was undertaken in a laboratory setting and it is likely that real-world performance would be poorer (Chowdhury et al. 2017). In a real-life environment, sometimes the activities are performed simultaneously and with greater complexity. For example, the arms and legs swing during walking activity, however someone might talk on the phone and hence one of the arms may not be swinging providing confounding inputs (Bui et al. 2018). Additionally, the data that represents the physical activity of the user for a certain period of time is likely more messy in comparison to the data collected in a laboratory (Dutta et al. 2018). This is because the user performs several activities throughout the day, and these activities are not performed in sequence.

Device use is also important, for instance the user might take off the device and then forget to wear it again (Kosmadopoulos et al. 2016). For example, if the device is not waterproof, the user may take it off before he/she takes a shower and then might forget to put the device again, hence information about the user's physical activity is lost. Electronic devices will need to be charged and when this is being done, information about physical activity is not collected during the charging the period. Again the users may forget to replace the device after charging or they might forget to charge their device, and therefore data will be lost (Rodgers et al. 2019).

For future research, raw accelerometry could be potentially used, however to do this a large dataset is essential as input in the deep learning algorithms. This is because these algorithms require large datasets in order to work successfully. Common machine learning algorithms take features as inputs instead of raw signals.

Since activity monitoring is an interesting and exciting topic for future work, a number of recommendations for future research are given.

The first relates to acquisition of underlying data, specifically to collect data from: 1) larger groups of healthy participants, 2) real patients with a range of different conditions and severity of walking impairments, and 3) activities conducted in a real-life environment.

On a more technical level, another aspect for future research to explore is associated with the step count algorithm developed for this thesis, template-matching using Dynamic Time Warping. The algorithm works well for the normal condition, and it works better than the existing algorithms from the literature for the simulated-pathological condition as well. However, there

is room for improvement when considering pathological gaits. The algorithm could be updated to account better for variability and refined to be person-specific. This could be achieved by first looking at how the algorithm can be completely automated in terms of the thresholds used. Currently, a constant which was identified for each specific activity through trial and error is used to calculate the threshold. The threshold can be considered partly-adaptive since the constant is multiplied with a parameter (peak distance) identified from the acceleration signal of each participant separately. For future work, it is suggested to have a completely automated threshold that could be identified solely from the input signal, resulting in no human input requirement. In case the volunteer performs different walking speeds, the algorithm should be able to adapt its essential thresholds according how each volunteer walked. Based on the results from the simulated-pathological condition, it is demonstrated that person-specific thresholds are required. This is because the simulated-pathological data were more variable and hence the adaptable thresholds for each activity were not representative for all 30 volunteers. For example, a constant was calculated based on the experimental results for each activity, and then an adaptable threshold was calculated for each volunteer. However, the constant given for each activity might differ from person to person, hence this constant could also be refined to be adaptable based on the data from each person. The variability of the simulated-pathological signals might be due to the fact that healthy people simulated the activities. This might result in the creation of inconsistent gait cycles, although inconsistency in gait is a hallmark of many disordered gait patterns (Esser et al. 2011; Del Din et al. 2019; Yamada et al. 2012).

In terms of the personalised healthcare, it may be useful to build an application to demonstrate the outputs to both clinicians and patients. For example clinicians could have access to data that shows the activity of the patient during a whole year. This could be of great help because clinicians will have a better understanding of how much active their patient was, and that means their decisions will be patient-specific and better informed. Step count is another important objective measure for the clinicians to understand how active their patient has been. It can be also used to identify the state of each patient in relation to other health variables. Also, steps can be used as a common metric among all clinicians, and hence the results can be used by any type of doctor (Bassett et al. 2017). Additionally, the patients might be able to access their data at any time, and hence take a decision about their behaviour associated with their disease and treatment. For instance, an RA patient could check his/her activity level through the app, if the activity level is low the patient has the chance to try to increase the activity

level at the desired daily goal. This can also be done by the number of steps taken daily and steps can also be used as a motivational tool. Also, the app might automatically send reminders and guidance to the patients about their activity levels. Additionally, wearables are available with many different types of sensors, such as accelerometer, magnetometer, gyroscope, global positioning system, pressure sensor etc. In a future study, multiple types of sensors could be used simultaneously to achieve better results in terms of activity monitoring. For example, a global positioning system could be used to locate the subject, and this might be used as another parameter in identifying the activity performed.

Further research on synthetic signal generation for pathological gaits is warranted. The method used for generating synthetic signals is not new, however it is new in the gait analysis field and this method has not been used before to generate walking acceleration signals. For the normal condition group, the method had worked well because of the nature and morphology of the signals as aforementioned. For the simulated-pathological group, the results could be improved by building a more complex model that includes the irregularity of the signals representing impaired gaits. For future work, it might be essential to study in more detail the irregularities of the signal formed by different pathological conditions. When this is understood, it might be possible to use mathematical functions to represent those irregularities accurately. Algorithms based on original data from real patients with a range of conditions will allow construction of more representative models especially when combined with refined models. High quality synthetic signal generation will potentially open many potential doors since data can be acquired/generated faster and without any ethical constraints.

7.3 Summary

Currently, most existing activity monitoring apps perform poorly for those with significantly abnormal gait. This study has provided the first steps to address this, and has provided a series of advances in understanding the technology that will provide a platform for future developments.

This project confirmed that wrist location is both preferable to patients and a technically viable option for developing algorithms for activity classification, condition classification and step count. For the activity classification, an accuracy score of 0.984 was reached. For the condition classification, an accuracy score of 0.949 was reached. For the step count, an average error for all the activities was 1.88, where the largest was 6.97 for the normal condition. Also,

an average error of 11.52 was calculated, where the largest error was 55.86 for the simulated-pathological condition. Using this location, the algorithm could differentiate the state of the user, healthy or pathological. This is an important feature because it might be used to track the progress of a patient (Trost and O'Neil 2014; Durstine et al. 2013; Pedersen and Saltin 2015). Then, using the appropriate algorithms the activities can be classified with high accuracy in both normal and pathological states, yielding accuracies of as 0.984 and 0.967 respectively. This means that clinicians can be informed about the activity of their patients without having to rely on their patients' memory (Trost and O'Neil 2014). With respect to the activity of the patients, clinicians should get also results with greater accuracy about the number of steps that each patient has performed even when the person is severely compromised.

There remains work to do in terms of algorithm development, since there is a need for development of person-specific algorithms, which will enable personalised healthcare, hence achieving better results for the patient. This project showed that possibilities exist in the realm of personalised healthcare. For example, the step count algorithm template-matching using Dynamic Time Warping showed that by using the morphology of each signal individually an algorithm has the potential to produce results with high accuracy, since the root mean square error was between 1.31 and 2.69. However, for the pathological gaits especially, the algorithm will need to have greater complexity and consider more parameters in order to yield accurate results. In comparison to the existing algorithms from the literature, the template-matching using Dynamic Time Warping achieved better results for both normal and simulated-pathological conditions, which demonstrates that this approach has the potential to be successful for the patient population as well.

The development of person-specific algorithms would enable more accurate measures of activity and step count to be collected. Therefore, remote monitoring can become a more standard way of measuring the physical activity of patients.

Appendix A

Stakeholder analysis

A.1 Introduction

The stakeholder analysis was led by the candidate with the help of the NIHR Leeds Biomedical Research Centre's Patient and Public Involvement team who provided a venue and practical support to the sessions.

A stakeholder is defined as "any group or individual who can effect or is affected by the achievement of the project's objective" (Freeman 1984). There are many benefits when seeing health interventions from different angles. Firstly, it is important to understand the perspective of a key decision maker. This will increase the chance of successfully implementing the project. Secondly, knowing the concerns and expectations of the end users will increase the possibility of having a successful product. Additionally, understanding multiple stakeholder perspectives provides the chance to refine interventions and think of more innovative ideas to meet the widest range of stakeholders needs. It is also possible to influence key stakeholders by knowing their needs. Lastly, key stakeholders share their views and this might improve the quality and change the way of thinking about a specific intervention (Hyder et al. 2010).

The main reason for conducting a stakeholder analysis for this project is to ensure understanding of the user requirements, in terms of the characteristics of the ideal activity monitoring device and platform. This section describes the process followed and the outcomes identified.

A.2 Methodology

The stakeholder analysis follows a set of steps that any individual, organisation, or company should follow, but each step can be executed with different methods. There are two basic steps that are ubiquitous; firstly to identify the stakeholders and their interest in the project, secondly to assess the importance and influence of each stakeholder (Jones 1976; Jepsen and Eskerod 2009).

Even though stakeholder analysis has been used widely, there is no agreed systematic approach to identify and analyse stakeholders (Bryson 2004). Some techniques are widely employed in stakeholder analysis (Ingen 2010; Schmeer 1999; Bryson 2004; Haleem 2008; Jepsen and Eskerod 2009) and a selection of these techniques, particularly those used for public/patient engagement are described below:

A.2.1 Identifying stakeholders

The most relevant stakeholders for the particular project are patients, physiotherapists/clinicians, engineers and scientists. Table A.1 illustrates the characteristics of the stakeholders in the current study.

A.2.1.1 Power versus interest matrix

This method is used to categorise the stakeholders and also to prioritise them (Bryson 2004). A power Vs interest diagram consists of x- and y-axes. The x-axis represents the power of stakeholders and the y-axis represents the interest of stakeholders over the project.

A.2.1.2 Stakeholder influence diagram

A power/influence matrix is developed to keep the project focus on the important stakeholders (Bryson 2004). Similarly to power Vs interest diagram, an influence diagram consists of x- and y-axes. In this case, y-axis represents the power of the stakeholders. And x-axis represents the influence of the stakeholders. These terms might seem very similar, however influence is voluntary, while power is forced.

A.2.1.3 Bases of power - directions of interest diagrams

The particular diagram identifies the "powers" of each stakeholder that can influence the project and their interests (Bryson 2004). This enables the team to identify any common interest between the stakeholders. A diagram for each of the main stakeholders is developed with the stakeholder's name written in the middle of the diagram. The bases of power are written below the stakeholder box with arrows pointing towards it. Lastly, the directions of interest are written above the stakeholder box with arrows from the box pointing towards the interests' box.

A.2.1.4 Stakeholder support versus opposition grids

This two-by-two matrix identifies the support, opposition and importance of each stakeholder about the device (Bryson 2004). The x-axis represents the power of the stakeholders. The y-axis represents the opposition and support of the stakeholders. The top row of the matrix is about the stakeholders who support the project and the bottom row is about the stakeholders who resist the project.

A.2.2 Interview and questionnaires

Hypotheses have been developed using different stakeholder analysis techniques targeting the needs of patients and clinicians. The interviews and questionnaires were used to identify their real perspective on different aspects about the wearable device.

Open-ended questions and questionnaires were developed for each stakeholder. The questions were reviewed by the research team three times until finalising the questions. Since the stakeholder analysis was performed to understand what the ideal device will look like, questions were mostly related to features, specifications and appearance of the device.

Patients and clinicians were contacted via email. For the patients, a focus group of people with RA was convened by the PPI manager of NIHR Leeds Biomedical Research Centre at Chapel Allerton Hospital.

A.2.2.1 Patients

A.2.2.1.1 Interview questions The first step in developing the exploration questions was to understand the reason for meeting with the patients. Based on the reason identified, several questions were prepared in advance by the research team. Often, focus group meetings last

for up to two hours. Therefore, this time duration was considered to identify an appropriate number of questions to have for the first part of the meeting, the discussion. Nine exploratory questions were designed for the RA PPI group discussion. The questions covered the following areas: (1) the wearability of the device, (2) the interaction of patients with the device, and (3) the captured information of the device.

A.2.2.1.2 Questionnaire A similar process was followed to develop the questions of the questionnaire. The only difference between the questions on the questionnaire and those in the discussion, were that the majority of those in the questionnaire were closed-ended questions. The choices of each answer were based on different literature findings. The patients' questionnaire included 8 closed-ended questions and 2 open-ended questions (see appendix C).

A.2.2.1.3 PPI group session A brief introduction will be given to the participants about the project. Then, the exploration questions will be followed to start discussing with the patients. The reason for asking the open-ended questions first was to give the opportunity to patients to think with their own opinion, instead of being influenced by the questionnaire's choices. And to ensure that everyone understood the project and that they did not have any misunderstanding. Then, the questionnaires were given to the participants to fill them. The researcher was there to supervise the patients if they required any help for completing the questionnaire.

A.2.2.2 Clinicians

A.2.2.2.1 Interview questions A similar process was followed to develop the exploration questions for the clinicians. Many different questions have been written down, and the research team revised them three times to finalise the questions. Since clinicians have tight schedules, four exploratory questions were selected for the discussion part of the session. In this case, the questions covered technical aspects of the wearables that might influence the clinicians.

A.2.2.2.2 Questionnaire A similar process was followed to develop the questions in the questionnaire in order to understand clinicians' perspective. However, several closed-ended questions were written, and the least important were removed. Finally, seven closed-ended questions and two open-ended questions constituted the final questionnaire. In contrast to the questions developed for the discussion, the questions in the questionnaire covered aspects related

to: (1) wearability, (2) interaction, (3) captured information, and (4) clinical information related to PA.

A second questionnaire was developed after the session with the PPI group. The reason for developing a second questionnaire was for the clinicians to rank patients' answers that were given during the discussion. Four closed-ended questions were developed covering the captured information, comfort, appearance, and features of the device.

A.2.2.2.3 One-to-one meetings The meetings with clinicians were one-on-one instead of a group meeting. The main reason for that was their busy schedules, therefore it was not practical to arrange a group discussion. At the beginning of each meeting, a summary of the project was given to clinicians. Then, the interview questions were asked to start a conversation with each clinician. The reason for starting with the exploration questions first was similar to the reason given for patients. Clinicians might be influenced from the choices of the questionnaires, and this was not the ideal case. Following that, the first questionnaire was given to the clinicians for completion. The questionnaire 1 was about clinicians' perspective for the wearable devices. Then, the questionnaire 2 was given which was the ranking of patients' answers.

A.3 Results

A.3.1 Identifying stakeholders

A priority list of stakeholders has been developed in a table to prioritise the stakeholders (see Table A.1). The list includes information about the number of interviewees and the reason and/or relation to the project. Additionally whether the stakeholder is internal or external to the project (Schmeer 1999).

Table A.1: General information on priority stakeholders to be interviewed.

Sector	Subsector	Internal/ External	Interviees	Relation to project
Patients (End user A)	People with RA, related to patients with RA and walking impairments	External	9	Identify: 1) the characteristics of device that meets their needs as a wearer, 2) the extra information they require the device to provide
Doctors / Physios (End user B)	Rheumatologists, Physiotherapists, Podiatrists	External	3	Identify the most important information that the device should capture

A.3.1.1 Power versus interest matrix

Clinicians and patients have high interest and power towards the project. Figure A.1 demonstrates the identified position of stakeholders. Clinicians have high interest because they will improve their decision making regarding their treatment plans. Additionally, clinicians have high power because if they decide not to use the system then the project might fail. On the other hand, patients have high interest as well since this project might positively impact their treatment. Another reason explaining their high power is that if patients decide not to use the device, then the project might fail.

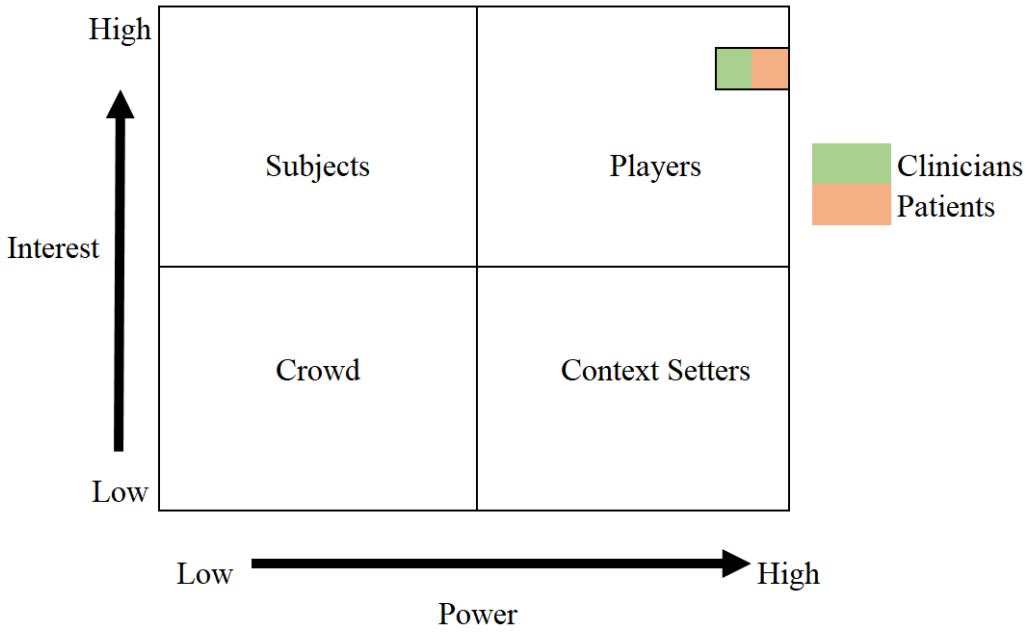


Figure A.1: Power Vs Interest Matrix

A.3.1.2 Stakeholder influence diagram

It has been prioritised that patients and clinicians were the most important stakeholders for this context. A power/influence matrix is developed to help to keep the project focus on the important stakeholders (Bryson 2004). Since patients and clinicians have high power over the project, they can significantly influence it. For example, they can influence the outcomes of the project through the stakeholder analysis. The research team had identified gaps in the literature in relation to wearable technology in RA patients. However, patients and clinicians might suggest different opinions from their perspective which have not been identified in literature. These opinions can inform the end product.

