Fall Risk Classification for People with Lower Extremity Amputations using Machine Learning and Smartphone Sensor Features from a 6-Minute Walk Test

Kyle Daines

Thesis submitted to the Faculty of Engineering
in partial fulfillment of requirements for the degree of

MASTER OF APPLIED SCIENCE
in
Biomedical Engineering

Ottawa Carleton Institute for Biomedical Engineering

University of Ottawa

Ottawa, Ontario

uOttawa

Abstract

Falls are a leading cause of injury and accidental injury death worldwide. Fall-risk prevention techniques exist but fall-risk identification can be difficult. While clinical assessment tools are the standard for identifying fall risk, wearable-sensors and machine learning could improve outcomes with automated and efficient techniques. Machine learning research has focused on older adults. Since people with lower limb amputations have greater falling and injury risk than the elderly, research is needed to evaluate these approaches with the amputee population.

In this thesis, random forest and fully connected feedforward artificial neural network (ANN) machine learning models were developed and optimized for fall-risk identification in amputee populations, using smartphone sensor data (phone at posterior pelvis) from 89 people with various levels of lower-limb amputation who completed a 6-minute walk test (6MWT). The best model was a random forest with 500 trees, using turn data and a feature set selected using correlation-based feature selection (81.3% accuracy, 57.2% sensitivity, 94.9% specificity, 0.59 Matthews correlation coefficient, 0.83 F1 score). After extensive ANN optimization with the best ranked 50 features from an Extra Trees Classifier, the best ANN model achieved 69.7% accuracy, 53.1% sensitivity, 78.9% specificity, 0.33 Matthews correlation coefficient, and 0.62 F1 score.

Features from a single smartphone during a 6MWT can be used with random forest machine learning for fall-risk classification in lower limb amputees. Model performance was similarly effective or better than the Timed Up and Go and Four Square Step Test. This model could be used clinically to identify fall-risk individuals during a 6MWT, thereby finding people who were not intended for fall screening. Since model specificity was very high, the risk of accidently misclassifying people who are a no fall-risk individual is quite low, and few people would incorrectly be entered into fall mitigation programs based on the test outcomes.

Table of Contents

List of Figures

List of Tables

Abbreviations and Definitions

6MWT	6-Minute Walk Test
ANN	Artificial Neural Network
AP	Anterior-Posterior
AS	All Steps (feature subset)
CFS	Correlation-based Feature Selection
ETC	Extra Trees Classifier feature selection
F1	F1 score
FFT	Fast-Fourier Transform
FN	False Negatives
FP	False Positives
FQFFT	First Quartile of Fourier Transform
FR	Fall-risk (regarding an individual who is or is not a fall risk)
FSST	Four Square Step Test
KNN	k-Nearest Neighbour
MCC	Matthew's Correlation Coefficient
ML	Medial-Lateral
NFR	No Fall-Risk (regarding an individual who is or is not a fall risk)
RelF	Relief F feature selection
REOH	Ratio of Even to Odd Harmonics
RMSProp	Root Mean Squared Propogation
S	Straightaway (feature subset)
S&T	Straightaway and Turn (feature subset)
SR	Summed Ranking
SVM	Support Vector Machine
T	Turn (feature subset)
TN	True Negatives
TOHRC	The Ottawa Hospital Rehabilitation Centre
TP	True Positives
TUG	Timed-Up and Go test

Acknowledgments

I would like to thank my supervisors, Dr. Edward D. Lemaire and Dr. Natalie Baddour for their guidance and support through the completion of this thesis, as well as their efforts to help me develop professionally through conferences, publications, and an international internship. I appreciate all that you have done for me during both of my degrees.

I would also like to thank Dr. Helena Burger and Dr. Andrej Bavec for their assistance with data collection and their clinical expertise, and Natural Sciences and Engineering Research Council of Canada Discovery and Create grants for funding the research in this thesis.

I would like to express my deepest appreciation for all my family who has supported me throughout my studies. Thank you to my wife Jenny, who worked hard to provide us a home and plan our wedding during my studies. Thank you to my mother, Kim Daines, father, Ian Daines, and stepmother, Shannon Daines for helping me through decades of schooling but keeping me well-rounded. Thank you to my sister Cassandra Daines for always believing in me.

Finally, thank you to all my friends and colleagues who made my experience at The Ottawa Hospital Rehabilitation Centre so enjoyable, and to Gabrielle Thibault and Pascale Juneau for helping with data collection and analysis.

1 Introduction

Falling can be a life altering incident that leads to injury, loss of independence, and reduced mobility confidence. This is especially true among at-risk populations such as people with lower limb amputations. According to the World Health Organization, falls are the second leading cause of accidental or unintentional injury deaths [1]. Fall-risk detection is needed so that falling can be mitigated through intervention and assistance [2]. Fall prevention programs such as functional training and strengthening exercises can reduce fall injury severity and the number of falls [2].

Unfortunately, detecting fall-risk individuals can be challenging, and tools used in the past have discriminated poorly between fallers and non-fallers [3]. Clinicians often use medical, fall risk, and mobility assessments to identify people at risk of falling [4]. Tests such as the Four Square Step Test (FSST) [5] and Timed-up and Go (TUG) have some success, but the best results often come from performing multiple tests, which can be time consuming [4]. Recently, techniques using wearable sensors have gained popularity because they can be easily applied at the point-of-care [6]–[10]. Wearable sensors systems can also be used to augment clinical mobility tests. Features can be extracted from inertial sensor data during clinical tests such as 6-minute walk tests (6MWT) [7], 10 meter walk tests [8], or other functional mobility tests. Inertial sensor applications for older adult fall-risk classification have achieved variable success, with accuracies between 62 and 100%, sensitivities between 55 and 99%, and specificities between 35 and 100% [11].

Multiple sensor locations on the body can provide additional data for fall-risk classification, but can increase set-up times, shift, fall off, and affect patient gait. The number of sensors should be minimized and have easy, reliable placement to be viable in clinic. This is an achievable goal, especially given rapidly developing mobile technology such as smartphones. Smartphones are accessible and powerful tools that could provide accurate and automated fall-risk classification.

Fall-risk classification literature has focused primarily on older adult populations because they are the largest at-risk population. However, lower extremity amputees are at higher risk of falling compared to able-bodied individuals [12], [13], and their risk of injury requiring medical care after a fall can be higher [14]. Research is needed to determine if results from existing models trained on older adult populations can also be achieved for people with lower limb amputations.

This thesis evaluated smartphone sensor-based (accelerometer and gyroscope) features on both straight and turn walking for their ability to classify fall-risk in amputees. Techniques and models were developed by building on previous successful studies on older adults [7], [8]. Features were determined for each step during both straight and turn walking, which required that foot strikes and turns be identified through both automated and manual pre-processing. Multiple feature selection techniques were performed to develop smaller, more effective feature sets. Feature sets were tested with random forests and fully connected feedforward artificial neural networks (ANN) on their ability to classify individuals as fall-risk (FR) or no fall-risk (NFR).