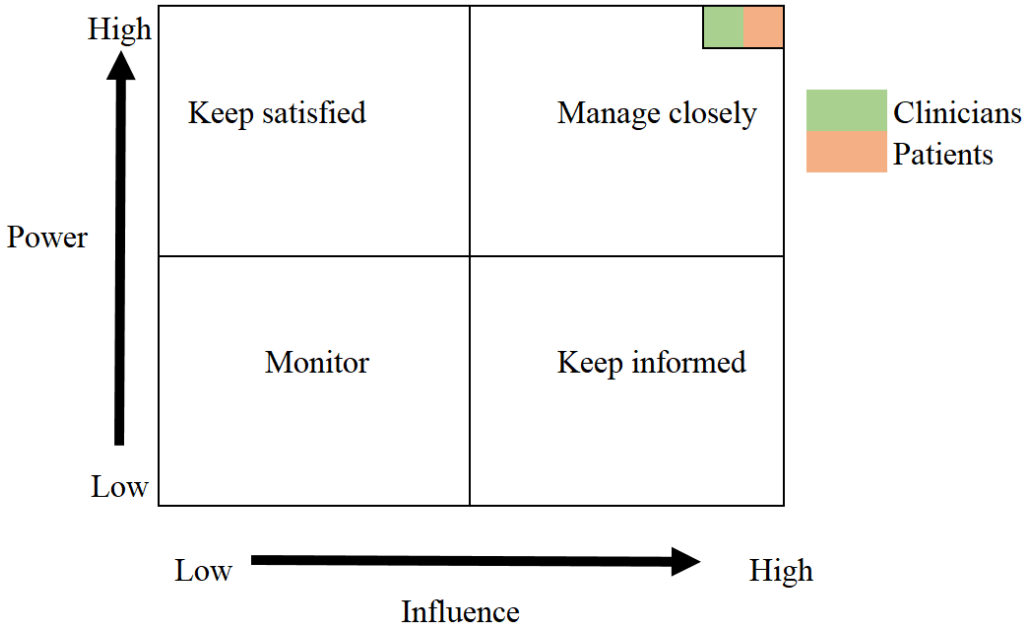


Figure A.2: Power Vs Influence Matrix

A.3.2 Creating ideas for strategic interventions

A.3.2.1 Bases of power - directions of interest diagrams

Two diagrams have been developed, one for each of the two most important stakeholder, PPI and clinicians. Regarding the PPI group, the reasons that have power were: (1) they are the end users, (2) the usability of the device, (3) they are experts for their condition and needs, and (4) the wearability of the device. Clinicians were end users as well, they used the outcomes from the device. Therefore, clinicians' power can be described as: (1) end users, (2) they have the medical knowledge about the patients' condition, and (3) the usability of the information

captured. These two groups shared similar interests. For example, they both care about the information that will be captured from the device. Additionally, they both have an investment in the features offered by the device. Patients were more interested about the comfort and appearance of the device. Clinicians were interested to enhance their decision by having extra information for the PA of the patients, which will be more accurate than questionnaires and discussion with patients on a single visit twice a year.

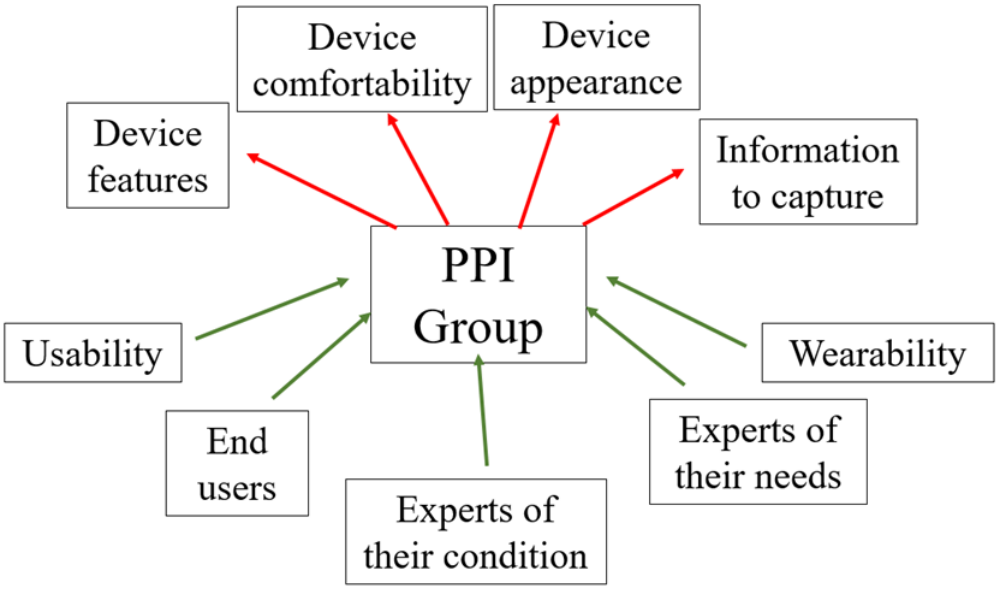


Figure A.3: Bases of power – Directions of interest Diagram - PPI Group

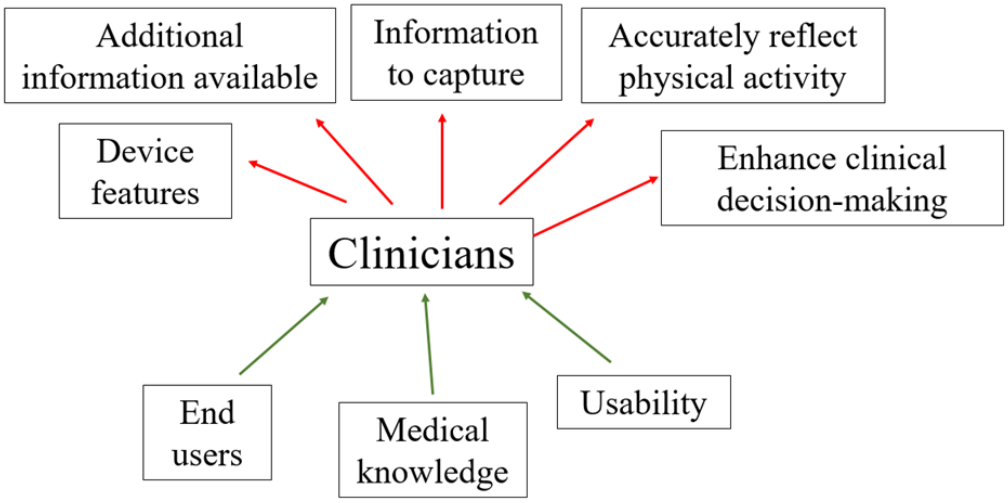


Figure A.4: Bases of power – Directions of interest Diagram - Clinicians

A.3.3 Techniques for proposal development review and adoption

A.3.3.1 Stakeholder support versus opposition grids

All stakeholders fall in the support category. Clinicians and patients are strong supporters since both might be influenced positively from the outcome of the project.

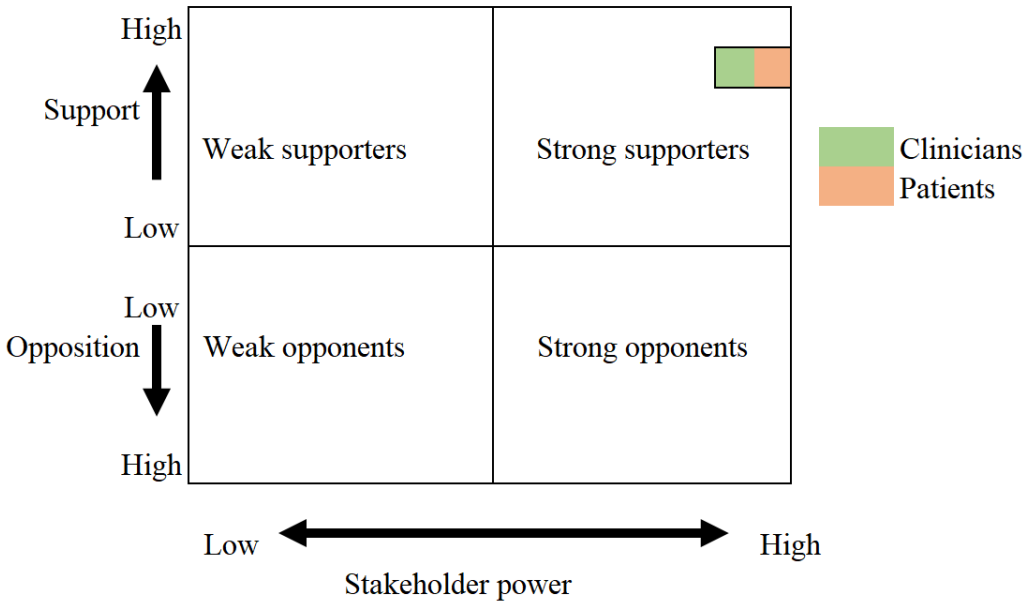


Figure A.5: Support Vs Opposition Matrix

A.3.4 Interview and questionnaires

A.3.4.1 PPI group

A.3.4.1.1 Interview Patients expressed several thoughts about their ideal wearable. All of their answers have been categorised in the three categories, (1) wearability, (2) interaction, and (3) captured information as demonstrated in Table A.2 . Even though each question was categorised in one of these sections, the answers of the patients were related to a mixture of these categories.

Table A.2: Patients' interview answers.

Wearability	Interaction	Captured information
Design	Features	Steps
Comfortable	Provide feedback	Activity duration
Accessible	Working/Recording/Battery light	Activity intensity
Compatible with patients	Audio options	Activity types
Robust	Be identifiable easily	Timing of activities e.g. morning
Waterproof	Motivation tool	Pain level
Slip-on strap	Linked to mobile	Heart rate
Magnetic strap	Battery life (once a week)	Blood pressure
Clip-on device	Data download (once a week)	Seasonal information
Colour options	Set goals	
Thin	Digital	
Not medical look	Design	
Watch	Buttons (usage)	
Light (weight)	Simplicity (usage)	
Material (not sensitive to skin)	Instructions/Communication	
Small (size)	Clear instructions (usage)	
Placement	Contact details available	
Neck	Clear benefits	
Not inflamed positions	Charging instructions	
No ankle	Alert signal	

A.3.4.1.2 Questionnaire Of the nine patients who contributed to the interviews, eight provided further feedback on the questionnaire. Based on the answers of the first questionnaire question, six of the participants were positive about using an activity monitor, one was negative, and the other one was ambivalent. Additionally, six of the participants had a smartphone, and two did not have a smartphone.

One of the questions was regarding the preferred placement of the device. Participants had the chance to select more than one placement. All respondents selected wrist, one selected waist as well, and another two selected chest as well as wrist. One of the options was ankle placement since is the most common and accurate location for step counting (Fortune et al. 2015), however, none of the participants selected this option. During the discussion, they strongly agreed that ankle placement is not suitable for them, which is mostly due to their joint involvement which makes it difficult to bend to apply and interact with a device at ankle level.

Regarding questions about the minimum number of days between battery charges, and also between days to download the data, the results were very informative. Five participants selected

that the minimum number of days to charge the device and download the data is between seven to nine days. Two selected four to six days to download the data, and three selected four to six days to charge the device. One of the participants selected two boxes, 4-6 and 7-9 for charging the device and for downloading the data.

Regarding the features that the accompanying software application might potentially offer, all participants selected to be notified in terms of their activity levels and to record their daily activity performance. Six of the participants also wished the application to record level of pain and feelings.

Participants ranked their perception of the importance of the different types of information, with "1" being the most important, and "7" being the least important. Seven options were provided, 1) active Vs sedentary, 2) number of steps, 3) walking, 4) sitting, 5) standing, 6) sit-to-stand, and 7) time spent on each activity daily. Some of the participants ranked more than one option with "1" or other number. Table A.3 demonstrates which activities were selected. The most important activities selected were the number of steps, walking, and time spent on each activity daily.

Table A.3: Information to be captured (1 – most important; 7 – least important).

Rank	1	2	3	4	5	6	7
Active Vs Sedentary	0	2	0	3	0	1	2
Number of steps	3	2	1	0	1	0	1
Walking	3	3	0	1	0	1	0
Sitting	0	0	2	3	2	0	1
Standing	0	1	2	1	0	4	0
Sit-to-stand	0	0	2	0	4	0	2
Time spent on each activity daily	3	1	2	0	0	1	1

The most important specification for the device as ranked by half of the participants was to be comfortable. Additionally, the appearance was quite important to them, hence they suggested that they prefer the device to not look like a medical device. Attachment and accessibility were also important. Table A.4 shows how participants ranked the specifications of the device.

Table A.4: Device specification (1 – most important; 5 – least important).

Rank	1	2	3	4	5
Comfort	4	2	1	0	0
Discreteness	0	0	0	4	3
Appearance	2	0	2	1	2
Accessibility	1	2	2	2	0
Attachment	1	3	2	0	1

A.3.4.2 Clinicians

A.3.4.2.1 Interview Similarly to PPI group, the interview answers given by the clinicians have been categorised in the following categories: wearability, interaction, and captured information.

Table A.5: Clinician's interview answers.

Wearability	Interaction	Captured information	
Robust	Download data automatically	Feelings	Overall activity level
Easy to put on/off	Set goals	Struggling activities	Steps
Placement options	Simple to use	Track progress based on interventions, medications, disease activity	Walking speed
Waterproof	Download data fast	Comparison system - patient A Vs patients RA Vs Norms	How much they do in one go

A.3.4.2.2 Questionnaire The following results are based on three clinician's perspective. Questionnaire 1 is about the clinicians' perspective for the device. Questionnaire 2 is about clinicians' perspective about patients' interview answers. Similarly to the patients' questionnaires, "1" is the most important, and the higher number (6 or 7 or 8) is the least important.

A.3.4.2.2.1 Questionnaire 1 Two of the clinicians selected that seven to nine was the minimum acceptable number of days to charge the battery of the wearable and to download the recorded data. And one of the clinicians selected that one to three was the minimum acceptable number of days to charge the battery of the wearable and to download the recorded data. Regarding the placement of the device, wrist and ankle were the preferred locations. Moreover, clinicians mentioned that they measure the physical activity of their patients subjectively using questionnaires. The records were stored on paper, electronically in a document format or

electronically in a platform. Additionally, the clinicians indicated that an appropriate cost for the device should be between £0-100. Lastly, the clinicians were asked whether they believe that this project might benefit them. They suggested that this project will aid to: (1) understand the areas of difficulty that the patient suffers from and the impact of different interventions, (2) provide accurate feedback to the patients about their activity, and (3) predict and prevent injuries since they often occur after a period of inactivity.

Table A.6: Information to capture (1 – most important; 7 – least important).

Rank	1	2	3	4	5	6	7
Active Vs Sedentary	3	0	0	0	0	0	0
Number of steps	0	1	1	1	0	0	0
Walking	0	0	1	1	0	1	0
Sitting	0	0	0	0	1	0	2
Standing	0	0	0	0	1	1	1
Sit-to-stand	0	0	0	1	1	1	0
Time spent on each activity daily	0	2	1	0	0	0	0

Table A.7: Activity monitor features (1 – most important; 7 – least important).

Rank	1	2	3	4	5	6	7
Screen	1	0	1	1	0	0	0
Water resistant	1	1	0	0	0	0	1
Sleep tracking	0	0	0	1	1	0	1
Heart rate monitor	0	0	0	1	2	0	0
Temperature sensor	0	0	0	0	0	2	1
USB connection	0	1	1	0	0	1	0
Bluetooth	1	1	1	0	0	0	0

Table A.8: Information to capture (1 – most important; 6 – least important).

Rank	1	2	3	4	5	6
Pain	0	2	0	0	1	0
Heart rate	0	0	1	0	2	0
Steps	0	0	0	2	0	1
Blood pressure	0	0	0	1	0	2
Activity duration	2	0	1	0	0	0
Activity intensity	1	1	1	0	0	0

Table A.9: Comfortability (1 – most important; 6 – least important).

Rank	1	2	3	4	5	6
Slip on wrist	1	0	0	2	0	0
Round the neck	0	0	3	0	0	0
Clip-on device	1	1	0	0	0	1
Location that it is not inflamed	0	1	0	1	1	0
Thin that can be worn a bit higher of the wrist	1	1	0	0	0	1
Materials that are not sensitive to the skin	0	0	0	0	2	1

Table A.10: Appearance (1 – most important; 7 – least important).

Rank	1	2	3	4	5	6	7
Something that does not look medical device	0	0	2	1	0	0	0
Colour options	0	0	0	0	0	1	2
Watch	0	1	1	1	0	0	0
Light (weight)	2	1	0	0	0	0	0
Material	0	0	0	0	1	1	1
Light (working/battery/recording)	1	1	0	0	0	1	0
Audio	0	0	0	2	1	0	0

Table A.11: Offered features (1 – most important; 8 – least important).