1.1 Rationale

Although clinical functional assessment tools can be effective for classifying fall risk, individual tests have been shown to vary in their reliability, indicating that multiple tests should be used for more trustworthy results [4]. However, these tests can be time-consuming and tiring for patients. Wearable sensors have the potential to make fall-risk classification more efficient through automation, therefore reducing the number of required tests. Patient comfort can also be improved through choice of sensor configurations. Sensor configurations used to develop existing research models can be complicated and cumbersome for patients in clinic. The number of sensors should be minimized to reduce their influence on patient movement and reduce the clinic time spent applying sensors to the body.

From a review of previous research, a single accelerometer placed on the lower back was the most common configuration [11]. Although using more than one sensor location can provide better results, fall-risk assessment can be performed using a single sensor located at the lower back [7]. In this research, a smartphone was used, thereby providing an accelerometer and gyroscope. Most previous fall risk studies used only accelerometers. Using both accelerometer and gyroscope can increase the number of relevant features.

Automated fall-risk classification has been achieved for older adult populations using wearable sensor technology to help augment existing clinical assessment tools [7]–[10], [15], [16]. Machine learning techniques have shown promising results for classifying individuals as either FR or NFR individuals using a variety of different tests. The most effective classifiers have been shown to be decision trees (or random forests) and ANN [7], [17]–[19]. However, none of this work has

been performed on amputee populations. Models developed for able-bodied and older adult populations do not necessarily translate to lower extremity amputee populations [11], which indicates a need to research the best features and models for fall-risk classification with amputees.

Previous research has shown that quality of turning in FR individuals is significantly compromised in recurrent fallers compared to non-fallers [20], and that FR individuals can have decreased stability and increased energy expenditure [21], [22]. Since previous research has shown that turn features can be better for classifying fall risk in older adults [7], further investigation using both straight and turn walking data on amputees is needed. A good test that includes both straight and turn walking is a 6MWT [23]. A 6MWT is suitable for fall-risk classification for reasons such as including multiple turns, having a long walking period which can cause people to tire, and therefore emphasize fall-risk behaviour, and already being included as a functional assessment tool in many clinics.

The main hypothesis for this thesis is that fall-risk classification using smartphone sensor data from a 6MWT, with random forests and ANN classifiers, can be achieved on amputee populations with the same or better results than previous studies on older adults.

1.2 Objective

1. Create and evaluate a viable machine learning model for predicting fall risk in people with lower extremity amputation using smartphone sensor data collected during a 6MWT.
   Hypotheses:

   1.1. Sensitivity and specificity for fall risk classification will be equivalent or better than metrics from the TUG [4], [23] and FSST [23], [24] clinical assessment tools for identifying fall risk in amputees and older adults.

   1.2. Sensitivity and specificity for fall risk classification will be equivalent or better than metrics from studies with an older adult cohort that used wearable inertial sensors [7], [8].

2. Compare random forests and ANN to determine which technique provides better results for fall-risk classification in lower limb amputees, and which feature selection techniques are most effective.

   Hypotheses:

   2.1. Random forests will have greater sensitivity and specificity than ANN when classifying fall-risk in amputees [6], [7].

   2.2. Both random forests and ANN will have better evaluation metrics when using feature selection strategies, compared to all features.

1.3 Thesis Contributions

This thesis contributed to fall-risk identification, artificial intelligence applications, and smartphone applications for clinical movement evaluation. I contributed to all aspects of the research except data collection; including, software modifications, data cleaning, data processing, feature selection, model development, and model evaluation. Specific thesis contributions are:

- Demonstrated that the 6MWT, a clinical test for functional capacity but not intended for fall-risk detection, can be used to identify people with lower extremity amputations who are at risk of falling
- Established that appropriate fall risk identification for people with lower extremity amputations can be achieved with IMU data from one smartphone, located at the posterior pelvis. This provides an efficient and accessible technique that can be applied clinically anywhere
- Specificity for the best random forest model was greater than specificities from other models and clinical tests for fall risk (TUG, FSST, etc.), therefore the approach in this thesis is better at avoiding unneeded interventions, saving time and resources
- Sensitivity was similar to other multiple-IMU-based approaches, but lower than clinical tests for fall risk identification (i.e., TUG, FSST). Therefore, while able to correctly identify approximately 60% of people at risk, other tests should be considered in practice if a clinician suspects fall risk
- A random forest classifier with correlation-based feature selection (CFS) was a superior machine learning approach compared with an optimized ANN

1.4 Thesis Outline

This thesis manuscript is divided into 5 chapters. Chapter 2 is a literature review describing existing techniques for fall-risk classification, and previous research that investigated machine learning approaches. Chapter 2 also explains machine learning strategies, presenting feature extraction, feature selection, model optimization, and model evaluation methods.

Chapter 3 contains a full manuscript submitted to PloS One, addressing objective 1 by testing random forests. This chapter evaluates random forests for their ability to classify fall-risk in amputees and compares them with existing clinical tests and previous studies using wearable sensors for fall-risk classification.

Chapter 4 contains a full manuscript that addresses objective 1 by testing neural networks. This chapter investigates a wide range of hyper parameters to optimize a neural network for classifying amputees as fall risk, and then evaluates the neural network by comparing outcomes to existing clinical tests and previous research.

Chapter 5 presents a thesis summary and compares the two approaches used in Chapters 3 and 4, therefore addressing objective 2. Chapter 5 compares random forests with neural networks for the purpose of fall-risk classification and presents suggestions for future work based on thesis conclusions.

2 Literature Review

2.1 Gait Analysis

Humans adapt gait over time to become efficient and stable, but each person has a slightly different gait pattern. By understanding how a person moves, healthcare professionals can instruct or assist people on how to ambulate as effectively and safely as possible. A variety of measurements can be used to quantify gait; such as, temporal spatial (step length, step time, etc.), kinematic (joint angles, limb accelerations, etc.), and kinetic (forces, moments, powers, etc.) measurements.

Walking can be divided into repeated cycles (Figure 1) [24]–[26], and sub-divided into stance and swing phases. Stance phase is when the foot is on the ground and is approximately 60% of the gait cycle. Stance begins with foot-strike (instant the foot touches the ground) and ends with foot-off (instant the foot leaves the ground). Body weight is supported by the limbs in stance. Swing phase is when the foot is in the air, where the foot is advanced to allow forward body progression. Swing begins with foot-off and ends with foot-strike.

Stance phase can be divided into five sub-phases:

1. Initial contact: Instant when the foot contacts the ground.
2. Loading response: Period following initial contact when the body absorbs landing impact by flexing the knee and the foot rotates to be flat on the ground.
3. Mid stance: Period where the body progresses over the foot, between loading response and terminal stance.
4. Terminal stance: Period where push off is performed to propel the body forward.
5. Pre-swing: Period where body weight is removed from the support limb and the foot prepares to be lifted off the ground.

Swing can be divided into three sub-phases:

1. Initial swing: The foot is lifted from the ground.
2. Mid swing: Period where the foot progresses forward.
3. Terminal swing: Period where the swing limb is decelerated, and the foot is positioned for the next ground contact.

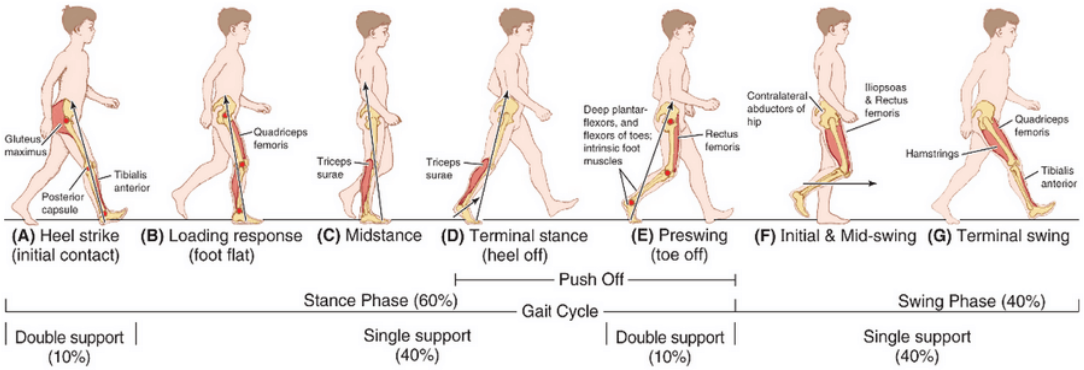

Figure 1 Gait cycle phases and sub phases [26]

2.2 Pathological Gait Leading to Fall Risk

Falls are defined as an event that results in someone coming to rest on the ground or a lower level inadvertently [1]. Existing fall-risk research primarily focuses on older adults because they are the largest fall-risk population. For older adults, a variety of factors can lead to increased fall risk; such as, loss of muscle mass, somatosensory impairment (specifically at the hip), chronic conditions such as arthritis and diabetes, visual impairment, and vitamin D deficiency [27]. These factors can also lead to increased risk of falls requiring medical care.

Interestingly, lower limb amputees have a higher risk of falling than older adults [12], [13]. A UK study found that approximately one third of amputee inpatients experienced falls [28], while a Canadian study found that around 20.5% of inpatients had experienced falls [29]. This demonstrates the need for fall-mitigation strategies for lower limb amputees.

A recent study reported that 44.7% of amputees who fall, fall while walking [30]. Amputee gait differs from able-bodied gait in a variety of ways. Amputee gait is typically asymmetrical; including spending more time in stance on their intact limb than their prosthetic limb, more load on their intact limb, and compensatory gait strategies to improve prosthetic limb clearance [31]–[33]. Additionally, loss of lower limb muscles combined with muscle weakness, mechanical limitations of prosthetic limbs, and postural instability predispose amputees to have an increased risk of falling when compared to able-bodied individuals [12], [34].

FR and NFR amputee populations differ; for example, NFR transtibial amputees have greater knee extensor moments throughout loading, indicating that they can compensate to maintain stability. NFR amputees also perform mechanically demanding tasks, such as stair ascent, more cautiously [35]. While we can identify differences in strategies during gait and activities of daily living, identifying these differences using clinical assessment tools can be difficult.

2.3 Fall-Risk Assessment and Prevention

Exercise programs can prevent falls for older adults living in the community [2]. Interventions recommended for preventing falls include individualized exercises aimed at improving strength, interventions for maximizing vision such as glasses and cataract surgery, and peripheral sensation counselling [36]. Individual exercise classes are a common technique to reduce fall risk, including exercises for conditioning, strength, flexibility, coordination, and balance. Clinical trials have shown that balance training in combination with strength training is an effective method of fall prevention [27].

The ability to predict fall risk is the first step to mitigate the risk of injury due to falls, therefore knowing who is at risk of falling before implementing appropriate interventions. Many tools for assessing fall risk are available to clinicians. While some are very effective, combining tools can improve the ability to evaluate elderly people [4]. Unfortunately, few studies have evaluated these tools on amputees, and the few studies focused on transtibial amputees. Tests such as a FSST [5] and TUG have been used to identify fall-risk in transtibial amputees. To assess for multiple fall risk (2 or more falls in past 6 months), FSST achieved 92% sensitivity and 93% specificity [37] on a convenience sample of transtibial amputees (n=47), while TUG had 85% sensitivity and 74% specificity [37]. When investigating older adults, FSST had a multiple fall risk sensitivity of 89% and specificity of 85% [38]. TUG had 76% and specificity of 49% with older adults when identifying single fall risk (one fall in the past 6 months) [39].

Clinical tests can be adjusted for use with amputees by modifying the interpretation thresholds. For example, TUG, a test requiring participants to stand-up, walk to and around a cone ten meters away, and then return and sit down, has a completion time threshold to identify fall-risk of 13.5 seconds with older adults but a threshold of 19 seconds with transtibial amputees [37]. There is less validated research for these clinical tests for amputees than for older adults. Similarly,

many of the preventative programs and fall-risk detection methods are aimed at older adults or able-bodied populations. Specific fall prevention techniques should be developed for transtibial amputees, such as focusing on improving prosthetic limb knee muscle strength to improve stability and progression during weight transfer onto the prosthetic limb [12].

2.3.1 Turn Relevance

Turn step analysis can indicate fall risk more easily than straight walking, by analyzing features such as peak turn speeds, turn durations, and steps per turn [20], [40]. Increased turn time and more steps per turn can indicate turning difficulties, which is a trait of recurrent fallers [41]. Both TUG and FSST incorporate a turn.

One useful tool to measure functional capacity that includes multiple turns is the 6MWT. 6MWT is widely used to measure response to therapeutic interventions [23], but is not intended for fall-risk detection. Because the 6MWT is widely used by clinicians, the 6MWT would be a convenient way to classify fall-risk at the same time as measuring functional capacity and possibly eliminating the need for other functional assessment tools. During a 6MWT, participants walk along a walkway (typically 20 meters) between two cones used to mark turnaround points. At the end of the walkway, the person walks around the cone (turning), and continues walking back and forth for 6 minutes [42]. While a 6MWT is not traditionally used for fall-risk detection, the inclusion of multiple straight and turn walking sections make the 6MWT a candidate for machine learning based fall-risk classification.

2.4 Inertial Sensor Based Data Collection Techniques

Wearable technology can collect movement data for clinical assessment, such as fall-risk identification or activity recognition. Sensor based technologies include accelerometers, gyroscopes, or physiological signal collectors [43]. Increased accessibility and sensor miniaturization have allowed for ubiquitous computing (computing anytime) when sensors are combined with embedded microprocessors, portable computers, or smartphones. These technologies have encouraged research on wearable sensor use in home monitoring, assisted living, rehabilitation, and clinical assessment [7], [44], [45].