Rank	1	2	3	4	5	6	7	8
Waterproof	0	0	0	1	1	0	0	1
Setting goals	0	2	1	0	0	0	0	0
Motivation tool	0	1	2	0	0	0	0	0
Daily feedback	1	0	0	1	1	0	0	0
Audio	0	0	0	0	1	1	1	0
Digital	1	0	0	0	0	1	1	0
Alert signal	1	0	0	1	0	0	0	1
Identifiable	0	0	0	0	0	1	1	1

A.3.4.2.2.2 Questionnaire 2

A.4 Discussion

A.4.1 Stakeholder analysis techniques

A.4.1.1 Identifying stakeholders

Patients and clinicians are the most important stakeholders since their opinion might influence the features of the end product and the lifetime of the project.

It is important to understand the perspective of patients because they are the end users of the device. There is no point in developing a device that it is unacceptable or useless to the patients. Understanding the perspective of clinicians is equally important. This is because clinicians are end users of the output of the device. It is important to ensure that the information that clinicians get is usable and useful to them.

A.4.1.2 Power versus interest matrix

As already mentioned, patients and clinicians have the power to undermine the project if their involvement is not prioritised. This is the reason that they have high power at the power Vs interest matrix. Similarly, clinicians and patients have great interest in the project. The extra information that clinicians will get, it might help them in their decision-making. And if clinicians take better decisions, then more effective treatment plans will be given to patients. Additionally, through the device, patients might be motivated to be more active.

A.4.1.3 Stakeholder influence diagram

As already demonstrated, patients and clinicians have high power and influence. Both groups have the power to fail the project by not agreeing to take part. But also, they can both influence positively the project by suggesting different ideas from their perspective. The research team will be able to develop a system that covers the needs of both groups, and therefore the usage of the system will be effective towards clinicians and patients.

A.4.1.4 Bases of power - directions of interest diagrams

The research team has hypothesised that the "PPI group" and "clinicians" share common powers and interests. The shared powers of these two groups are that both will be end users, and the usability of each user will be important. If the majority of each group does not use the device/application, then the project might fail. Patients also have the power of being experts regarding their condition and needs. Clinicians though have the medical knowledge about the condition of the patients. Therefore, they know what type of information is important to know for enhancing the treatment of patients. Clinicians also have the power to use or ignore this extra information that has been provided to them.

Additionally, the two groups share similar interests, such as information to capture and the offered features of the device/application. Since patients will be using the device, their first

interest might be about the information that wearable will capture and offer to them. Clinicians might be interested about the offered features of the device since these features might influence the captured and offered information. Moreover, they might be interested to enhance their decision-making on patients' treatment. This will be done using the extra available information, which will be accurately recorded.

A.4.1.5 Stakeholder support versus opposition grids

Both stakeholders are supporters for this project. Clinicians and patients are strong supporters, because both groups will be positively influenced through this project.

A.4.2 Interview and questionnaires

A.4.2.1 PPI Group

A.4.2.1.1 Interview The WHO suggested that the FITT principle can be used to monitor how active you are (Verlaan et al. 2015). Patients have included three of the four aspects of this principle in the information that they like the device to capture. Intensity, type and time have been mentioned, and frequency not.

They also suggested that the time of the day and the season is important to them. This is because their pain, fatigue, and stiffness levers are changing based on these two factors (Rojkovich and Gibson, 1998; Feldthusen et al., 2016). There is also data to show PA varies by season and weather changes (Tucker and Gilliland 2007).

Patients would also like to know their blood pressure. At the moment, no such sensor has been developed for this purpose. This makes it difficult for the research team to implement this idea. Regarding the heart rate sensor, other studies have combine accelerometers with heart rate sensors. Their findings demonstrated positive results, therefore this suggestion will be considered.

Moreover, patients have asked to record their pain level. Even though there is no sensor available to record pain, this project might provide patients a device or an application to track their pain. This might be useful for both patients and clinicians.

In terms of the features of the device, patients suggested several ideas that already exist in the market. For example, setting goals, used as a motivation tool, the device to be digital, be able

to connect with smartphones and have audio options.

They suggested also that they would like the device to have lights showing when it is working, recording, and charging. This is because they are not familiar with technology. Additionally, the device should have a feature that makes easy to identify if lost.

RA patients experience joint problems, therefore many have reduced dexterity. They suggested any buttons on the device should be simple to use. Additionally, the devices might fall several times and hence it has to be robust. Several things that patients expressed about the device did not come to my attention in the literature. First of all they were all unwilling to wear the activity monitor on their ankle. They prefer wrist location, with a slip-on or magnetic strap instead of buckle. Additionally, the device should be thin and light weight in order to be placed a bit higher than the wrist.

These were the most important findings from the discussion with patients. Some of their suggestions are applicable, but some others are difficult to implement. However, all answers will be carefully considered.

A.4.2.1.2 Questionnaire Patients agreed that seven to nine days is a sensible number of days to download or transfer data. However, charging the device every seven to nine days will be difficult. This is because the device will be used continuously throughout the day, and most devices in the market can be left uncharged for maximum three/four days. Although it is a challenging problem, when the algorithms will be developed the battery life of the device will be carefully considered. The ideal scenario is to produce algorithms of low power consumption and increased battery life.

Comfort was the device specification most highly ranked by patients. They also expressed that appearance is quite important to them. In the early stage of the project, appearance might be ignored since the aim of the project is to develop a device that works accurately. Appearance can be modified at the end stage of the project based on patients' ideas and needs.

A.4.2.2 Clinicians

A.4.2.2.1 Interview Equally important was to understand clinicians' perspective and identify any similarities or differences between clinicians and patients' views. The comparison system would give the opportunity to clinicians to compare the PA of the individual patient with the

average PA of a group of patients. Each group will consist of subgroups with their own average values based on age and gender. This will help them to understand whether a treatment that they provided is working or not. Additionally, the system will be able to compare the individual patient with norms. Since clinicians try to help patients to have a "normal life", the system will help them to compare patients with norms. The clinician suggested the device to be position independent. This is to give patients the opportunity to place the device in their ideal position, and hence clinicians will still get the desired PA information. Additionally, the suggestion of having a system that can measure the quality of the exercise of each patient would be valuable. For example, being able to identify the quality of walking activity of a patient with walking impairments.

A.4.2.2.2 Questionnaire Two of the clinicians selected seven-to-nine days and one clinician selected one-to-three for downloading data and charging the device and also noted that wrist and ankle were their preferred location for wearing the device.

The clinicians validated literature in terms of how they measure PA. They do not measure PA objectively, but subjectively through patients' reported activity. Therefore, currently clinicians do not have an accurate overview of the PA of the patients.

The top three answers for the most important information to capture were: 1) how active or sedentary the patients were, 2) the time spent on each different activity daily and 3) the number of steps. These factors will inform clinicians about the PA of patients. The benefits of being active are well known, however patients still tend to be inactive. This is because they might experience pain, joint stiffness, and/or fatigue. When clinicians have an overall overview of the PA of patients, they can advise them what they should do to become more active, and vice versa.

Questionnaire 2 was given to the clinician to rank the patients' answers with a clinical view. This is done to prioritise the importance of the patients' answers. In terms of the captured information, excluding PA tracking, measuring activity duration was the most important. Clinicians ranked as least important blood pressure. If this is the case at the end of the stakeholder analysis, then this sensor will not be considered. Moreover, clinicians agreed that the device should be thin, light and be worn as slip-on. It will be easier for the patients to wear the device, and it will be more comfortable. The device is preferred to have a watch look rather than a

medical device appearance, and also have a light that shows to the wearer whether the device is working or not. Lastly, the top ranked features from the clinicians were the device to be used as a motivational tool and to set goals. This will help the patients while they are using the device, and also to keep using the device.

A.5 Summary

Stakeholder analysis has helped understand the end users perspectives and needs in order to develop a usable device. Several stakeholder techniques have been followed prior the meeting with the stakeholders. This was essential to identify the key stakeholders, and how they could be approached. Additionally, several interview questions and questionnaires regarding each stakeholder have been developed.

The three most important findings from the stakeholder analysis until now are:

1. Active Vs sedentary, step counts, and time spent on each activity is clinically important
2. Patients will not wear a monitor on their ankle
3. Clinicians would benefit from a system that compares a RA patient, with a RA group and norms

Appendix B

Stakeholder analysis: Interviews and questionnaires

Interview questions for clinicians:

1. What information would be useful for you to know regarding the physical activity of the patients?
2. What technical specifications would you like the wearable device to have?
3. How do you think a wearable device should be in order to be usable for a clinician?
4. What would be an appropriate amount to buy such a wearable device?

Stakeholder analysis interview questions

Clinicians (RA focused) / Physiotherapists

Please tick and complete where appropriate

A. Rheumatoid arthritis related questions

1. How often do RA patients visit the hospital for a check-up/medical appointment in a year?

1	2	>5	>10

2. Which activities do the patients commonly carry out in their everyday life?

_____

_____

_____

3. What is the most common range of age that suffers from the RA disease/condition?

<10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	>81

B. Technology related questions

4. Do you currently use any technologies in your work area?

_____

_____

_____

5. Do you currently use any technology in your medical appointments to measure PA? If yes, please state.

_____

_____

_____

6. How do you currently store the patients' record?

On paper	Electronically – in a document format	Electronically – Platform

7. Which of these monitors do you believe is the most suitable for the RA patients and why? (Valeria will provide monitor information to the clinicians)

A	B	C	D	E	F	G	H	I

C. General questions

8. Have you heard something similar to this project?

Yes	No

9. Do you believe that this project can benefit you? If yes, please explain.

_____

_____

_____

Stakeholder analysis interview questions

Clinicians' perspective on PPI group views

1. Information to capture. Please rank your preferences (1-6,[1:most important, 6:least important])

Measure pain	Measure heart rate	Measure steps	Measure blood pressure	Measure activity duration	Measure activity intensity

2. Comfortability. Please rank your preferences (1-6,[1:most important, 6:least important])

Slip on	Round the neck	Clip-on device	Location that it is not inflamed	Thin that can be worn a bit higher of the wrist	Materials that are not sensitive to the skin

3. Appearance. Please rank your preferences (1-7,[1:most important, 7:least important])

Something that does not look medical device	Colour options	Watch	Light (weight)	Material made of	Light (working/battery /recording)	Audio

4. Offered features. Please rank your preferences (1-8,[1:most important, 8:least important])

Waterproof	Setting goals	Motivation tool	Daily feedback	Audio	Digital	Alert signal	Identifiable

Interview questions for PPI group:

1. How can we make sure that this device will be used by people with RA?
2. What would be useful/helpful for you this device to capture?
3. What do you think of this idea?
4. In terms of comfortability, do you have something in mind?
5. In terms of appearance, do you have something in mind?
6. What features do you expect the device to offer?
7. What is the minimum number of days that you would like to charge your device?
8. Do you mind if you need to download the data every seven days?
9. Is there any reason that you would not wear such a device? If yes, please explain.

Stakeholder analysis interview questions

Patients' perspective on physical activity monitors

1. Would you use an activity monitor?

Yes	No

2. Do you have a smartphone?

Yes	No

3. Where would you prefer to wear the activity monitor?

Wrist	Waist	Hip	Ankle	Chest

Other suggestions:

_____

_____

4. What aspects of the appearance of the monitor would affect your willingness to wear it?

_____

_____

_____

5. Which specification is the most important to you? Please rank your preferences (1:most important)

Comfort	Discreteness	Appearance	Accessibility	Attachment

Other suggestions:

_____

_____

6. What would be the minimum acceptable number of days between battery charges?

1-3	4-6	7-9	10-12	13-15	16-18	19-21	22-24	25<

7. What would be the minimum acceptable number of days between data downloads?

1-3	4-6	7-9	10-12	13-15	16-18	19-21	22-24	25<

8. What information you would like the activity monitor to capture? Please rank your preferences (1:most important)

Active Vs Sedentary Periods	Number of steps	Walking	Sitting	Standing	Sit-to-stand transitions	Time spent on each activity daily

Other suggestions:

_____

_____

9. What features would you like an associated mobile app to include?

Notifications in terms of the activity level	Record level of pain	Record feelings	Record daily activities

Other suggestions:

_____

_____

10. How do you believe that activity monitoring might benefit you?

_____

_____

_____

Appendix C

Collection of activity monitor data

Activity description document

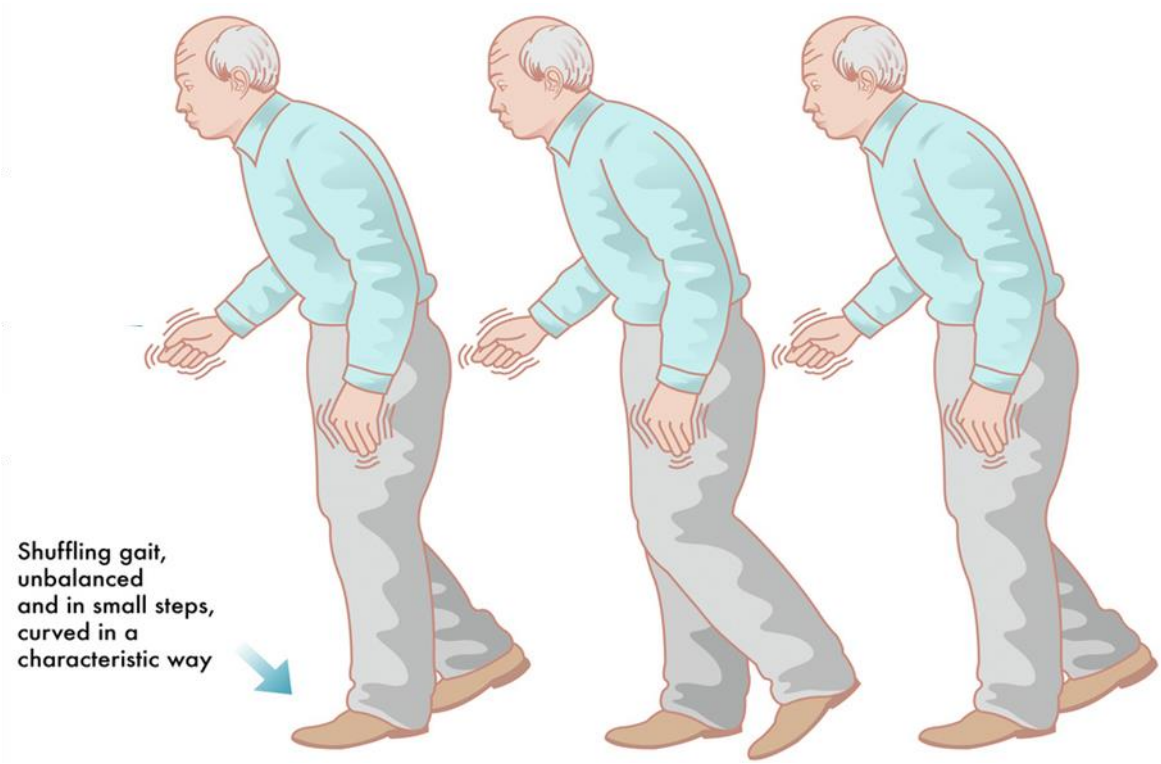
Stand in a relaxed position whilst the researcher attaches activity monitors to the wrist and ankle – elastic straps/bands will be used to keep in place the monitors

Jump at the start and end point of the selection of activities (Jump at the start (before lying down) and at the end (after stairs) of the activities)

You will perform a predetermined selection of activities of daily living – **healthy** (normal)

You will perform a predetermined selection of activities of daily living – **simulated patient** (abnormal)

- Normal Vs Abnormal (shuffling gait)
- **Shuffling gait:** A gait in which the foot is moving forward at the time of initial contact or during mid-swing, with the foot either flat or at heel strike, usually accompanied by **shortened steps, reduced arm swing and forward flexed posture**.



- Activities are carried out **continuously**, not in isolation.
    - Lying down (30 seconds)
    - Sitting (30 seconds)
    - Standing (30 seconds)
    - Sit-to-Stand (5 times)
    - Slow walking (0.4-0.5 m/s) - (self-paced)
    - Normal (self-paced) walking
    - Brisk walking (0.8 m/s) - (self-paced)
    - Stairs (ascending/descending – 12 steps)

Appendix D

Manuscript submitted to EMBC 2020

Capturing accelerometer outputs in healthy volunteers under normal and simulated-pathological conditions using ML classifiers*

Filippou. V., Redmond. A.C., Bennion. J., Backhouse. M.R., and Wong. D.