More training and testing data are always preferred for machine learning applications. While individual sensors can provide rich data streams, multiple sensors can provide knowledge about multiple body segments and different types of data, thereby improving machine learning predictions. However, more sensors also lead to more complicated, expensive, and time demanding data collections. In the context of clinical functional assessments, drawbacks from additional sensors can lead to difficulties regarding patient comfort and clinic efficiency. Therefore, data collection during functional assessments such as a 6MWT should be as efficient and simple as possible by having only one sensor attached to the patient.

Fall-risk research for older adults has investigated many wearable sensor locations, including: pelvis, head, upper back, sternum, shoulders, elbow, wrists, hips, thighs, knees, shanks, ankle, and foot [11]. The best and most common single sensor location is the lower back (pelvis) for gait related fall risk identification [7], [8], [11], [46]. A lower back inertial sensor can provide measurements such as Harmonic Ratio (quantifies smoothness of walking), balance shifting in turns, gait speed, step duration, gait variability, gait speed, and activity level to identify fall risk in older adults [47]. These measurements have resulted in fall-risk detection during TUG, treadmill walking, 10 meter walk tests, 6MWT, and dancing [7], [48]–[50].

In a busy clinic, a wearable device that not only collects movement data but also processes and provides immediate feedback is ideal. A smartphone meets this need since data collection, processing, and display is available to clinicians in a portable and familiar device [51], [52]. Applications can use the smartphone's accelerometer, gyroscope, altimeter, and any other sensors for clinical assessment [53]. Prototype smartphone applications for physical activity and fall-risk monitoring are available [50]. Ideally, data can be processed and reported immediately using the smartphone microprocessor and display.

2.5 Extracted Features

A review of 40 studies investigating fall-risk assessment in geriatric populations found that the most common variables to investigate fall-risk were temporal (23.1%), linear acceleration measures such as peak amplitudes (20%), frequency domain measures (15.4%) and angular velocity measures (11.5%) [11]. Temporal features are common in fall-risk assessment because gait variability and variation between steps have been associated with fall risk [20], [54], and with

amputees, step time has been associated with fall risk [55]. Temporal data is also the primary measure for TUG.

Frequency domain features determined from the absolute value of the Fast Fourier Transform (FFT) are relevant for fall risk classification; specifically, peak parameters from lower-back accelerometers, First Quartile of the Fast Fourier Transform (FQFFT), and the Ratio of Even to Odd Harmonics (REOH) magnitudes [11], [40], [56]. FQFFT is the percentage of frequencies within the first quartile of the Nyquist frequency. Lower FQFFT values indicate more high frequency components, which has been linked to instability [56]. REOH is the ratio of frequencies in the even harmonics compared to the odd harmonics, using stride time as the fundamental frequency. Lower REOH values have been associated with fall risk in previous studies [7], [57].

2.6 Feature Selection

While having multiple features can improve model accuracies, some features can be redundant and may adversely affect fall-risk models. Feature selection can improve the model's power while reducing complexity. Feature selection is the process of choosing a subset of all the features to reduce dimensionality and redundancy [58]–[60], which can result in simpler, more comprehensible models. By reducing the number of features, feature selection can increase computational efficiency, decrease memory storage, and build models that better generalize to the population [58]. A similar technique to feature selection called feature extraction projects the original features onto a new, potentially lower dimensionality feature space. However, feature extraction modifies the features so that they lose their physical meaning, which is why feature selection is often preferred, especially clinically.

When training with many features, a critical issue known as the curse of dimensionality can arise. With each additional feature added to the feature set (i.e., an increase in dimensionality), the problem becomes exponentially more complicated [61]. This is especially true for classifiers that investigate the relationships between features, such as ANN. As the number of features increases, so do the number of interactions. In addition to increasing complexity, increasing the number of features can lead to overfitting. Overfitting means the classifier success is skewed toward the training and/or testing set, but won't be as effective on unseen data [58].

Feature selection techniques can be divided into supervised and unsupervised. With supervised feature selection, the aim is to discriminate samples from different classes. In this way, feature selection can determine feature relevance via its correlation to class labels. Unsupervised feature selection is generally designed for clustering problems. These methods are more popular when labeling data as classes is expensive, and instead seek different criteria to define feature relevance. For this thesis, labels are available, hence supervised learning is preferred. Within supervised learning, three types of feature selection are available: filter, wrapper, and embedded.

2.6.1 Types of Feature Selection

Filter methods develop feature subsets independent of any learning algorithms [58]–[60]. Feature importance is assessed by first ranking the features based on evaluation criteria, and then filtering out the lower ranked features. Filters can rank features using univariate or multivariate means. Univariate methods rank features independent of other features, while multivariate methods rank features as batches [58]. Filters are the most computationally efficient of the three methods because they do not require models to be built. Another benefit of filters being model independent is that feature selection techniques can be directly compared. Filter methods use various statistical tests such as Pearson's correlation, linear discriminant analysis, and ANOVA tests to rank features.

The second feature selection method is wrappers. Wrapper methods rely on a specific learning algorithm's predictive performance to evaluate feature effectiveness [58], [59]. Wrappers first search for a random subset of features, and then evaluate the selected features on the algorithm. The process is repeated on various feature subsets while retaining the best learning performance, until the desired number of features is retained. Wrappers then return the feature subset with the highest learning performance. These methods can be computationally expensive since the search space for d features is 2^d, which may be impractical with larger number of features [62]. As a result, wrapper methods are less common in practice than filter methods.

Embedded methods are a tradeoff between filters and wrappers, where feature selection is embedded into the algorithm's execution, making them more computationally efficient then wrappers [58]. First, a machine learning model is trained, then feature importance in making predictions from the model is derived [63]. Using derived feature importance, non-important features are removed. Embedded methods can be used prior to or during model development.

2.6.2 **Relief F**

Relief-F (RelF) is a common supervised filtering method that detects feature dependencies by using the concept of nearest neighbors to derive feature statistics that indirectly account for interactions [60]. RelF selects features by weighting the parameter's relative strength to separate instances between classes. Less relevant features are eliminated based on a feature score [58]. However, redundant features are retained, since RelF is a univariate feature selection process.

2.6.3 **Correlation-Based Feature Selection**

CFS is a supervised, multivariate, filter-based method that identifies a subset of features by calculating the feature's "merit" based on pair-wise correlations [64]–[66]. By calculating the feature's merit, subsets are developed with features that are correlated to class labels (i.e., FR or NFR), but uncorrelated to other parameters. This allows CFS to develop a subset that has no irrelevant or redundant features, only adding features that improve the subset's merit.

2.6.4 **Extra Trees Classifier Ensemble Method**

Decision trees are a commonly used embedded method because they are good at providing feature importance, and can be applied with wearable sensor data [63], [67]. Decision trees select a feature in each recursive step of tree growth and calculate feature importance as the decrease in node impurity, weighted by the probability of reaching that node [68]. A node's importance can be calculated using Gini Importance, which assume only two child nodes per parent node [68].

By training multiple decision trees, each with different random samples of training data, an ensemble of trees can be developed to determine the feature importance by taking a vote across multiple trees [69], [70]. Using multiple decision trees, or a forest, to decide, allows each individual tree to only search subsets of the randomly selected features. This adds variation among trees, allowing each tree to make different mistakes, and combine to one stronger classifier.

Two types of forests can be considered, random forests and Extra Trees Classifiers (ETC). In both forests, each feature is ordered based on their Gini importance and the best features can be selected. With random forests, each node in a tree is formed by searching the subset for the best feature split. ETC splits nodes at random [70], [71], making ETC less computationally expensive,

provide diversified trees [72], and reduce model variance. Reduced variance reduces overfitting, making ETC preferable for feature selection [71], [72].

2.7 Machine Learning

Machine learning is an application of artificial intelligence with the goal of allowing computers to learn automatically. Automatic learning reduces the need for human intervention and subjectivity. This computational learning scheme can predict and make decisions on a wide range of applications, including research and data mining.

A wide variety of machine learning techniques have been developed, each with their own strengths and weaknesses. Examples include support vector machines (SVM), k-nearest neighbour (KNN), Naïve Bayes, decision trees and forests, and multilayer perceptrons (also known as artificial neural networks). These machine learning techniques can be used to organize samples of data into classes, a process known as classification. As with feature selection, there can be either supervised or unsupervised machine learning. With supervised learning, samples have labeled classes, allowing classifiers to learn from existing knowledge. Unsupervised learning is when a classifier makes its own inferences to find hidden patterns and groupings within unlabeled data.

As technology advances, more effective and efficient machine learning can be performed. Modern computing makes machine learning a more practical option than in the past due to reduced computing times, and better data collection for more precise classification. Although a variety of classifiers were investigated during this research, Random Forests and ANN were pursued in depth.

2.7.1 Support Vector Machines

SVM are linear classifiers that create a decision boundary or hyperplane in an N-dimensional space, where N is the number of features. The hyperplane aims to distinctly separate data points into classes with the greatest distance between data points of both classes. Maximizing the distance between the hyperplane and datapoints helps reduce the chance of error on future data [73]. The function used to maximize the margin is called hinge loss [74]. A two-dimensional example of an SVM can be seen in Figure 2.

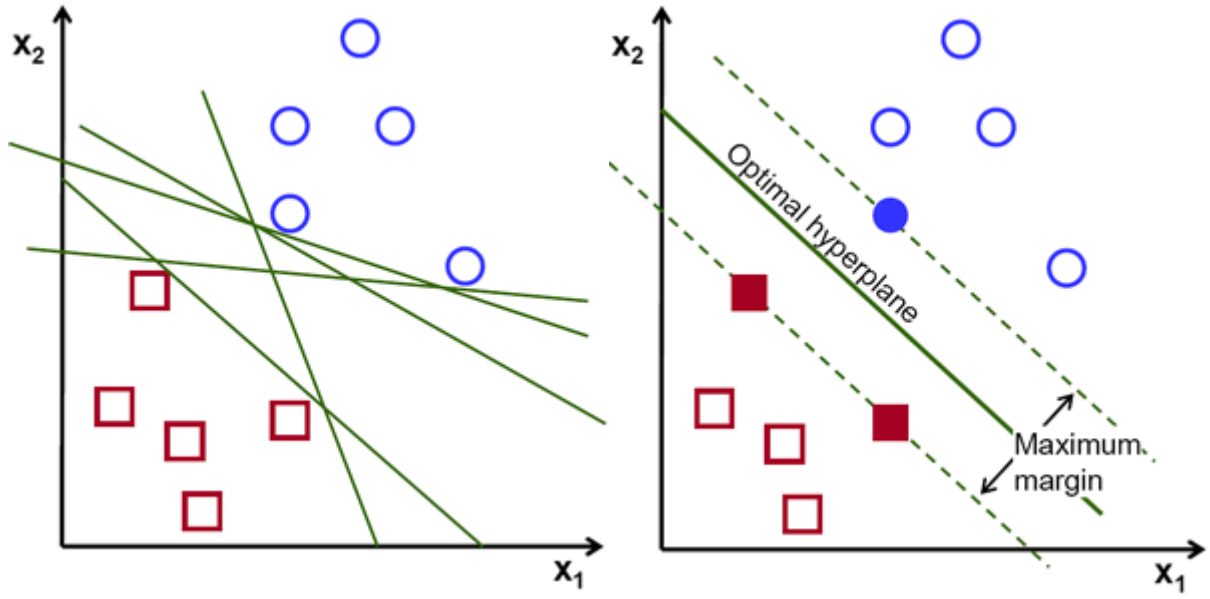

Figure 2 2D representation of SVM. Although many hyperplanes could divide classes into two groups (left), the optimal hyperparameter has a maximum margin from the nearest samples in each class (right) [73].

2.7.2 K-Nearest Neighbour

KNN is a supervised machine learning method that classifies samples based on the nearest *k* neighbours for any given feature space. KNN algorithms assume that similar classes will exist in close proximity, and classifies points based on their distance to their nearest neighbours in N-dimensions (Figure 3) [75]. The distance between points is determined through a distance metric such as a simple Euclidean distance. The Euclidean distance in two dimensions is described as

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \qquad \text{Eq. 1}$$

where *k* is the number of neighbours, and *x* and *y* are the distances between two samples for the two features. By using larger values for *k*, and increasing the number of neighbours to check, the predictions become more stable due to a majority voting. However, after a certain number of neighbours, the algorithm starts to look far away from the sample's location, and makes errors by using further neighbours of other classes [75].

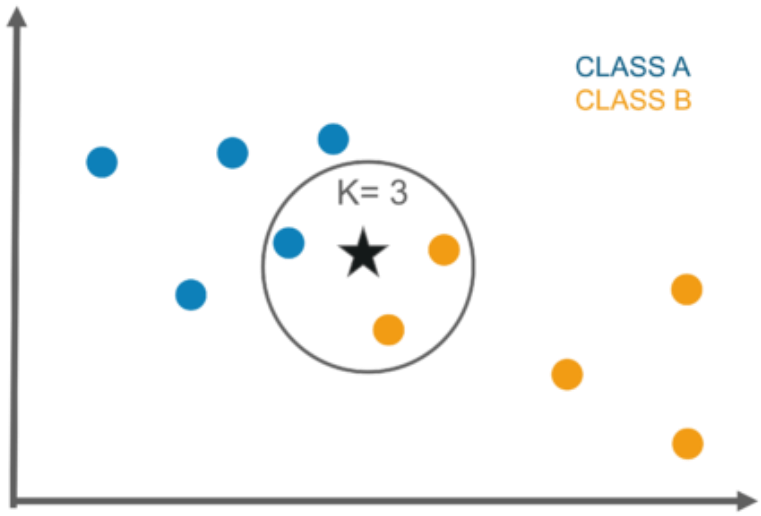

Figure 3 2D KNN example demonstrating a three nearest neighbour classification. Majority voting would classify the star as Class B [76].