Abstract— **Wearable devices offer a possible solution for acquiring objective measurements of physical activity. Most current algorithms are derived using data from healthy volunteers. It is unclear whether such algorithms are suitable in specific clinical scenarios, such as when an individual has altered gait. We hypothesized that algorithms trained on healthy population will result in less accurate results when tested in individuals with altered gait. We further hypothesized that algorithms trained on simulated-pathological gait would prove better at classifying abnormal activity.**

We studied healthy volunteers to assess whether activity classification accuracy differed for those with healthy and simulated-pathological conditions. Healthy participants (n=30) were recruited from the University of Leeds to perform nine pre-defined activities under healthy and simulated-pathological conditions. Activities were captured using a wrist-worn MOX accelerometer (Maastricht Instruments, NL). Data were analyzed based on the Activity-Recognition-Chain process. We trained a Neural-Network, Random-Forests, k-Nearest-Neighbors (k-NN), Support-Vector-Machines (SVM) and Naive Bayes models to classify activity. Algorithms were trained four times; once with 'healthy' data, and once with 'simulated-pathological data' for each of activity-type and activity-task classification.

In activity-type instances, the SVM provided the best results; the accuracy was 98.4% when the algorithm was trained and then tested with unseen data from the same group of healthy individuals. Accuracy dropped to 52.8% when tested on simulated-pathological data. When the model was retrained with simulated-pathological data, prediction accuracy for the corresponding test set was 96.7%. Algorithms developed on healthy data are less accurate for pathological conditions. When evaluating pathological conditions, classifier algorithms developed using data from a target sub-population can restore accuracy to above 95%.

Clinical Relevance— **This method remotely establishes health-related data of objective outcome measures of activities of daily living.**

I. Introduction

Physical activity (PA) significantly influences people's health and well-being, and helps prevent and delay onset of several chronic non-communicable diseases [1]. Several methods have been used previously to measure levels of activity in people. Such methods include large and expensive laboratory systems [2], and inexpensive, but time-consuming, subjective measures such as questionnaires, surveys and diaries [3].

Recent advances in commercial wearable technology has led to multiple devices that can enable PA to be assessed objectively. Of these, the accelerometer is commonly used for quantifying activity intensity and counting the number of steps [4]. Accelerometers are inexpensive, easy to use and long-lasting. However, common algorithms, including those used in consumer devices, are designed to be accurate for an archetypal healthy user and so may not be representative of subgroups such as those with chronic diseases that affect gait [5], [6]. Research to date has used accelerometers to classify activities and number of steps in moderately healthy patient populations [7], [8].

Our aim was to carry out a proof of concept study to investigate the performance of activity recognition algorithms using accelerometer data when trained on healthy individuals, but tested under healthy as well as unusual (*simulated-pathological*) gait conditions. We used a simulated-pathological condition, since recruiting actual patients was considered infeasible and impractical, especially given the exploratory nature of the current work.

We hypothesized that automated algorithms trained to identify types of physical activities in healthy participants would perform less well on participants when simulating a pathological gait.

II. Methods

A. Recruitment process

Participants were recruited via email and word of mouth from the staff and students of the University of Leeds. Participants were considered eligible for inclusion if they could walk freely without pain for two minutes. All participants were healthy, without any musculoskeletal condition or any condition affecting their gait. Participants 18+ years of age were recruited and, all participants gave informed written consent. Local ethical approval was provided by the University of Leeds (Ref #: MREC16-172).

B. Data acquisition

1) Data Sources

Each participant wore a MOX tri-axial accelerometer (Maastricht Instruments, Maastricht, NL) (dimensions:

*This study was supported by the Engineering and Physical Sciences Research Council (EPSRC) Centre for Doctoral Training in Tissue Engineering and Regenerative Medicine—EP/L014823/1.

Filippou. V is with the Institute of Medical and Biological Engineering, University of Leeds, Leeds, UK (phone: 07500481379; e-mail: mn12vf@leeds.ac.uk).

Redmond. A is with the Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds, Leeds, UK(email: A.Redmond@leeds.ac.uk).

Backhouse. M.R. is with the York Trials Unit, University of York, York, UK (email: mike.backhouse@york.ac.uk).

Wong. D is with the Centre for Health Informatics and Department of Computer Science, University of Manchester, Manchester, UK (email: david.wong@manchester.ac.uk).

Jacqueline. B is with the Royal Free London NHS Foundation Trust, London, UK (email: Jacqueline.Bennion@nhs.net)

35×35×10mm, weight: 11g). The device was held in place on the non-dominant wrist by an elasticated strap. The accelerometer had a measurement range of ±8g and a sampling frequency of 100 Hz. Recorded signals were stored locally on the accelerometer's internal memory (2GB) as a binary file that was downloaded upon the completion of each participant trial.

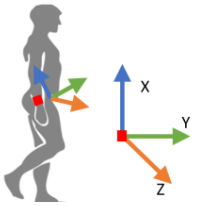


Figure 2: Accelerometer location and axis orientation

Our gold standard was a video recording of each participant. We used slow motion playback of videos to label the accelerometer data with the number of steps and to define the start and end time of each activity. This was cross-verified by an independent observer three times. The camera followed at approximately 2m from the participants.

2) Experimental protocol and set-up

Before attaching the activity monitor, participants were instructed that they would be performing nine activities: lie down, sit, stand, stand-to-sit, slow walk, normal walk, fast walk, walk upstairs, walk downstairs. Upon monitor attachment, the participant was asked to jump once to facilitate alignment of the video and accelerometer. After the jump, the participant performed the nine activities sequentially, and was reminded of each task. Participants were asked to jump once again after activities had been completed.

Each set of activities were performed twice, once under healthy conditions, and once under simulated-pathological conditions. For the simulated-pathological conditions, participants were asked to repeat the series of activities using a shuffling gait and to perform the activities more slowly. A shuffling gait was defined as when the foot is moving forward at the time of initial contact or during mid-swing, with the foot either flat or at heel strike, usually accompanied by shortened steps, reduced arm swing and forward flexed posture [9]. Such gait is a common marker of diseases such as severe rheumatoid arthritis and stroke. A written description, figure and video of shuffling gait was given to the participants prior to data collection. Participants were free to practice before data acquisition began.

C. Data processing

1) Data extraction

The binary files from the accelerometer were imported into Python™ (v3.6) for analysis. The extracted text files contained three columns of acceleration data, representing acceleration along the three principal axes.

To reduce the impact of high frequency random noise generated during data capture (caused, for instance, by muscle contraction), the accelerometer signal was filtered using a 6th order Butterworth filter with a 3Hz cutoff. The frequency of human activity is between 0-20 Hz and almost all of the signal energy is contained below 3 Hz [10]–[12].

We then derived five continuous signals from the 3-axis accelerometer data: dynamic accelerations, total magnitude, jerk, angular velocity and inclination angles.

Dynamic accelerations were calculated by averaging the readings on each direction, and then subtracting the corresponding average value from the raw acceleration signal. *Total magnitude* was calculated as:

$$acc = \sqrt{x^2 + y^2 + z^2}$$

Jerk is the rate of change of acceleration. A first order approximation was estimated from the acceleration signal as:

$$jerk = (acc_{t+T} - acc_t)/T$$

Where T is the sampling period. *Angular velocity* was identified by calculating the angle between the acceleration vectors in the current and the previous point. The accelerometer registers the data at equal time intervals. Therefore the angle between the vectors provides the angular velocity :

$$\cos(i, i+1) = \frac{(x_i x_{i+1} + y_i y_{i+1} + z_i z_{i+1})}{(\sqrt{x_i^2 + y_i^2 + z_i^2} \times \sqrt{x_{i+1}^2 + y_{i+1}^2 + z_{i+1}^2})}$$

Inclination angle was calculated for each direction.

$$\phi_x = \arccos(x^2/acc)$$

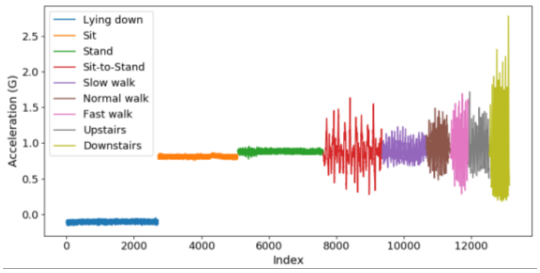


Figure 1: Time-series acceleration signal

The continuous data were split into a series of short time windows, in which the signal may be approximated as stationary. We used windows of 200 samples, corresponding to a time period of 2 seconds, exceeding the Nyquist limit required to detect slower gait and within the range of window lengths proposed in prior research [13].

Each window was manually labelled with a specific activity task and assigned to one of three broader activity types (static, dynamic, transition) using the video gold standard. Each activity task corresponded to an activity type. Dynamic activity tasks were slow walk, normal walk, fast walk, ascending and descending stairs. Static activity tasks were lying, sitting, standing. The transition activity type comprised the stand-to-sit task only.

2) Feature extraction and selection

From the acceleration time series in each window, we extracted a set of 120 summary features to represent the acceleration (x, y, z, t), jerk (x, y, z, t), angular velocity and inclination angle (x, y, z) signals. The features derived were time-domain (mean, standard deviation, kurtosis, skewness, root mean square, interquartile range, power spectral density) and frequency-domain (energy, max frequency, max 2nd frequency, mean frequency, entropy). A reduced number of linear combinations of these features were selected using principal component analysis (PCA). A cut-off total explained variance of 0.95 was set on the explained variance. By reducing the dimensionality of the feature set, we limited the risk of overfitting subsequent classification models. The features were reduced to 25 and 30 principal components for healthy and *simulated-pathological* groups respectively.

Table 1: Machine learning (ML) algorithm parameters

Parameters	Activity type		Activity task	
	H/H	S/S	H/H	S/S
k-NN (K neighbors)	4	4	4	4
NN (neurons)	35	55	60	75
RF (trees, min_samples_split*)	4, 12	4, 12	4, 12	4, 12
SVM (C, gamma)	1, 1	10, 1	1, 1	10, 1
* minimum number of data required to split an internal node				

The PCA feature set was then used as input to a selection of five machine learning classifiers: back-propagation Neural Networks (NN), Random Forests (RF), Support Vector Machines (SVM), k-Nearest Neighbours (kNN), and Naive Bayes (GB). These classifiers have been commonly used for clinical classification problems [14]–[16]. The parameters used for each algorithm are shown in table 1.

Each classifier was assessed on its ability to classify both activity type and activity task. We conducted three algorithmic scenarios:

1. trained on healthy data; tested on healthy data (H/H)
2. trained on *simulated-pathological* data; tested on *simulated-pathological* data (S/S)
3. trained on healthy data; tested on *simulated-pathological* data (H/S).

Performance was assessed using accuracy [14]. For scenarios (1) and (2), performance was estimated using 10-fold cross validation, and we report the mean performance. For scenario (3) all relevant data were used for training and testing.

III. Results

The mean age of participants was 32.7 years (s.d 12.7). Of the 30 participants, 14 identified as female. Their mean height was 171.5 cm (s.d 7.1) and their mean weight was 69.2 kg (s.d 13.6).

The highest level of accuracy for activity classification was achieved using SVM and k-NN in activity-type and activity-task groups respectively (Table 2). All ML approaches demonstrated higher accuracies for the broader activity-type identification than for specific activity-task identification. The SVM and k-NN classifiers achieved an accuracy of 98.4% and 94.3% for activity-type and activity-task identification respectively in classifiers trained on healthy data (H/H). When these classifiers were applied to *simulated-pathological* data (H/S), to replicate real world use of wearable accelerometers, accuracy fell between 31.3%-52.8%. Training the algorithms using simulated pathological data and then identifying simulated pathological activities (S/S) improved the accuracy to 96.7% and 84.5% for activity-type and activity-task identification respectively.

Confusion matrices are performance measurements which were developed to visualize accuracy and other metrics (figures 3-4). Figure 3 shows that static, stand-to-sit and slow walk activities achieved high individual recall scores, with lying achieving the highest recall score as 0.996. Fast walk obtained the worst recall performance, which was 0.796. In terms of the precision score, static, stand-to-sit and downstairs activities achieved scores greater than 0.940. Normal walk obtained the worst precision score which was 0.798. Figure 4 demonstrates that static activities had the three greatest recall scores, while lying, sitting and stand-to-sit had the three highest precision scores.

Table 2: Machine learning algorithm evaluation (accuracy)

ML algorithms	Group (Train/Test)		
	H/H	S/S	H/S
	Activity type: Static, Dynamic, Transition		
NN	0.983(0.982-0.983)	0.957 (0.956-0.958)	
RF	0.953 (0.952-0.954)	0.921 (0.920-0.923)	
k-NN	0.983 (0.982-0.983)	0.960 (0.959-0.961)	
GB	0.897 (0.896-0.898)	0.834 (0.832-0.836)	
SVM	0.984 (0.983-0.984)	0.967 (0.966-0.968)	0.528
	Activity task: Specific activities		
NN	0.926 (0.924-0.927)	0.770 (0.767-0.772)	
RF	0.873 (0.871-0.875)	0.689 (0.687-0.691)	
k-NN	0.943 (0.941-0.944)	0.845 (0.843-0.846)	0.313
GB	0.749 (0.746-0.751)	0.516 (0.514-0.518)	
SVM	0.926 (0.925-0.928)	0.838 (0.836-0.840)	

IV. Discussion

Earlier studies have attempted activity recognition using machine learning classifiers similar to those used here. In healthy volunteers, results were similar. All classifiers that were tested, except Naïve Bayes, had accuracies ranging from 68% to 98% [7], [14], [17]–[21]. Naïve Bayes provided poorer results than the other algorithms [14], [17]–[20].

Our results demonstrated high levels of accuracy when the classifier was trained and tested with data from a similar group. However, when the tested data (*simulated-pathological*) differed from the training data (healthy), the accuracy dropped dramatically.

The difference in mean accuracy is likely due to the fact that volunteers were asked to make significant changes to their motions under *simulated-pathological* conditions. Although we attempted to train participants to replicate compromised motion, we could not be certain that their movements accurately reflected real pathological motion. Indeed, participants may have interpreted the instructions on how to mimic the pathological activities slightly differently. This means that the accuracies reported can only be considered a reasonable initial estimate of the performance of ML algorithms on real patients.

Previous studies have assessed whether algorithms trained on data from healthy populations were suitable for pathological populations. They conclude, like us, that large differences between groups means that algorithms should be trained for specific target groups [22]–[24].

One potential limitation is that we have reported accuracies as our overall performance metric. It is well known that accuracy can be a poor metric of overall performance in the presence of unbalanced data.

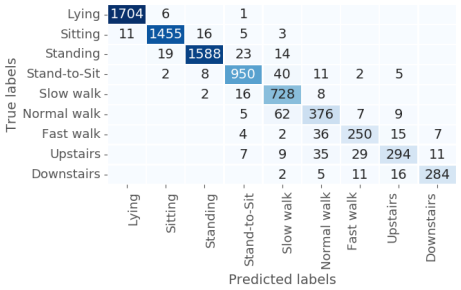


Figure 3: Confusion matrix of H/H group for tasks of activity

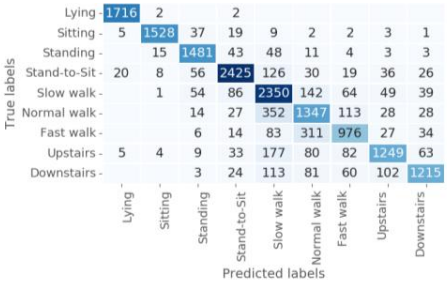


Figure 4: Confusion matrix of S/S group for tasks of activity

This problem was of lower concern here, in which quantity of each of activity was similar. Reporting the accuracy also allowed direct comparison to other work, and exact classifications are shown in table 2.

Another limitation in this study is human error for labeling the activities. Even though there was a gold standard video, the activity labeling was completed manually and subject to human error. However, the authors ensured thorough steps were taken to minimize this by using slow motion analysis, replaying analysis and triple counting each activity set.

One aspect of the study was to act as a baseline for developing activity classifiers for the healthy population. These classifiers will be further updated to suit the pathological population with walking impairments, and used by clinicians to evaluate the daily activity performance of chronic condition patients. Clinicians will be able to have a more informed view about the activity of their patients, hence provide them with better and patient-specific treatment plans and medications.

Additionally, a range of different accelerometer devices could be used, and their results compared to check the accuracy of the devices and the wider utility of the ML approach.

V. Conclusion

In this study, we used five machine learning algorithms to classify nine daily living activities. Activities were performed by healthy and *simulated-pathological* conditions. Furthermore, activities were classified into two groups, general activity type and specific activity task. The SVM and k-NN classifiers outperformed all other algorithms in activity-type and activity-task classifications respectively. All algorithms performed well when the training and test sets both came from the same population. Conversely, when the algorithms were trained with healthy data and tested with simulated-pathological data, as would usually occur in the real-world, the accuracy demonstrated was poor. It may therefore be possible to develop more accurate and clinically useful activity classification algorithms based on a person's or a sub-population's accelerometer gait signal.

References

[1] S. M. Phillips, L. Cadmus-Bertram, D. Rosenberg, M. P. Buman, and B. M. Lynch, "Wearable technology and physical activity in chronic disease: opportunities and challenges," *Am. J. Prev. Med.*, vol. 54, no. 1, pp. 144–150, 2017.

[2] S. J. Strath *et al.*, "Guide to the assessment of physical activity: Clinical and research applications," *Circulation*, vol. 128, no. 20, pp. 2259–2279, 2013.

[3] J. M. Broderick, J. Ryan, D. M. O. Donnell, and J. Hussey, "A guide to assessing physical activity using accelerometry in cancer patients," *Support Care Cancer*, vol. 22, no. 4, pp. 1121–1130, 2014.

[4] R. K. Walker, A. M. Hickey, and P. S. Freedson, "Advantages and limitations of wearable activity trackers: Considerations for patients and clinicians," *Clin. J. Oncol. Nurs.*, 2016.

[5] M. R. Backhouse, E. M. A. Hensor, D. White, A. Keenan, P. S. Helliwell, and A. C. Redmond, "Concurrent validation of activity monitors in patients with rheumatoid arthritis," *Clin. Biomech.*, vol. 28, no. 4, pp. 473–479, 2013.

[6] P. J. Mancuso, M. Thompson, M. Tietze, S. Kelk, and G. Roux, "Can patient use of daily activity monitors change nurse practitioner practice?," *J. Nurse Pract.*, vol. 10, no. 10, pp. 787–793, 2014.

[7] A. Mannini, S. S. Intille, M. Rosenberger, A. M. Sabatini, and W. Haskell, "Activity recognition using a single accelerometer placed at the wrist or ankle," *Med Sci Sport. Exerc*, vol. 45, no. 11, pp. 2193–2203, 2013.

[8] X. Kang, B. Huang, and G. Qi, "A novel walking detection and step counting algorithm using unconstrained smartphones," *Sensors*, vol. 18, no. 1, pp. 297–311, 2018.

[9] B. Salzman, "Gait and balance disorders in older adults," *Am. Fam. Physician*, vol. 82, no. 1, pp. 61–68, 2010.

[10] W. Z. Wang, B. Y. Huang, and L. Wang, "Analysis of filtering methods for 3D acceleration signals in body sensor network," *Int. Symp. Bioelectron. Bioinforma.*, pp. 263–266, 2011.

[11] M. Merryn, "Monitoring and interpreting human movement patterns using a triaxial accelerometer," 2003.

[12] E. K. Antonsson and R. W. Mann, "The frequency content of gait," *J. Biomech.*, vol. 18, no. 1, pp. 39–47, 1985.

[13] O. Banos, J.-M. Galvez, M. Damas, H. Pomares, and I. Rojas, "Window size impact in human activity recognition," *Sensors*, vol. 14, no. 4, pp. 6474–6499, 2014.

[14] I. Cleland *et al.*, "Optimal placement of accelerometers for the detection of everyday activities," *Sensors (Basel).*, vol. 13, no. 7, pp. 9183–9200, 2013.

[15] B. Erdaş, I. Atasoy, K. Açici, and H. Oğul, "Integrating features for accelerometer-based activity recognition," *Procedia Comput. Sci.*, vol. 58, pp. 522–527, 2016.

[16] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, 2008.

[17] A. Mannini and A. M. Sabatini, "Machine learning methods for classifying human physical activity from on-body accelerometers," *Sensors*, vol. 10, no. 2, pp. 1154–1175, 2010.

[18] Y. Saez, A. Baldominos, and P. Isasi, "A comparison study of classifier algorithms for cross-person physical activity recognition," *Sensors*, vol. 17, no. 1, pp. 66–91, 2017.

[19] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, "How accurately can your wrist device recognize daily activities and detect falls?," *Sensors (Switzerland)*, vol. 16, no. 6, pp. 800–820, 2016.

[20] S. Zhang, A. V Rowlands, P. Murray, and T. L. Hurst, "Physical activity classification using the GENEA wrist-worn accelerometer," *Med. Sci. Sport. Exerc.*, vol. 44, no. 4, pp. 742–748, 2012.

[21] K. Lee and M. P. Kwan, "Physical activity classification in free-living conditions using smartphone accelerometer data and exploration of predicted results," *Comput. Environ. Urban Syst.*, vol. 67, pp. 124–131, 2018.

[22] L. Lonini, A. Gupta, K. Kording, and A. Jayaraman, "Activity recognition in patients with lower limb impairments: Do we need training data from each patient?," *Eng. Med. Biol. Soc.*, pp. 3265–3268, 2016.

[23] N. A. Capela, E. D. Lemaire, and N. Baddour, "Feature selection for wearable smartphone-based human activity recognition with able bodied, elderly, and stroke patients," *PLoS One*, vol. 10, no. 4, pp. 1–18, 2015.

[24] M. B. Del Rosario *et al.*, "A comparison of activity classification in younger and older cohorts using a smartphone," *Physiol. Meas.*, vol. 35, no. 11, pp. 2269–2286, 2014.

References

Abdull Sukor, A. S., Zakaria, A., and Abdul Rahim, N. (2018). "Activity recognition using accelerometer sensor and machine learning classifiers". In: *2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA)*, pp. 233–238. DOI: 10.1109/CSPA.2018.8368718.

Agarana, M C. and Akinlabi, E T. (2018). "Mathematical modelling and analysis of human arm as a triple pendulum system using Euler - Lagragian model". In: *IOP Conference Series: Materials Science and Engineering* 413, p. 012010. ISSN: 1757899X. DOI: 10.1088/1757-899X/413/1/012010.

Ailisto, H J., Lindholm, M., Mantyjarvi, J., Vildjiounaite, E., and Makela, S. (2005). "Identifying people from gait pattern with accelerometers". In: *Proceedings of SPIE - The International Society for Optical Engineering* March, p. 7. ISSN: 0277786X. DOI: 10.1117/12.603331. URL: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.603331.

Al-Eidan, R M., Al-Khalifa, H., and Al-Salman, A. (2018). "A review of wrist-worn wearable: sensors, models, and challenges". In: *Journal of Sensors* 2018. ISSN: 16877268. DOI: 10.1155/2018/5853917.

Al-zu, L A., Al-tamimi, A A., Al-momani, T D., Alkarala, A J., and Alzawahreh, M A. (2012). "Modeling and simulating human arm movement using a 2 dimensional 3 segments coupled pendulum System". In: *International Conference on Biomedical Engineering (ICBE 2012)* 71, pp. 1372–1377.

Alaqtash, M., Sarkodie-Gyan, T., Yu, H., Fuentes, O, Brower, R., and Abdelgawad, A. (2011). "Automatic classification of pathological gait patterns using ground reaction forces and machine learning algorithms". In: *Annu Int Conf IEEE Eng Med Biol Soc* 2011, pp. 453–457. DOI: 10.1109/IEMBS.2011.6090063.

Alharbi, F., Ouarbya, L., and Ward, J A. (2020). "Synthetic sensor data for human activity recognition". In: *Proceedings of the International Joint Conference on Neural Networks*. DOI: 10.1109/IJCNN48605.2020.9206624.

Alinia, P., Cain, C., Fallahzadeh, R., Shahrokni, A., Cook, D., and Ghasemzadeh, H. (2017). "How accurate is your activity tracker? A comparative study of step counts in low-intensity physical activities". In: *JMIR Mhealth Uhealth* 5.8, e106. DOI: 10.2196/mhealth.6321.

Almasi, A., Shamsollahi, M., and Senhadji, L. (2011). "A dynamical model for generating synthetic Phonocardiogram signals". In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 2011, pp. 5686–5689. DOI: 10.1109/IEMBS.2011.6091376.

Alzantot, M., Chakraborty, S., and Srivastava, M. (2017). "SenseGen: a deep learning architecture for synthetic sensor data generation". In: *2017 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2017*, pp. 188–193. DOI: 10.1109/PERCOMW.2017.7917555.

Andreu-Perez, J., Garcia-Gancedo, L., McKinnell, J., Van der Drift, A., Powell, A., Hamy, V., Keller, T., and Yang, G. (2017). "Developing fine-grained actigraphies for rheumatoid arthritis patients from a single accelerometer using machine learning". In: *Sensors (Switzerland)* 17.9, p. 2113. DOI: 10.3390/s17092113.

Ann, O. and Theng, L. (2014). "Human activity recognition: a review". In: *2014 IEEE International Conference on Control System, Computing and Engineering (ICCSCE 2014)*, pp. 389–393. DOI: 10.1109/ICCSCE.2014.7072750.

Anwary, A., Yu, H., and Vassallo, M. (2018). "An automatic gait feature extraction method for identifying gait asymmetry using wearable sensors". In: *Sensors (Basel)* 18.2, p. 676. DOI: 10.3390/s18020676.

Ao, B., Wang, Y., Liu, H., Li, D., Song, L., and Li, J. (2018). "Context impacts in accelerometer-based walk detection and step counting". In: *Sensors (Basel, Switzerland)* 18.11, p. 3604. DOI: 10.3390/s18113604.

Arif, M., Kattan, A, and Ahamed, S I. (2017). "Classification of physical activities using wearable sensors". In: *Intelligent Automation and Soft Computing* 23.1, pp. 21–30. DOI: 10.1080/10798587.2015.1118275.

Atallah, L., Lo, B., King, R., and Yang, G. (2010). "Sensor placement for activity detection using wearable accelerometers". In: *2010 International Conference on Body Sensor Networks*, pp. 24–29. DOI: 10.1109/BSN.2010.23.

Atallah, L., Lo, B., King, R., and Yang, G Z. (2011). "Sensor positioning for activity recognition using wearable accelerometers". In: *IEEE Transactions on Biomedical Circuits and Systems* 5.4, pp. 320–329. DOI: 10.1109/TBCAS.2011.2160540.

Attal, F., Mohammed, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L., and Amirat, Y. (2015). "Physical human activity recognition using wearable sensors". In: *Sensors* 15.12, pp. 31314–31338. DOI: 10.3390/s151229858.

Avci, A., Bosch, S., Marin-Perianu, M., Marin-Perianu, R., and Havinga, P. (2010). "Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: a survey". In: *23th International Conference on Architecture of Computing Systems 2010*, pp. 1–10. DOI: 10.1007/978-3-319-13105-4_17.

Backhouse, M R., Hensor, E M A., White, D., Keenan, A., Helliwell, P S., and Redmond, A C. (2013). "Concurrent validation of activity monitors in patients with rheumatoid arthritis". In: *Clinical Biomechanics (Bristol, Avon)* 28.4, pp. 473–479. DOI: 10.1016/j.clinbiomech.2013.02.009.

Baig, M M., GholamHosseini, H., Moqeem, A A., Mirza, F., and Lindén, M. (2017). "A systematic review of wearable patient monitoring systems – current challenges and opportunities for clinical adoption". In: *Journal of Medical Systems* 41.7, pp. 1–27. DOI: 10.1007/s10916-017-0760-1.

Banos, O., Damas, M., Pomares, H., Rojas, F., Delgado-Marquez, B., and Valenzuela, O. (2013). "Human activity recognition based on a sensor weighting hierarchical classifier". In: *Soft Computing* 17, pp. 333–343. DOI: 10.1007/s00500-012-0896-3.

Banos, O., Galvez, J., Damas, M., Pomares, H., and Rojas, I. (2014). "Window size impact in human activity recognition". In: *Sensors (Basel, Switzerland)* 14.4, pp. 6474–6499. DOI: 10.3390/s140406474.

Bao, L. and Intille, S S. (2004). "Activity recognition from user-annotated acceleration data". In: *Pervasive Computing*, pp. 1–17. DOI: 10.1007/978-3-540-24646-6_1.

Bassett, D R., Toth, L P., LaMunion, S R., and Crouter, S E. (2017). "Step counting: a review of measurement considerations and health-related applications". In: *Sports Medicine* 47.7, pp. 1303–1315. DOI: 10.1007/s40279-016-0663-1.

Bassett, D R., Troiano, R P., Mcclain, J J., and Wolff, D L. (2015). "Accelerometer-based physical activity: total volume per day and standardized measures". In: *Medicine and Science in Sports and Exercise* 47.4, pp. 833–838. DOI: 10.1249/MSS.0000000000000468.

Batista, G E A P A. and Monard, M C. (2003). "An analysis of four missing data treatment methods for supervised learning". In: *Applied Artificial Intelligence* 17.5-6, pp. 519–533. DOI: 10.1080/713827181.

Beevi, F H A., Miranda, J., Pedersen, C F., and Wagner, S. (2016). "An evaluation of commercial pedometers for monitoring slow walking speed populations". In: *Telemed J E Health* 22.5, pp. 441–449. DOI: 10.1089/tmj.2015.0120.

Bemdt, D J. and Clifford, J. (1994). "Using dynamic time warping to find patterns in time series". In: *AAAIWS'94: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, pp. 359–370.

BHF (2017). *Physical inactivity and sedentary behaviour report 2017*. Tech. rep., p. 10. URL: https://www.bhf.org.uk/publications/statistics/physical-inactivity-report-2017.

Bland, J M. and Altman, D G. (1999). "Measuring agreement in method comparison studies". In: *Stat Methods Med Res* 8.2, pp. 135–160. DOI: 10.1177/096228029900800204.

Booth, F W., Roberts, C K., and Laye, M J. (2012). "Lack of exercise is a major cause of chronic diseases". In: *Comprehensive Physiology* 2.2, pp. 1143–1211. DOI: 10.1002/cphy.c110025.

Boser, B E., Guyon, I M., and Vapnik, V N. (1992). "A training algorithm for optimal margin classifiers". In: *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152. DOI: 10.1145/130385.130401.

Broderick, J M., Ryan, J., Donnell, D M O., and Hussey, J. (2014). "A guide to assessing physical activity using accelerometry in cancer patients". In: *Support Care Cancer* 22.4, pp. 1121–1130. DOI: 10.1007/s00520-013-2102-2.

Bryson, J M. (2004). "What to do when Stakeholders matter". In: *Public Management Review* 6.1, pp. 21–53. DOI: 10.1080/14719030410001675722.

Buchowski, M S. (2014). "Doubly labeled water is a validated and verified reference standard in nutrition research". In: *Journal of nutrition* 144.5, pp. 573–574. DOI: 10.3945/jn.114.191361.

Bui, D T., Nguyen, N D., and Jeong, G M. (2018). "A robust step detection algorithm and walking distance estimation based on daily wrist activity recognition using a smart band". In: *Sensors (Basel)* 18.7, p. 2034. DOI: 10.3390/s18072034.

Bulling, A., Blanke, U., and Schiele, B. (2014). "A tutorial on human activity recognition using body-worn inertial sensors". In: *ACM Computing Surveys (CSUR)* 33, pp. 1–33. DOI: http://dx.doi.org/10.1145/2499621.

Bunn, J A., Jones, C., Oliviera, A, and Webster, M J. (2019). "Assessment of step accuracy using the consumer technology association standard". In: *Journal of Sports Sciences* 37.3, pp. 244–248. DOI: 10.1080/02640414.2018.1491941.

Bunn, J A., Navalta, J W., Fountaine, C J., and Reece, J D. (2018). "Current state of commercial wearable technology in physical activity monitoring 2015-2017". In: *International Journal of Exercise Science* 11.7, pp. 503–515. URL: http://www.intjexersci.com.

Butler, P J., Green, J A., Boyd, I L., and Speakman, J R. (2004). "Measuring meatabolic rate in the fiels: the pros and cons of the doubly labeled water and heart rate methods". In: *Functional Ecology* 18.2, pp. 168–183. DOI: 10.1111/j.0269-8463.2004.00821.x.

Butte, N F., Ekelund, U., and Westerterp, K R. (2012). "Assessing physical activity using wearable monitors: measures of physical activity". In: *Medicine and Science in Sports and Exercise* 44.1 SUPPL 1, S5–12. DOI: 10.1249/MSS.0b013e3182399c0e.

Calabró, M A., Lee, J., Saint-Maurice, P F., Yoo, H., and Welk, G J. (2014). "Validity of physical activity monitors for assessing lower intensity activity in adults". In: *Int J Behav Nutr Phys Act* 11, p. 119. DOI: 10.1186/s12966-014-0119-7.

Capela, N A., Lemaire, E D., and Baddour, N. (2015a). "Feature selection for wearable smartphone-based human activity recognition with able bodied, elderly, and stroke patients". In: *PLoS ONE* 10.4, e0124414. DOI: 10.1371/journal.pone.0124414.

– (2015b). "Novel algorithm for a smartphone-based 6-minute walk test application: algorithm, application development, and evaluation". In: *Journal of neuroengineering and rehabilitation* 12, p. 19. DOI: 10.1186/s12984-015-0013-9.

Caron, A., Ayala, A., Damián, J., Rodriguez-Blazquez, C., Almazán, J., Castellote, J M., Comin, M., Forjaz, M J., and Pedro, J. de (2017). "Physical activity, body functions and disability among middle-aged and older Spanish adults". In: *BMC Geriatrics* 17.1, p. 150. DOI: 10.1186/s12877-017-0551-z.

Caspersen, C J., Powell, K E., and Christenson, G M. (1985). "Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research". In: *Public health reports (Washington, D.C. : 1974)* 100.2, pp. 126–31. DOI: 10.2307/20056429.

Castillo, J C., Carneiro, D., Serrano-Cuerda, J., Novais, P., Fernández-Caballero, A., and Neves, J. (2014). "A multi-modal approach for activity classification and fall detection". In: *International Journal of Systems Science* 45.4, pp. 810–824. DOI: 10.1080/00207721.2013.784372.

Chandel, V., Choudhury, A D., Ghose, A., and Bhaumik, C. (2014). "AcTrak-unobtrusive activity detection and step counting using smartphones". In: *Mobile and Ubiquitous Systems: Computing, Networking, and Services. MobiQuitous 2013. Lecture Notes of the Institute for*

Computer Sciences, Social Informatics and Telecommunications Engineering 131, pp. 447–459. DOI: 10.1007/978-3-319-11569-6_35.