2.7.3 Naïve Bayes

Naïve Bayes uses Bayes Theorem to find the probability of A happening, given that B has occurred. In Eq. 2, B is the evidence and A is the hypothesis [77]. The algorithm is called Naïve because it assumes that the presence of one feature does not affect the other, meaning the features are independent. Naïve Bayes makes fast predictions, performs well with multi-class predictions, and works with fewer training samples. However, fall-risk classification with accelerometers found that Naïve Bayes was not as good as other methods, such as random forests and ANN [8].

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{Eq. 2}$$

2.7.4 Decision Trees and Random Forests

Decision trees are effective classifiers that are composed of many threshold-based decisions. As discussed in section 2.6.4, decision trees divide into two choices at each node based on a threshold. The benefit of decision trees as classifiers is that they need minimal data, are robust to noisy data, and are easy to visualize (Figure 4). Decision trees continue splitting until nodes have only one class remaining, or until a desired complexity is reached. Complexity can be controlled by setting a minimum number of training inputs per node, or a maximum depth.

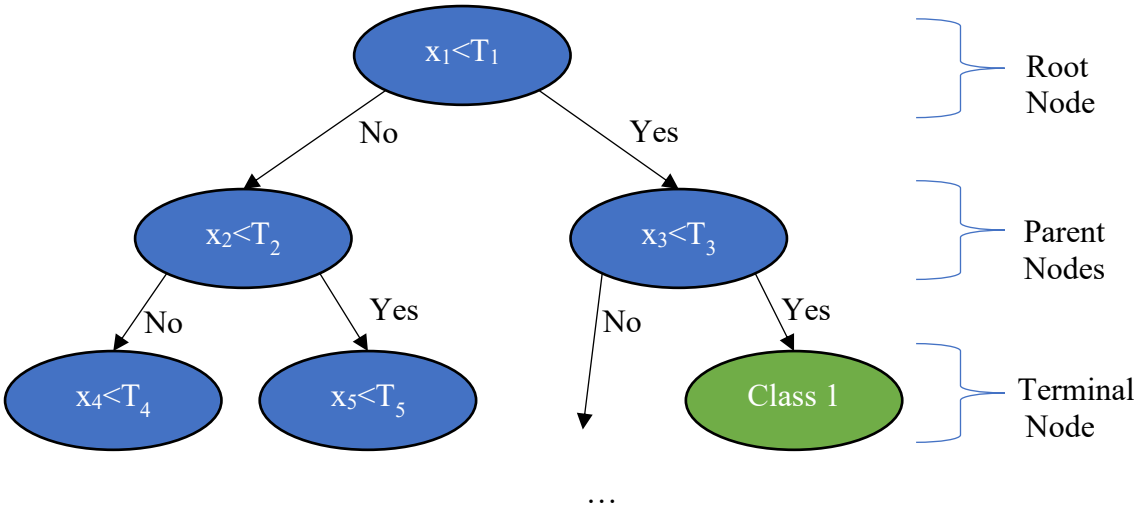

Figure 4 Visual representation of a decision tree. Beginning with a root node, the tree splits using binary decision thresholds, T_n for features x_n. Parent nodes continue splitting until the desired complexity is reached, and each branch ends in terminal nodes.

Unfortunately, due to the ability of decision trees to develop highly complex thresholds across many nodes, a trained model may not generalize to new data. Decision trees have high variance, which may lead to overfitting; therefore, methods to reduce variance are often required to reduce overfitting. One way to reduce variance is to use random forests rather than a single decision tree. As discussed in section 2.6.4, random forests are an ensemble of multiple decision trees, making them less prone to overfitting. Each tree in a random forest is constructed using randomly bootstrapped subsets of training data [78]. Randomly sampling the training data to build decision trees allows random forests to perform better than a single decision tree, with more stable predictions and less overfitting [70]. The models are more robust because they classify new data using a voting technique (Figure 5). These benefits often lead to much better model performance than single decision trees.

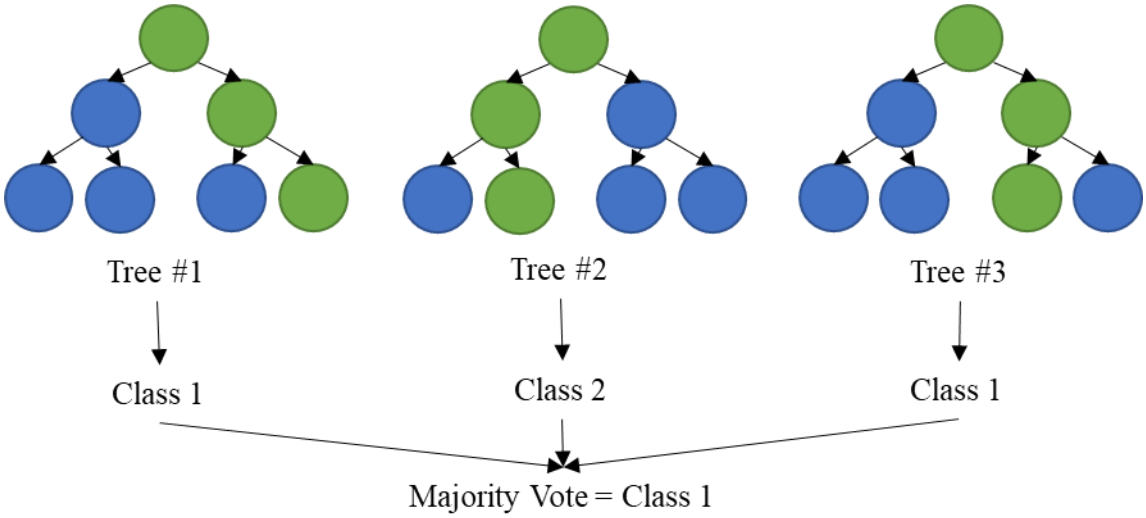

Figure 5 Visual representation of a random forest. Multiple decision trees are built and used to make predictions on a given sample. The majority vote across all trees is used as the prediction.