Chandrasekar, A., Hensor, E M A., Mackie, S L., Backhouse, M R., and Harris, E. (2018). "Preliminary concurrent validity of the Fitbit-Zip and ActiGraph activity monitors for measuring steps in people with polymyalgia rheumatica". In: *Gait and Posture* 61, pp. 339–345. DOI: 10.1016/j.gaitpost.2018.01.035.

Chernbumroong, S., Atkins, A S., and Hongnian, Y. (2011). "Activity classification using a single wrist-worn accelerometer". In: *2011 5th International Conference on Software, Knowledge Information, Industrial Management and Applications (SKIMA) Proceedings*, pp. 1–6. DOI: 10.1109/SKIMA.2011.6089975.

Cho, Y., Cho, H., and Kyung, C. (2016). "Design and implementation of practical step detection algorithm for wrist worn devices". In: *IEEE Sensors Journal* 16.21, pp. 7720–7730. DOI: 10.1109/JSEN.2016.2603163.

Chow, J J., Thom, J M., Wewege, M A., Ward, R E., and Parmenter, B J. (2017). "Accuracy of step count measured by physical activity monitors: the effect of gait speed and anatomical placement site". In: *Gait Posture* 57, pp. 199–203. DOI: 10.1016/j.gaitpost.2017.06.012.

Chowdhury, A K., Tjondronegoro, D., Chandran, V., and Trost, S G. (2018). "Physical activity recognition using posterior-adapted class-based fusion of multiaccelerometer data". In: *IEEE Journal of Biomedical and Health Informatics* 22.3, pp. 678–685. DOI: 10.1109/JBHI.2017.2705036.

Chowdhury, E A., Western, M J., Nightingale, T E., Peacock, O J., and Thompson, D. (2017). "Assessment of laboratory and daily energy expenditure estimates from consumer multisensor physical activity monitors". In: *PLoS ONE* 12.2, pp. 1–15. DOI: 10.1371/journal.pone.0171720.

Chu, A H Y., Ng, S H X., Paknezhad, M., Gauterin, Al., Koh, D., Brown, M S., and Müller-Riemenschneider, F. (2017). "Comparison of wrist-worn Fitbit Flex and waist-worn ActiGraph for measuring steps in free-living adults". In: *PLoS ONE* 12.2. DOI: 10.1371/journal.pone.0172535.

Cleland, I., Kikhia, B., Nugent, C., Boytsov, A., Hallberg, J., Synnes, K., McClean, S., and Finlay, D. (2013). "Optimal placement of accelerometers for the detection of everyday activities". In: *Sensors (Basel)* 13.7, pp. 9183–9200. DOI: 10.3390/s130709183.

Clifford, G D. (2006). "A novel framework signal representation and source separation: Applications to filtering and segmentation of biosignals". In: *Journal of Biological Systems* 14.2, pp. 169–183. DOI: 10.1142/S0218339006001830.

Cola, G., Avvenuti, M., Musso, F., and Vecchio, A. (2016). "Gait-based authentication using a wrist-worn device". In: *MOBIQUITOUS 2016: Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pp. 208–217. DOI: 10.1145/2994374.2994393.

– (2017). "Personalized gait detection using a wrist-worn accelerometer". In: *2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks, BSN 2017*, pp. 173–177. DOI: 10.1109/BSN.2017.7936035.

Cola, G., Avvenuti, M., Vecchio, A., Yang, G Z., and Lo, B. (2015). "An on-node processing approach for anomaly detection in gait". In: *IEEE Sensors Journal* 15.11, pp. 6640–6649. DOI: 10.1109/JSEN.2015.2464774.

Collins, S H. and Kuo, A D. (2013). "Two independent contributions to step variability during over-ground human walking". In: *PLoS ONE* 8.8, e73597. DOI: 10.1371/journal.pone.0073597.

Culhane, K M., O'Connor, M., Lyons, D., and Lyons, G M. (2005). "Accelerometers in rehabilitation medicine for older adults". In: *Age Ageing* 34.6, pp. 556–560. DOI: 10.1093/ageing/afi192.

Del Din, S., Elshehabi, M., Galna, B., Hobert, M A., Warmerdam, E., Suenkel, U., Brockmann, K., Metzger, F., Hansen, C., Berg, D., Rochester, L., and Maetzler, W. (2019). "Gait analysis with wearables predicts conversion to parkinson disease". In: *Annals of Neurology* 86.3, pp. 357–367. ISSN: 15318249. DOI: 10.1002/ana.25548.

Del Rosario, M B., Wang, K., Wang, J., Liu, Y., Brodie, M., Delbaere, K., Lovell, N H., Lord, S R., and Redmond, S J. (2014). "A comparison of activity classification in younger and

older cohorts using a smartphone". In: *Physiological Measurement* 35.11, pp. 2269–2286. DOI: 10.1088/0967-3334/35/11/2269.

Dirican, A C. and Aksoy, S. (2017). "Step counting using smartphone accelerometer and fast Fourier transform". In: *Sigma J Eng & Nat Sci* 8.2, pp. 175–182.

Dumith, S C., Hallal, P C., Reis, R S., and Kohl, H W. (2011). "Worldwide prevalence of physical inactivity and its association with human development index in 76 countries". In: *Preventive Medicine* 53.1-2, pp. 24–28. DOI: 10.1016/j.ypmed.2011.02.017.

Durstine, J L., Gordon, B., Wang, Z., and Luo, X. (2013). "Chronic disease and the link to physical activity". In: *Journal of Sport and Health Science* 2.1, pp. 3–11. DOI: 10.1016/j.jshs.2012.07.009.

Dutta, A., Ma, O., Toledo, M., Pregonero, A Fl., Ainsworth, B E., Buman, M P., and Bliss, D W. (2018). "Identifying free-living physical activities using lab-based models with wearable accelerometers". In: *Sensors (Switzerland)* 18.11, pp. 1–14. DOI: 10.3390/s18113893.

Ehrler, F., Weber, C., and Lovis, C. (2016). "Influence of pedometer position on pedometer accuracy at various walking speeds: a comparative study". In: *Journal of Medical Internet Research* 18.10, pp. 1–9. DOI: 10.2196/jmir.5916.

Erdaş, B., Atasoy, I., Açici, K., and Oğul, H. (2016). "Integrating features for accelerometer-based activity recognition". In: *Procedia Computer Science* 98, pp. 522–527. DOI: 10.1016/j.procs.2016.09.070.

Esfahani, M I M. and Nussbaum, M A. (2019). "Using smart garments to differentiate among normal and simulated abnormal gaits". In: *Journal of Biomechanics* 93, pp. 70–76. DOI: 10.1016/j.jbiomech.2019.06.009.

Esser, P., Dawes, H., Collett, J., Feltham, M G., and Howells, K. (2011). "Assessment of spatio-temporal gait parameters using inertial measurement units in neurological populations". In: *Gait and Posture* 34.4, pp. 558–560. ISSN: 09666362. DOI: 10.1016/j.gaitpost.2011.06.018. URL: http://dx.doi.org/10.1016/j.gaitpost.2011.06.018.

Feng, Y., Wong, C K., Janeja, V., Kuber, R., and Mentis, H M. (2017). "Comparison of tri-axial accelerometers step-count accuracy in slow walking conditions". In: *Gait Posture* 53, pp. 11–16. DOI: 10.1016/j.gaitpost.2016.12.014.

Fortune, E., Lugade, V., Amin, S., and Kaufman, K R. (2015). "Step detection using multi- versus single tri-axial accelerometer-based systems". In: *Physiological Measurement* 36.12, pp. 2519–2535. DOI: 10.1088/0967-3334/36/12/2519.

Fortune, E., Lugade, V., Morrow, M., and Kaufman, K. (2014). "Validity of using tri-axial accelerometers to measure human movement - Part II: step counts at a wide range of gait velocities". In: *Medical Engineering and Physics* 36.6, pp. 659–669. DOI: 10.1016/j.medengphy.2014.02.006.

Freeman, R E. (1984). *Strategic management: a stakeholder approach*. Vol. 1, p. 276. DOI: 10.2139/ssrn.263511.

Fulk, G D., Combs, S A., Danks, K A., Nirider, C D., Raja, B., and Reisman, D S. (2014). "Accuracy of 2 activity monitors in detecting steps in people with stroke and traumatic brain injury". In: *Physical Therapy* 94.2, pp. 222–229. DOI: 10.2522/ptj.20120525.

Gemperle, F., Kasabach, C., Stivoric, J., Bauer, M., and Martin, R. (1998). "Design for wearability". In: *Digest of Papers. Second International Symposium on Wearable Computers (Cat. No.98EX215)*, pp. 116–122. DOI: 10.1109/ISWC.1998.729537.

Genovese, V., Mannini, A., and Sabatini, A M. (2017). "A smartwatch step counter for slow and intermittent ambulation". In: *IEEE Access* 5, pp. 13028–13037. DOI: 10.1109/ACCESS.2017.2702066.

Giavarina, D. (2015). "Understanding Bland Altman analysis". In: *Biochemia Medica* 25.2, pp. 141–151. DOI: 10.11613/BM.2015.015.

Gilmore, S J., Davidson, M., Hahne, A J., and McClelland, J A. (2020). "The validity of using activity monitors to detect step count after lumbar fusion surgery". In: *Disability and Rehabilitation* 42.6, pp. 863–868. DOI: 10.1080/09638288.2018.1509140.

Gjoreski, H. and Gams, M. (2011). "Accelerometer data preparation for activity recognition". In: *International Multiconference Information Society* 1014, p. 1014.

Gjoreski, M., Gjoreski, H., Luštrek, M., and Gams, M. (2016). "How accurately can your wrist device recognize daily activities and detect falls?" In: *Sensors (Basel)* 16.6, p. 800. DOI: 10.3390/s16060800.

Godfrey, A., Conway, R., Meagher, D., and ÓLaighin, G. (2008). "Direct measurement of human movement by accelerometry". In: *Medical Engineering and Physics* 30.10, pp. 1364–1386. DOI: 10.1016/j.medengphy.2008.09.005.

Godfrey, A., Morris, R., Hickey, A., and Del Din, S. (2016). "Beyond the front end: investigating a thigh worn accelerometer device for step count and bout detection in Parkinson's disease". In: *Med Eng Phys* 38.12, pp. 1524–1529. DOI: 10.1016/j.medengphy.2016.09.023.

Gomes, A L G N. (2014). "Human activity recognition with accelerometry: novel time and frequency features". PhD thesis, p. 127. URL: https://run.unl.pt/bitstream/10362/14040/1/Gomes{_}2014.pdf.

González, K., Fuentes, J., and Márquez, J L. (2017). "Physical inactivity, sedentary behavior and chronic diseases". In: *Korean J Fam Med* 38.3, pp. 111–115. DOI: 10.4082/kjfm.2017.38.3.111.

Goodfellow, I J., Pouget-Abadie, J, Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). "Generative Adversarial Nets". In: DOI: 10.1109/ICCVW.2019.00369.

Gu, F., Khoshelham, K., Shang, J., Yu, F., and Wei, Z. (2017). "Robust and accurate smartphone-based step counting for indoor localization". In: *IEEE Sensors Journal* 17.11, pp. 3453–3460. DOI: 10.1109/JSEN.2017.2685999.

Gupta, Piyush Puneet and Dallas, Tim (2014). "Feature selection and activity recognition system using a single triaxial accelerometer". In: *Biomedical Engineering, IEEE Transactions on* 61.6, pp. 1780–1786. ISSN: 0018-9294. DOI: 10.1109/TBME.2014.2307069.

Haleem, A. (2008). "Stakeholder Analysis Report". In: November.

Hans, P. (2015). "Critical review of method comparison studies for the evaluation of estimating glomerular filtration rate equations". In: *Int J Nephrol Kidney Failure* 1.1. DOI: 10.16966/2380-5498.102.

Hassan, M M., Uddin, M Z., Mohamed, A., and Almogren, A. (2018). "A robust human activity recognition system using smartphone sensors and deep learning". In: *Future Generation Computer Systems* 81, pp. 307–313. DOI: 10.1016/j.future.2017.11.029.

Hassouni, A., Hoogendoorn, M., and Muhonen, V. (2018). "Using generative adversarial networks to develop a realistic human behavior simulator". In: *In: Miller T., Oren N., Sakurai Y., Noda I., Savarimuthu B., Cao Son T. (eds) PRIMA 2018: Principles and Practice of Multi-Agent Systems. PRIMA 2018. Lecture Notes in Computer Science* 11224. DOI: 10.1007/978-3-030-03098-8_32.

Hees, V T. van, Gorzelniak, L., Dean León, E C., Eder, M., Pias, M., Taherian, S., Ekelund, U., Renström, F., Franks, PW., Horsch, A., and Brage, S. (2013). "Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity". In: *PLoS ONE* 8.4. DOI: 10.1371/journal.pone.0061691.

Hills, A P., Mokhtar, N., and Byrne, N M. (2014). "Assessment of physical activity and energy expenditure: an overview of objective measures". In: *Frontiers in Nutrition* 1.5. DOI: 10.3389/fnut.2014.00005.

Ho, N H., Truong, P H., and Jeong, G M. (2016). "Step-detection and adaptive step-length estimation for pedestrian dead-reckoning at various walking speeds using a smartphone". In: *Sensors (Basel))* 16.9, p. 1423. DOI: 10.3390/s16091423.

Huang, Y., Zheng, H., Nugent, C., McCullagh, P., Black, N., Burns, W., Tully, M A., and McDonough, S M. (2012). "An orientation free adaptive step detection algorithm using a smart phone in physical activity monitoring". In: *Health and Technology* 2.4, pp. 249–258. DOI: 10.1007/s12553-012-0035-2.

Hyder, A., Syed, S., Puvanachandra, P., Bloom, G., Sundaram, S., Mahmood, S., Iqbal, M., Hongwen, Z., Ravichandran, N., Oladepo, O., Pariyo, G., and Peters, D. (2010). "Stakeholder

analysis for health research: case studies from low- and middle-income countries". In: *Public Health* 124.3, pp. 159–166. DOI: 10.1016/j.puhe.2009.12.006.

Ignatov, A. (2018). "Real-time human activity recognition from accelerometer data using convolutional neural networks". In: *Applied Soft Computing* 62, pp. 915–922. DOI: 10.1016/j.asoc.2017.09.027.

Ingen, T. van (2010). *Report on Stakeholder Analysis and Strategies for Stakeholder Engagement*. Tech. rep., p. 235. URL: http://www.wetwin.eu/downloads/D2-1.pdf.

Israel, G D. (1992). "Determining sample size". In: *Program Evaluation and Organizational Development, IFAS, University of Florida* Nov. ISSN: 1394195X.

Jang, Y., Kim, S., Kim, K., and Lee, D. (2018). "Deep learning-based classification with improved time resolution for physical activities of children". In: *PeerJ* 6, e5764. DOI: 10.7717/peerj.5764.

Jarchi, D., Pope, J., Lee, T K M., Tamjidi, L., Mirzaei, A., and Sanei, S. (2018). "A Review on accelerometry-based gait analysis and emerging clinical applications". In: *IEEE Reviews in Biomedical Engineering* 11, pp. 177–194. DOI: 10.1109/RBME.2018.2807182.

Jepsen, A L. and Eskerod, P. (2009). "Stakeholder analysis in projects: challenges in using current guidelines in the real world". In: *International Journal of Project Management* 27.4, pp. 335–343. DOI: 10.1016/j.ijproman.2008.04.002.

John, D., Sasaki, J., Staudenmayer, J., Mavilia, M., and Freedson, P S. (2013). "Comparison of raw acceleration from the GENEA and ActiGraph^{TM} GT3X+ activity monitors". In: *Sensors (Basel, Switzerland)* 13.11, pp. 14754–14763. DOI: 10.3390/s131114754.

Jones, J C. (1976). "Tools for development". In: *The Round Table* 66.264, pp. 331–341. DOI: 10.1080/00358537608453236.

Julious, S A. (2005). "Sample size of 12 per group rule of thumb for a pilot study". In: *Pharmaceutical Statistics* 4.4, pp. 287–291. DOI: 10.1002/pst.185.

Kang, X., Huang, B., and Qi, G. (2018). "A novel walking detection and step counting algorithm using unconstrained smartphones". In: *Sensors (Basel)* 18.1, p. 297. DOI: 10.3390/s18010297.

Kaptein, R G., Wezenberg, D., Ijmker, T., Houdijk, H., Beek, P J., Lamoth, C J., and Daffertshofer, A. (2014). "Shotgun approaches to gait analysis: insights and limitations". In: *J Neuroeng Rehabil* 11, p. 120. DOI: 10.1186/1743-0003-11-120.

Kavanagh, J J. and Menz, H B. (2008). "Accelerometry: a technique for quantifying movement patterns during walking". In: *Gait Posture* 28.1, pp. 1–15. DOI: 10.1016/j.gaitpost.2007.10.010.

Kelleher, John D, Namee, Brian Mac, and D'arcy, Aoife (2015). *Fundamentals of machine learning for predictive data analytics*. ISBN: 9780262029445.

Kirtley, C. (2006). *Clinical Gait Analysis. Theory and Practice*, p. 328. DOI: https://doi.org/10.1016/B978-0-443-10009-3.X5001-2.

Klassen, T D., Simpson, L A., Lim, S B., Louie, D R., Parappilly, B., Sakakibara, B M., Zbogar, D., and Eng, J J. (2016). ""Stepping up" activity poststroke: ankle-positioned accelerometer can accurately record steps during slow walking". In: *Physical Therapy* 96.3, pp. 355–360. DOI: 10.2522/ptj.20140611.

Kohl, HW., Craig, C L., Lambert, E V., Inoue, S., Alkandari, J R., Leetongin, G., and Kahlmeier, S. (2012). "The pandemic of physical inactivity: global action for public health". In: *Lancet* 380.9838, pp. 294–305. DOI: 10.1016/S0140-6736(12)60898-8.

Koo, H. and Lee, S. (2016). "Gait analysis on the condition of arm swing in healthy young adults". In: *Phys Ther Rehabil Sci* 5.3, pp. 149–154. DOI: 10.14474/ptrs.2016.5.3.149.

Korjus, K., Hebart, M N., and Vicente, R. (2016). "An efficient data partitioning to improve classification performance while keeping parameters interpretable". In: *PLoS ONE* 11.8, e0161788. DOI: 10.1371/journal.pone.0161788.

Kosmadopoulos, A., Darwent, D., and Roach, G D. (2016). "Is it on? An algorithm for discerning wrist-accelerometer non-wear times from sleep/wake activity". In: *Chronobiology International* 33.6, pp. 599–603. DOI: 10.3109/07420528.2016.1167720.

Laarhoven, S N V., Lipperts, M., Bolink, S A A N., Senden, R., Heyligers, I C., and Grimm, B. (2016). "Validation of a novel activity monitor in impaired, slow-walking, crutch-supported patients". In: *Annals of Physical and Rehabilitation Medicine* 59.5-6, pp. 308–313. DOI: 10.1016/j.rehab.2016.05.006.

Lamont, R M., Daniel, H L., Payne, C L., and Brauer, S G. (2018). "Accuracy of wearable physical activity trackers in people with Parkinson's disease". In: *Gait Posture* 63, pp. 104–108. DOI: 10.1016/j.gaitpost.2018.04.034.

Lara, O D. and Labrador, M A. (2013). "A survey on human activity recognition using wearable sensors". In: *IEEE Communications Surveys & Tutorials* 15.3, pp. 1192–1209. DOI: 10.1109/SURV.2012.110112.00192.

Lara, Ó D., Prez, A J., Labrador, M A., and Posada, J D. (2012). "Centinela: A human activity recognition system based on acceleration and vital sign data". In: *Pervasive and Mobile Computing* 8.5, pp. 717–729. DOI: 10.1016/j.pmcj.2011.06.004.

Larkin, L., Nordgren, B., Purtill, H., Brand, C., Fraser, A., and Kennedy, N. (2016). "Criterion Validity of the activPAL Activity Monitor for Sedentary and Physical Activity Patterns in People Who Have Rheumatoid Arthritis". In: *Physical Therapy* 96.7, pp. 1093–1101. DOI: 10.2522/ptj.20150281.

Lee, H H., Choi, S., and Lee, M J. (2015). "Step detection robust against the dynamics of smartphones". In: *Sensors (Basel))* 15.10, pp. 27230–27250. DOI: 10.3390/s151027230.

Lee, H S. (2019). "Application of dynamic time warping algorithm for pattern similarity of gait". In: *J Exerc Rehabil* 15.4, pp. 526–530. DOI: 10.12965/jer.1938384.192.

Lee, I., Shiroma, E J., Lobelo, F., Puska, P., Blair, S N., and Katzmarzyk, P T. (2012). "Impact of physical inactivity on the world's major non-communicable diseases". In: *Lancet* 380.9838, pp. 219–229. DOI: 10.1016/S0140-6736(12)61031-9.

Lee, K. and Kwan, M P. (2018). "Physical activity classification in free-living conditions using smartphone accelerometer data and exploration of predicted results". In: *Computers, Environment and Urban Systems* 67, pp. 124–131. DOI: 10.1016/j.compenvurbsys.2017.09.012.

Li, F., Shirahama, K., Nisar, M., Köping, L., and Grzegorzek, M. (2018). "Comparison of feature learning methods for human activity recognition using wearable sensors". In: *Sensors* 18.3, p. 679. ISSN: 1424-8220. DOI: 10.3390/s18020679. URL: http://www.mdpi.com/1424-8220/18/2/679.

Lingesan, B. and Rajesh, R. (2018). "Tilt angle detector using 3-axis accelerometer". In: *Int J Sci Research in Sci and Tech* 4.2, pp. 784–791.

Lipperts, M., Laarhoven, S V., Senden, R., Heyligers, I., and Grimm, B. (2017). "Clinical validation of a body-fixed 3D accelerometer and algorithm for activity monitoring in orthopaedic patients". In: *Journal of Orthopaedic Translation* 11, pp. 19–29. DOI: 10.1016/j.jot.2017.02.003.

Lipstein, E A., Dodds, C M., Lovell, DJ., Denson, L A., and Britto, M T. (2016). "Making decisions about chronic disease treatment: a comparison of parents and their adolescent children". In: *Health Expectations* 19.3, pp. 716–726. DOI: 10.1111/hex.12210.

Lonini, L., Gupta, A., Deems-Dluhy, S., Hoppe-Ludwig, S., Kording, K., and Jayaraman, A. (2017). "Activity recognition in individuals walking with assistive devices: the benefits of device-specific models". In: *JMIR Rehabilitation and Assistive Technologies* 4.2, e8. DOI: 10.2196/rehab.7317.

Lonini, L., Gupta, A., Kording, K., and Jayaraman, A. (2016). "Activity recognition in patients with lower limb impairments: Do we need training data from each patient?" In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3265–3268. DOI: 10.1109/EMBC.2016.7591425.

Maher, D., Ford, N., and Unwin, N. (2012). "Priorities for developing countries in the global response to non-communicable diseases". In: *Globalization and Health* 8, p. 14. DOI: 10.1186/1744-8603-8-14.

Majumder, S., Mondal, T., and Deen, M J. (2017). "Wearable sensors for remote health monitoring". In: *Sensors (Basel, Switzerland)* 17.1, p. 130. DOI: 10.3390/s17010130.

Mancuso, P J., Thompson, M., Tietze, M., Kelk, S., and Roux, G. (2014). "Can patient use of daily activity monitors change nurse practitioner practice?" In: *The Journal for Nurse Practitioners* 10.10, 787–793.e4. DOI: 10.1016/j.nurpra.2014.09.002.

Mannini, A., Intille, S S., Rosenberger, M., Sabatini, A M., and Haskell, W. (2013). "Activity recognition using a single accelerometer placed at the wrist or ankle". In: *Med Sci Sports Exerc* 45.11, pp. 2193–2203. DOI: 10.1249/MSS.0b013e31829736d6.Activity.

Mannini, A., Rosenberger, M., Haskell, W L., Sabatini, A M., and Intille, S S. (2017). "Activity Recognition in Youth Using Single Accelerometer Placed at Wrist or Ankle". In: *Med Sci Sports Exerc* 49.4, pp. 801–812. DOI: 10.1249/MSS.0000000000001144.

Mannini, A. and Sabatini, A M. (2010). "Machine learning methods for classifying human physical activity from on-body accelerometers". In: *Sensors (Basel)* 10.2, pp. 1154–1175. DOI: 10.3390/s100201154.

Mannini, A., Trojaniello, D., Cereatti, A., and Sabatini, A M. (2016). "A machine learning framework for gait classification using inertial sensors: application to elderly, post-stroke and huntington's disease patients". In: *Sensors (Basel)* 16.1, p. 134. DOI: 10.3390/s16010134.

Mantilla, J., Oudre, L., Barrois, R., Vienne, Á., and Ricard, D. (2017). "Template-DTW based on inertial signals: preliminary results for step characterization". In: *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2267–2270. ISSN: 1557170X. DOI: 10.1109/EMBC.2017.8037307.

Marschollek, M., Goevercin, M., Wolf, K., Song, B., Gietzelt, M., Haux, R., and Steinhagen-Thiessen, E. (2008). "A performance comparison of accelerometry-based step detection algorithms on a large, non-laboratory sample of healthy and mobility-impaired persons". In: *Annu Int Conf IEEE Eng Med Biol Soc* 2008, pp. 1319–1322. DOI: 10.1109/IEMBS.2008.4649407.

Mathie, M J., Coster, A C F., Lovell, N H., and Celler, B G. (2004). "Accelerometry: providing an integrated, practical method for long-term, ambulatory monitoring of human movement". In: *Physiological Measurement* 25.2, R1–20. DOI: 10.1088/0967-3334/25/2/R01.

McPhail, S M., Schippers, M., and Marshall, A L. (2014). "Age, physical inactivity, obesity, health conditions, and health-related quality of life among patients receiving conservative management for musculoskeletal disorders". In: *Clinical Interventions in Aging* 9, pp. 1069–1080. DOI: 10.2147/CIA.S61732.

McSharry, P E., Clifford, G D., Tarassenko, L., and Smith, L A. (2003). "A Dynamical Model for Generating Synthetic Electrocardiogram Signals". In: *IEEE Transactions on Biomedical Engineering* 50.3, pp. 289–294. DOI: 10.1109/TBME.2003.808805.

Menz, H B., Lord, S R., and Fitzpatrick, R C. (2003). "Acceleration patterns of the head and pelvis when walking on level and irregular surfaces". In: *Gait Posture* 18.1, pp. 35–46. DOI: 10.1016/S0966-6362(02)00159-5.

Mertins, A. (1999). *Signal analysis wavelets, filter banks, time-frequency transforms and applications*, p. 330. ISBN: 978-0-471-98626-3.

Micó-Amigo, M E., Kingma, I., Ainsworth, E., Walgaard, S., Niessen, M., Van Lummel, R C., and Van Dieën, J H. (2016). "A novel accelerometry-based algorithm for the detection of step durations over short episodes of gait in healthy elderly". In: *Journal of NeuroEngineering and Rehabilitation* 13, p. 38. DOI: 10.1186/s12984-016-0145-6.

Mikov, A., Moschevikin, A., Fedorov, A., and Sikora, A. (2013). "A localization system using inertial measurement units from wireless commercial hand-held devices". In: *International Conference on Indoor Positioning and Indoor Navigation*, pp. 1–7. DOI: 10.1109/IPIN.2013.6817924.

Miller, G D., Jakicic, J M., Rejeski, W J., Whit-Glover, M C., Lang, W., Walkup, Mi P., and Hodges, M L. (2013). "Effect of varying accelerometry criteria on physical activity: the look ahead study". In: *Obesity (Silver Spring)* 21.1, pp. 32–44. DOI: 10.1002/oby.20234.

Moe-Nilssen, R. and Helbostad, J L. (2004). "Estimation of gait cycle characteristics by trunk accelerometry". In: *J Biomech* 37.1, pp. 121–126. DOI: 10.1016/S0021-9290(03)00233-1.

Montes, J., Tandy, R., Young, J., Lee, S., and Navalata, J W. (2020). "Step count reliability and validity of five wearable technology devices while walking and jogging in both a free motion setting and on a treadmill". In: *International Journal of Exercise Science* 13.7, pp. 410–426.

Montoye, A H K., Pivarnik, J M., Mudd, L M., Biswas, S., and Pfeiffer, K A. (2016). "Validation and comparison of accelerometers worn on the hip, thigh, and wrists for measuring physical activity and sedentary behavior". In: *AIMS Public Health* 3.2, pp. 298–312. DOI: 10.3934/publichealth.2016.2.298.

Motl, R W., Snook, E M., and Agiovlasitis, S. (2011). "Does an accelerometer accurately measure steps taken under controlled conditions in adults with mild multiple sclerosis?" In: *Disability and Health Journal* 4.1, pp. 52–57. DOI: 10.1016/j.dhjo.2010.02.003.

Mynarski, W., Psurek, A., Borek, Z., Rozpara, M., Grabara, M., and Strojek, K. (2012). "Declared and real physical activity in patients with type 2 diabetes mellitus as assessed by the International Physical Activity Questionnaire and Caltrac accelerometer monitor: a potential tool for physical activity assessment in patients with type 2 dia". In: *Diabetes Research and Clinical Practice* 98.1, pp. 46–50. DOI: 10.1016/j.diabres.2012.05.024.

Nayak, M. and Panigrahi, B S. (2011). "Advanced signal processing techniques for feature extraction in data mining". In: *International Journal of Computer Applications* 19.9, pp. 30–37. ISSN: 09758887. DOI: 10.5120/2387-3160.

Ndahimana, D. and Kim, E. (2017). "Measurement methods for physical activity and energy expenditure: a review". In: *Clinical Nutrition Research* 6.2, pp. 68–80. DOI: 10.7762/cnr.2017.6.2.68.

Ogbuabor, G. and La, R. (2018). "Human activity recognition for healthcare using smartphones". In: *ICMLC 2018: Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, pp. 41–46. DOI: 10.1145/3195106.3195157.

Orphanidou, C. and Wong, D. (2017). "Machine learning models for multidimensional clinical data". In: *In: Khan S., Zomaya A., Abbas A. (eds) Handbook of Large-Scale Distributed Computing in Smart Healthcare. Scalable Computing and Communications*, pp. 177–216. DOI: 10.1007/978-3-319-58280-1_8.

Oudre, L., Barrois-Müller, R., Moreau, T., Truong, C., Vienne-Jumeau, A., Ricard, D., Vayatis, N., and Vidal, P P. (2018). "Template-based step detection with inertial measurement units". In: *Sensors (Basel)* 18.11, p. 4033. DOI: 10.3390/s18114033.

Palshikar, G K. (2009). "Simple algorithms for peak detection in time-series". In: DOI: 10.1109/IEMBS.2002.1134453.

Park, Y., Ghosh, J., and Shankar, M. (2013). "Perturbed gibbs samplers for generating large-scale privacy-safe synthetic health data". In: *2013 IEEE International Conference on Healthcare Informatics*, pp. 493–498. DOI: 10.1109/ICHI.2013.76.

Pate, R R., O'Neill, J R., and Lobelo, F. (2008). "The evolving definition of "sedentary"". In: *Exercise and Sport Sciences Reviews* 36.4, pp. 173–178. DOI: 10.1097/JES.0b013e3181877d1a.

Pedersen, B K. and Saltin, B. (2015). "Exercise as medicine - evidence for prescribing exercise as therapy in 26 different chronic diseases". In: *Scandinavian Journal of Medicine and Science in Sports* 25.Suppl 3, pp. 1–72. DOI: 10.1111/sms.12581.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). "Scikit-learn: machine learning in python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830. DOI: 10.1145/2786984.2786995.

Petitjean, F., Ketterlin, A., and Gançarski, P. (2011). "A global averaging method for dynamic time warping, with applications to clustering". In: *Pattern Recognition* 44.3, pp. 678–693. DOI: 10.1016/j.patcog.2010.09.013.

Pham, M H., Elshehabi, M., Haertner, L., Del Din, S., Srulijes, K., Heger, T., Synofzik, M., Hobert, M A., Faber, G S., Hansen, C., Salkovic, D., Ferreira, J J., Berg, D., Sanchez-Ferro, Á., Van Dieën, J H., Becker, C., Rochester, L., Schmidt, G., and Maetzler, W. (2017). "Validation of a step detection algorithm during straight walking and turning in Patients with Parkinson's disease and older adults using an inertial measurement unit at the lower back". In: *Front Neurol* 8, p. 457. DOI: 10.3389/fneur.2017.00457.

Pham, V T., Duc, N A., Dang, N D., Pham, H H., Tran, V A., Sandrasegaran, K., and Tran, D. (2018). "Highly accurate step counting at various walking states using low-cost inertial measurement unit support indoor positioning system". In: *Sensors* 18.10, p. 3186. DOI: 10.3390/s18103186.

Phillips, S M., Cadmus-Bertram, L., Rosenberg, D., Buman, M P., and Lynch, B M. (2018). "Wearable technology and physical activity in chronic disease: opportunities and challenges". In: *Am J Prev Med* 54.1, pp. 144–150. DOI: 10.1016/j.amepre.2017.08.015.

Pogorelc, B., Bosnić, Z., and Gams, M. (2012). "Automatic recognition of gait-related health problems in the elderly using machine learning". In: *Multimed Tools Appl* 58, pp. 333–354. DOI: 10.1007/s11042-011-0786-1.

Ponce, H., Miralles-Pechuán, L., and Martínez-Villaseñor, M. (2016). "A flexible approach for human activity recognition using artificial hydrocarbon networks". In: *Sensors* 16.11, p. 1715. DOI: 10.3390/s16111715.

Preece, S J., Goulermas, J Y., Kenney, L P J., Howard, D., Meijer, K., and Crompton, R. (2009). "Activity identification using body-mounted sensors—a review of classification techniques". In: *Physiological Measurement* 30.4, R1–33. DOI: 10.1088/0967-3334/30/4/R01.

Racic, V. and Brownjohn, J M W. (2012). "Mathematical modelling of random narrow band lateral excitation of footbridges due to pedestrians walking". In: *Computers and Structures* 90-91, pp. 116–130. DOI: 10.1016/j.compstruc.2011.10.002.