2.8 Machine Learning with Artificial Neural Networks

Feedforward ANN are a machine learning strategy that can be loosely compared to the human brain. Neurons are interconnected to create networks that can take multiple inputs, perform calculations on the inputs, and provide any number of outputs. An arrangement of neurons feeding forward to a single layer of output neurons is known as a perceptron. Perceptrons take several binary inputs, and produce one binary output [79]. Outputs from a perceptron are based on the weighted importance (w_j) of the inputs (x_j) (Eq. 3) (Figure 6) [80]. Neural networks are composed of layers of perceptrons, with each layer's perceptrons providing outputs to the next layer: an input layer, hidden layers, and an output layer [81]. The input layer is the initial data being provided to the network. The output layer produces the desired results from the input (i.e., final classification). The hidden layers are the layers where most of the decision making is performed.

$$\text{output} = \begin{cases} 0 & \text{if } \sum_j w_j x_j \leq \text{threshold} \\ 1 & \text{if } \sum_j w_j x_j > \text{threshold} \end{cases} \quad \text{Eq. 3}$$

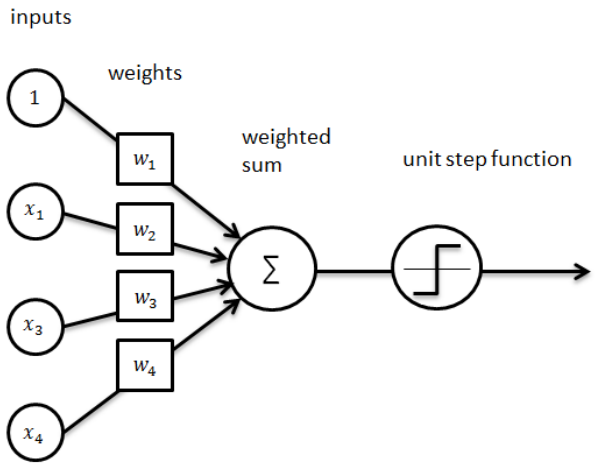

Figure 6 Single perceptron in neural network [80]

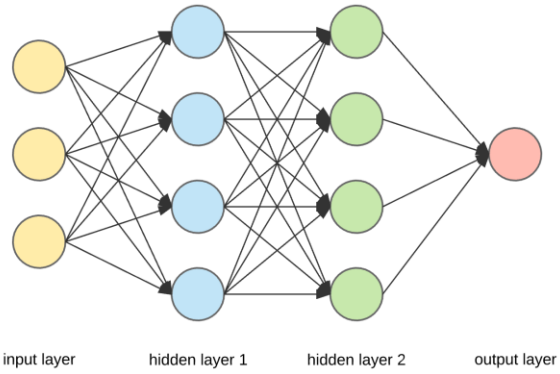

Figure 7 Neural network layers [80]

The number of hidden layers and the nodes in each layer influence the ANN's performance by increasing model complexity. Increasing model complexity can increase the model's capacity, potentially allowing the model to learn a larger set of mapping functions. While a model can hypothetically learn to approximate any mapping function with a single layer given enough nodes, adding layers can reduce total number of nodes and computation time while achieving the same model capacity [81]. This is because deeper layers can detect patterns between inputs that a single layer might not. Adjusting the number of layers, and the number of nodes in each layer, are the most intuitive and common parameter adjustments to optimize ANN models.

Although multiple hidden layers can be used, only one input and output layer are present (Figure 7). By varying the weights and thresholds within each perceptron, the overall model for decision making in the ANN changes. These weights are adjusted according to an optimization algorithm and loss function, which adjusts weights to reduce error while training.

2.8.1 **Optimization Algorithms**

Calculating the exact weights that achieve the best results in a neural network is unrealistic because there are too many unknowns [82]. Instead, the problem of selecting weights in a neural network is treated as a numerical methods type optimization problem. The goal when training ANN is to minimize a loss function by finding the optimized values for weights. However, the algorithm should also generalize well. Gradient descent, an iterative machine learning optimization algorithm to reduce the cost function [83], indicates the direction needed to reach a minimum in the loss

function. Optimizers update parameters in the negative gradient direction to minimize the loss function and determine the best weights.

An important aspect of gradient descent is learning rate. Learning rate is the size of step used to reach the global minimum [84]. Larger learning rates train models with larger changes to weights and bias values for each iteration. If the learning rate is too high, the iteration may overshoot the loss function's global minimum. If the learning rate is small, convergence takes longer and convergence may be at a local minima rather than the global minimum [84].

Many types of optimizers have been developed, each with their own advantages and disadvantages. The main differences are how the gradient is calculated and the way learning rates are applied.

2.8.1.1 **Adagrad**

Adagrad is an adaptive gradient algorithm that adapts the learning rate for each parameter individually. In this way, Adagrad reduces the need to manually tune learning rate. Unfortunately, Adagrad accumulates the sum of squares for the past gradients in the denominator and, since each term is positive, the learning rate can eventually become infinitely small [83].

2.8.1.2 **RMSProp**

Root Mean Square Propagation (RMSProp) tries to improve on Adagrad by using a moving average of the squared gradient. Because RMSProp is based on Adagrad, this optimizer automatically adjusts the learning rate per parameter, meaning the initially chosen learning rate is modified over the course of training and is less important than with other algorithms.

2.8.1.3 **Adam**

The most popular and often considered most effective optimizer is called Adaptive Moment Estimation (Adam) [84]–[86]. Adam combines Adagrad and RMSProp, but uses an exponential moving average of gradients to scale the learning rate instead of a simple average as in Adagrad [83]. Adam is computationally efficient, needs very little memory, is appropriate for noisy or sparse gradients, and has intuitive hyper parameters typically requiring little tuning [86].

2.8.2 Loss Functions

To evaluate error in an ANN solution, a loss function (i.e., the loss) is calculated [82]. By minimizing the loss, ANN classification results should ideally improve. This function must be mapped onto many dimensions; therefore, the loss function depends on the type of problem being addressed. Three classification types relate to loss function selection: regression, multi-class, and binary class [87].

In regression predictive modelling, a real-valued quantity is predicted. The default loss for regression problems is a Mean Squared Error Loss [87], the average of the squared differences between the predicted and actual values. Because the result is squared, larger mistakes result in larger errors, therefore punishing the larger mistakes.

For multi and binary class problems, discrete labels are assigned to each sample. The most common loss function for binary classification is a binary cross-entropy loss [87], [88]. Cross-entropy loss calculates a score that summarizes the average difference between the actual and predicted probability distributions for predicting class 1. The model can be trained by minimizing this score, meaning a perfect entropy value is 0. If multiple classes exist, categorical cross-entropy can be used [87].

2.8.3 Additional Hyperparameters

A variety of hyperparameters can be chosen or modified to influence ANN success such as, number of layers, number of nodes per layer, optimizer, learning rate, and loss function. However, other parameters can also be adjusted.

2.8.3.1 Initializers

When assigning initial weights and biases to a neural network, the values can be either set (usually zeroes or ones) or randomized [89]. When assigning weights, random values are preferred since weights initialized too high or too low can be mapped with small gradients, and learning can take longer.

2.8.3.2 **Batch size**

When training an ANN, the number of training samples used before the model's internal parameters are updated is called the batch size [90]. By adjusting the batch size, the number of samples used to calculate the loss and update model weights is changed. Regardless of batch size, each sample is used once per epoch to recalculate weights. Decreasing the batch size results in noisier results, which reduces generalization error but increases model training times [90], [91]. Larger batch sizes tend to converge on minima that are not generalizable.