Racic, V. and Morin, J B. (2014). "Data-driven modelling of vertical dynamic excitation of bridges induced by people running". In: *Mechanical Systems and Signal Processing* 43.1-2, pp. 153–170. DOI: 10.1016/j.ymssp.2013.10.006.

Racic, V. and Pavic, A. (2010a). "Mathematical model to generate near-periodic human jumping force signals". In: *Mechanical Systems and Signal Processing* 24.1, pp. 138–152. DOI: 10.1016/j.ymssp.2009.07.001.

– (2010b). "Stochastic approach to modelling of near-periodic jumping loads". In: *Mechanical Systems and Signal Processing* 24.8, pp. 3037–3059. DOI: 10.1016/j.ymssp.2010.05.019.

Rhudy, M B. and Mahoney, J M. (2018). "A comprehensive comparison of simple step counting techniques using wrist- and ankle-mounted accelerometer and gyroscope signals". In: *J Med Eng Technol* 42.3, pp. 236–243. DOI: 10.1080/03091902.2018.1470692.

Rodgers, M M., Alon, G., Pai, V M., and Conroy, R S. (2019). "Wearable technologies for active living and rehabilitation: current research challenges and future opportunities". In: *Journal of Rehabilitation and Assistive Technologies Engineering* 6, p. 205566831983960. DOI: 10.1177/2055668319839607.

Rodríguez, G., Casado, F E., Iglesias, R., Regueiro, C V., and Nieto, A. (2018). "Robust step counting for inertial navigation with mobile phones". In: *Sensors (Basel)* 18.9, p. 3157. DOI: 10.3390/s18093157.

Saez, Y., Baldominos, A., and Isasi, P. (2016). "A comparison study of classifier algorithms for cross-person physical activity recognition". In: *Sensors (Basel)* 17.1, p. 66. DOI: 10.3390/s17010066.

Santaniello, S., Fiengo, G., Glielmo, L., and Catapano, G. (2006). "Dynamic modeling and statistical characterization of subthalamic nucleus neural activity in Parkinson's disease patients". In: *2006 American Control Conference* 2006, pp. 4812–4817. DOI: 10.1109/acc.2006.1657482.

Santos-Lozano, A., Hernández-Vicente, A., Pérez-Isaac, R., Santín-Medeiros, F., Cristi-Montero, C., Casajús, J A., and Garatachea, N. (2017). "Is the SenseWear armband accurate enough to quantify and estimate energy expenditure in healthy adults?" In: *Annals of Translational Medicine* 5.5, p. 97. DOI: 10.21037/atm.2017.02.31.

Sasaki, J E., Hickey, A., Staudenmayer, J., John, D., Kent, J A., and Freedson, P S. (2016). "Performance of activity classification algorithms in free-living older adults". In: *Med Sci Sports Exerc* 48.5, pp. 941–50. DOI: 10.1249/MSS.0000000000000844.

Sasko, B., Thiem, U., Christ, M., Trappe, H J., Ritter, O., and Pagonas, N. (2018). "Size matters: an observational study investigating estimated height as a reference size for calculating tidal volumes if low tidal volume ventilation is required". In: *PLoS ONE* 13.6, e0199917. DOI: 10.1371/journal.pone.0199917.

Sazonov, E., Hedge, N., Browning, R C., Melanson, E L., and Sazonova, N A. (2014). "Posture and activity recognition and energy expenditure prediction in a wearable platform". In: *2014*

36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 4163–4167. DOI: 10.1109/JBHI.2015.2432454.

Schmeer, K. (1999). "Guidelines for conducting a stakeholder analysis". In: *MD: Partnerships for Health Reform*, pp. 1–42.

Schrack, J A., Cooper, R., Koster, A., Shiroma, E J., Murabito, Joanne M., Rejeski, W J., Ferrucci, L., and Harris, T B. (2016). "Assessing daily physical activity in older adults: unraveling the complexity of monitors, measures, and methods". In: *Journals of Gerontology - Series A Biological Sciences and Medical Sciences* 71.8, pp. 1039–1048. DOI: 10.1093/gerona/glw026.

Sejdic, E., Lowry, K A., Bellanca, J., Perera, S., Redfern, M S., and Brach, J S. (2016). "Extraction of stride events from gait accelerometry during treadmill walking". In: *IEEE J Transl Eng Health Med* 4, p. 2100111.

Seo, J., Chiang, Y., Laine, T H., and Khan, A M. (2015). "Step counting on smartphones using advanced zero-crossing and linear regression". In: *IMCOM '15: Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication* 106, pp. 1–7. DOI: 10.1145/2701126.2701223.

Shany, T., Redmond, S J., Narayanan, M R., and Lovell, N H. (2012). "Sensors-based wearable systems for monitoring of human movement and falls". In: *IEEE Sensors Journal* 12.3, pp. 658–670. DOI: 10.1109/JSEN.2011.2146246.

Sliepen, M., Mauricio, E., Lipperts, M., Grimm, B., and Rosenbaum, D. (2018). "Objective assessment of physical activity and sedentary behaviour in knee osteoarthritis patients – beyond daily steps and total sedentary time". In: *BMC Musculoskeletal Disorders* 19.1, p. 64. DOI: 10.1186/s12891-018-1980-3.

Smith, S W. (1999). *The scientist and engineer's guide to digital signal processing*. Second. ISBN: 0966017676.

Stansfield, B., Hajarnis, M., and Sudarshan, R. (2015). "Characteristics of very slow stepping in healthy adults and validity of the activPAL3^{TM} activity monitor in detecting these steps". In: *Med Eng Phys* 37.1, pp. 42–47. DOI: 10.1016/j.medengphy.2014.10.003.

Stefanovska, A., Lotric, M B., Strle, S., and Haken, H. (2001). "The cardiovascular system as coupled oscillators?" In: *Physiological Measurement* 22.3, pp. 535–550. DOI: 10.1088/0967-3334/22/3/311.

Storm, F A., Heller, B W., and Mazzà, C. (2015). "Step detection and activity recognition accuracy of seven physical activity monitors". In: *PLoS ONE* 10.3, e0118723. DOI: 10.1371/journal.pone.0118723.

Strath, S J., Kaminsky, L A., Ainsworth, B E., Ekelund, U., Freedson, P S., Gary, R A., Richardson, C R., Smith, D T., and Swartz, A M. (2013). "Guide to the assessment of physical activity: clinical and research applications: a scientific statement from the American Heart Association". In: *Circulation* 128.20, pp. 2259–2279. DOI: 10.1161/01.cir.0000435708.67487.da.

Strath, S J., Kate, R J., Keenan, K G., Welch, W A., and Swartz, A M. (2015). "Ngram time series model to predict activity type and energy cost from wrist, hip and ankle accelerometers: Implications of age". In: *Physiological Measurement* 36.11, pp. 2335–2351. DOI: 10.1088/0967-3334/36/11/2335.

Strath, S J. and Rowley, T W. (2018). "Wearables for promoting physical activity". In: *Clinical Chemistry* 64.1, pp. 53–63. DOI: 10.1373/clinchem.2017.272369.

Suto, J., Oniga, S., and Pop Sitar, P. (2016). "Feature analysis to human activity recognition". In: *International Journal of Computers Communications & Control* 12.1, p. 116. ISSN: 1841-9836. DOI: 10.15837/ijccc.2017.1.2787. URL: http://univagora.ro/jour/index.php/ijccc/article/view/2787.

Sylvia, L G., Bernstein, E E., Hubbard, J L., Keating, L., and Anderson, E J. (2014). "A practical guide to measuring physical activity". In: *Journal of the Academy of Nutrition and Dietetics* 114.2, pp. 199–208. DOI: 10.1016/j.jand.2013.09.018.

Tao, W., Liu, T., Zheng, R., and Feng, H. (2012). "Gait analysis using wearable sensors". In: *Sensors (Basel)* 12.2, pp. 2255–2283. DOI: 10.3390/s120202255.

Taraldsen, K., Chastin, SF M., Riphagen, I I., Vereijken, B., and Helbostad, J L. (2012). "Physical activity monitoring by use of accelerometer-based body-worn sensors in older adults:

a systematic literature review of current knowledge and applications". In: *Maturitas* 71.1, pp. 13–19. DOI: 10.1016/j.maturitas.2011.11.003.

Taylor, D. (2014). "Physical activity is medicine for older adults". In: *Postgraduate Medical Journal* 90.1059, pp. 26–32. DOI: 10.1136/postgradmedj-2012-131366.

Thanh, P V., Thi, A N., Thuy, Q T T., Phuong, D C T., Mau, V., and Tran, D. (2017). "A novel step counter supporting for indoor positioning based on inertial measurement unit". In: *In Proceedings of the 7th International Conference on Integrated Circuits, Design, and Verification (ICDV)*, pp. 69–74. DOI: 10.1109/ICDV.2017.8188641.

Tophøj, K H., Petersen, M G., Sæbye, C., Baad-Hansen, T., and Wagner, S. (2018). "Validity and reliability evaluation of four commercial activity trackers' step counting performance". In: *Telemed J E Health* 24.9, pp. 669–677. DOI: 10.1089/tmj.2017.0264.

Toth, L., Park, S., Pittman, W., Sarisaltik, D., Hibbing, P., Morton, A., Springer, C., Crouter, S., and Bassett, D. (2018). "Validity of activity tracker step counts during walking, running, and activities of daily living". In: *Translational Journal of the American College of Sports Medicine* 3.7, pp. 52–59. DOI: 10.1249/TJX.0000000000000057.

Treacy, D., Hassett, L., Schurr, K., Chagpar, S., Paul, S S., and Sherrington, C. (2017). "Validity of different activity monitors to count steps in an inpatient rehabilitation setting". In: *American Physical Therapy Association* 97.5, pp. 581–588. DOI: 10.1093/ptj/pzx010.

Trost, S G. and O'Neil, M. (2014). "Clinical use of objective measures of physical activity". In: *British journal of sports medicine* 48.3, pp. 178–81. DOI: 10.1136/bjsports-2013-093173.

Tucker, P. and Gilliland, J. (2007). "The effect of season and weather on physical activity: A systematic review". In: *Public Health* 121.12, pp. 909–922. DOI: 10.1016/j.puhe.2007.04.009.

Twomey, N., Fafoutis, X., Elsts, A., McConville, R., Flach, P., and Craddock, I. (2018). "A comprehensive study of activity recognition using accelerometers". In: *Informatics* 5.2, p. 27. DOI: 10.3390/informatics5020027.

Ummels, D., Beekman, E., Theunissen, K., Braun, S., and Beurskens, A J. (2018). "Counting steps in activities of daily living in people with a chronic disease using nine commercially available fitness trackers: cross-sectional validity study". In: *JMIR Mhealth Uhealth* 6.4, e70. DOI: 10.2196/mhealth.8524.

Vanhees, L., Lefevre, J., Philippaerts, R., Martens, M., Huygens, W., Troosters, T., and Beunen, G. (2005). "How to assess physical activity? How to assess physical fitness?" In: *European Journal of Cardiovascular Prevention and Rehabilitation* 12.2, pp. 102–114. DOI: 10.1097/01.hjr.0000161551.73095.9c.

Veldhuijzen van Zanten, J J C S., Rouse, P C., Hale, E D., Ntoumanis, N., Metsios, G S., Duda, J L., and Kitas, G D. (2015). "Perceived barriers, facilitators and benefits for regular physical activity and exercise in patients with rheumatoid arthritis: A review of the literature". In: *Sports Medicine* 45.10, pp. 1401–1412. DOI: 10.1007/s40279-015-0363-2.

Verlaan, L., Bolink, S A A N., Laarhoven, S N V., Lipperts, M., Heyligers, I C., Grimm, B., and Senden, R. (2015). "Accelerometer-based physical activity monitoring in patients with knee osteoarthritis: objective and ambulatory assessment of actual physical activity during daily life circumstances". In: pp. 157–163.

Walker, R K., Hickey, A M., and Freedson, P S. (2016). "Advantages and limitations of wearable activity trackers: considerations for patients and clinicians". In: *Clin J Oncol Nurs* 20.6, pp. 606–610. DOI: 10.1188/16.CJON.606-610.

Waltenegus, D. (1999). "Analysis of time and frequency domain features of accelerometer measurements". In: 32.3, pp. 90–102. ISSN: 0024-7766. DOI: 10.1109/ICCCN.2009.5235366. URL: http://www.ncbi.nlm.nih.gov/pubmed/10494521.

Wang, W Z., Huang, B Y., and Wang, L. (2011). "Analysis of filtering methods for 3D acceleration signals in body sensor network". In: *International Symposium on Bioelectronics and Bioinformatics*, pp. 263–266. DOI: 10.1109/ISBB.2011.6107697.

Wang, Z., Myles, P., and Tucker, A. (2019). "Generating and evaluating synthetic UK primary care data: Preserving data utility & patient privacy". In: *2019 IEEE 32nd International*

Symposium on Computer-Based Medical Systems (CBMS) Generating, pp. 126–131. DOI: 10.1109/CBMS.2019.00036.

Wannamethee, S G. and Shaper, A G. (2001). "Physical activity in the prevention of cardiovascular disease: an epidemiological perspective". In: *Sports medicine* 31.2, pp. 101–14. DOI: 10.2165/00007256-200131020-00003.

Warburton, D E R., Nicol, C W., and Bredin, S S D. (2006). "Health benefits of physical activity: the evidence". In: *Canadian Medical Association journal* 174.6, pp. 801–809. DOI: 10.1503/cmaj.051351.

Weiss, R J., Wretenberg, P., Stark, A., Palmblad, K., Larsson, P., Gröndal, L., and Broström, E. (2008). "Gait pattern in rheumatoid arthritis". In: *Gait Posture* 28.2, pp. 229–234. DOI: 10.1016/j.gaitpost.2007.12.001.

Westerterp, K R. (2009). "Assessment of physical activity: A critical appraisal". In: *European Journal of Applied Physiology* 105.6, pp. 823–828. DOI: 10.1007/s00421-009-1000-2.

Whittle, M W. (2007). *Gait analysis: An introduction*. Fourth. Butterwirth Heinemann Elsevier, p. 255.

WHO (2017). *Physical activity*. URL: http://www.who.int/mediacentre/factsheets/fs385/en/.

– (2018). *Noncommunicable diseases country profiles 2018*. Tech. rep., p. 223.

Wise, J M. and Hongu, N. (2014). "Pedometer, accelerometer, and mobile technology for promoting physical activity". In: *College of agriculture & life sciences*, pp. 1–4.

Wong, C K., Mentis, H M., and Kuber, R. (2018). "The bit doesn't fit: Evaluation of a commercial activity-tracker at slower walking speeds". In: *Gait and Posture* 59, pp. 177–181. DOI: 10.1016/j.gaitpost.2017.10.010.

Wu, X., Kumar, V., Quinlan, J R., Ghosh, J., Yang, Q., Motoda, H., Mclachlan, G J., Ng, A., Liu, Bing, Yu, Philip S, Zhou, Zhi-hua, Steinbach, Michael, Hand, David J, and Steinberg, Dan (2008). "Top 10 algorithms in data mining". In: *Knowl Inf Syst* 14, pp. 1–37. DOI: 10.1007/s10115-007-0114-2.

Xu, C., He, J., Zhang, X., Wang, C., and Duan, S. (2017). "Detection of freezing of gait using template-matching-based approaches". In: *Journal of Sensors* 2017. DOI: 10.1155/2017/1260734.

Yamada, M., Aoyama, T., Mori, S., Nishiguchi, S., Okamoto, K., Ito, T., Muto, S., Ishihara, T., Yoshitomi, Hiroyuki, and Ito, Hiromu (2012). "Objective assessment of abnormal gait in patients with rheumatoid arthritis using a smartphone". In: *Rheumatology International* 32.12, pp. 3869–3874. ISSN: 01728172. DOI: 10.1007/s00296-011-2283-2.

Yan, L., Zhen, T., Kong, J L., Wang, L M., and Zhou, X L. (2020). "Walking gait phase detection based on acceleration signals using voting-weighted integrated neural network". In: *Complexity* 2020. DOI: 10.1155/2020/4760297.

Yang, C. and Hsu, Y. (2010). "A review of accelerometry-based wearable motion detectors for physical activity monitoring". In: *Sensors (Basel)* 10.8, pp. 7772–7788. DOI: 10.3390/s100807772.

Zeng, Q., Zhou, B., Jing, C., Kim, N., and Kim, Y. (2015). "A novel step counting algorithm based on acceleration and gravity sensors of a smart-phone". In: *International Journal of Smart Home* 9.4, pp. 211–224. DOI: 10.14257/ijsh.2015.9.4.22.

Zeng, W., Liu, F., Wang, Q., Wang, Y., Ma, L., and Zhang, Y. (2016). "Parkinson's disease classification using gait analysis via deterministic learning". In: *Neurosci Lett* 633, pp. 268–278. DOI: 10.1016/j.neulet.2016.09.043.

Zhang, S., Rowlands, A V., Murray, P., and Hurst, T L. (2012). "Physical activity classification using the GENEA wrist-worn accelerometer". In: *Med Sci Sports Exerc* 44.4, pp. 742–748.