2.8.3.3 **Epochs**

An epoch is comprised of one or more batches, such that once per epoch each sample is used to update the model, the loss function's gradient is calculated, and weights are reassigned to the model. The number of epochs required to train a model depends on the desired results. Training for longer should result in greater training accuracies and a smaller training loss. However, training for too long can lead the model to overfit to the training data, thus reducing performance on testing data. Optimizing the number of epochs can reduce the overall training time required and reduce overfitting. Using fewer epochs can improve model generalizability on validation data.

2.8.3.4 **Dropout**

Dropout is an approach to reduce overfitting by randomly ignoring or dropping out individual nodes [92]. When using dropout, each update during training is performed on a different layer configuration, which forces nodes within a layer to take on varying amounts of responsibility. This makes the training process noisy and reduces capacity but reduces overfitting by limiting the co-dependency of neurons and improving the individual power of each neuron.

2.9 Evaluation Techniques for Machine Learning Classifiers

Whether using random forests or deep learning neural networks, the goal of model evaluation is to provide a realistic prediction of how the model will perform on future data. Ideally, models should generalize, and perform on future data similarly to the training data. Two techniques for evaluating model performance are holdout and cross-validation. For both methods, data used to train the models are not used to evaluate the model. If all the data is used to train the model, the

model may simply remember the training set when being evaluated since it is the same data, leading to misleadingly good results [93], [94].

Holdout is a simple technique where some data is held back as a validation set (usually 20%) [95]. The training set typically includes most of the data. Training data is used to build the predictive models and the validation set is either used as one set or divided into a validation and test data set. Validation data is used to assess model performance and often to optimize the model. Model parameters can be tuned to maximize the validation set's evaluation metrics. While validation data is not used in model training, using validation results to optimize the model could lead to overfitting to the validation data. Therefore, a test data set is often used, which was not used for training or validation. If the model performs much worse on the test data set, then overfitting has likely occurred to both the training and validation data sets [96]. While holdout is simple to apply, an optimal split between test and validation data could be chosen by chance, which would result in much better or worse results than could be expected in practice.

Cross-fold validation is an improvement on standard hold-out techniques because it gives an average error across multiple hold-out tests [95], [97]. By dividing the data into k subsets, k different models can be built. For each model, one of the k subsets is used as a validation data set, and the remaining subsets as training data. The advantage of this is that each sample is used to build multiple models, and each sample is tested in one model. The logical extreme of a k-fold is to use all but one sample to construct a model, and then test on the remaining sample. This technique, called leave-one-out cross validation, is effective because it trains models using nearly all the data. The disadvantage to any cross-fold validation is the time required to test the model, which is approximately k times greater than the time required to test a model using holdout.

Stratified subsets should be used when dividing data for both holdout and cross-fold validation. Stratified subsets have a ratio between classes that is the same as the full data set. The goal of stratification is to ensure each fold is a good representative of the whole data set [98]. This becomes especially important when there is class imbalance or fewer samples, because models should be built using training sets that resemble validation sets as closely as possible.

2.9.1 Classification Metrics

Several metrics are available to determine the effectiveness of machine learning models. The most common metric is accuracy [99], determined by dividing the number of correct classifications by the total number of samples. However, accuracy does not assess model effectiveness on specific classes. A model can correctly classify all of one class, but none of another, and still have a good accuracy. This is especially true for imbalanced classes.

A confusion matrix can be used to visualize evaluation data (Figure 8) [99]. In an example using fall risk, a square contains true positives (TP) as FR correctly classified as FR, true negatives (TN) as NFR correctly classified as NFR, false positives (FP) as participants misclassified as FR, and false negatives (FN) as participants misclassified as NFR.

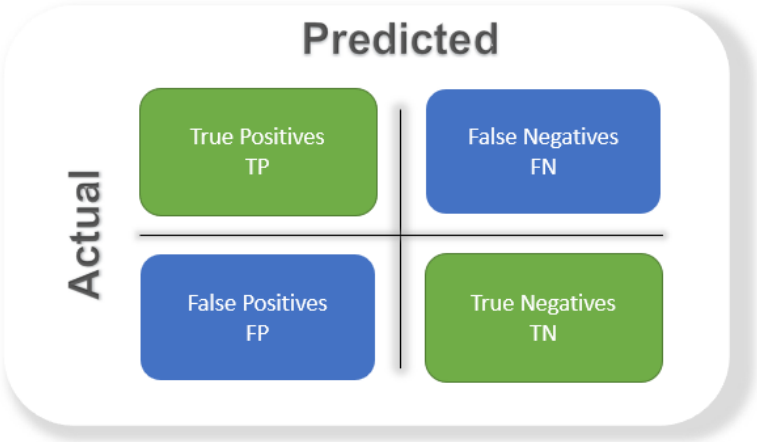

Figure 8 Example confusion matrix for binary classification problem.

Other evaluation metrics include sensitivity, specificity, Matthews Correlation Coefficient (MCC), and F1-score. Sensitivity, also called the recall, is an indicator of the model's ability to correctly classify the positive class and is calculated as the number of true positives divided by the total number of actual positives. Specificity indicates the model's ability to classify the negative class and is calculated by the number of true negatives divided by the total number of actual negatives. Sensitivity and specificity provide a clearer view of a model's strengths.

Precision describes how often the model is correctly classifying the positive class when it says it is classifying the positive class. Precision is calculated as the number of true positives

divided by the number of total positive predictions [99]. The harmonic mean of precision and sensitivity (or recall) is called the F1 Score and can be used to describe the classifier's ability to classify the positive class. F1 score does not consider the number of correctly classified negative samples [99] and is calculated as:

$$F1 = \frac{2 * precision * recall}{precision + recall} \quad \text{Eq. 4}$$

These metrics have various difficulties. As mentioned, accuracy can be misleading with class imbalance. Precision and F1-score can have asymmetric results [100], where the class selected as the "positive class" changes the results. Therefore, it is important to use metrics such as sensitivity and specificity since they remain the same regardless of which class is labeled positive, and only swap values if the positive class is redefined as the negative class.

MCC [100] remains the same regardless of which class is chosen as the positive class, and considers both classes equally regardless of class imbalance. MCC is a value between -1 and 1, where the closer the value is to -1 or 1, the better the correlation and the better the classifier prediction. MCC is calculated as

$$MCC = \frac{(TP * TN - FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad \text{Eq. 5}$$

By using combinations of metrics, an overall summary of classifier success can be made. A summed ranking method [8] can be used, where classifiers are ranked from best (1) to worst (N where N=total number of classifiers) for each metric, the scores for each classifier are summed, and then the classifiers are sorted based on the sum of ranks. The lowest summed ranking is then deemed the best classifier. Using this technique, classifiers can be chosen for overall effectiveness without undue influence by class imbalance.

2.10 Literature Review Summary

Through this literature review, the need for fall-risk assessment models specific to amputee populations was presented. Wearable technology systems could be developed to augment and improve results of existing clinical tests. These systems should minimize the number of sensor locations to make point-of-care analysis efficient and easy. Smartphone are a viable approach for this application, where a single smartphone can collect both straight and turning data during a 6MWT. This is important since previous research has demonstrated that fall-risk prediction improves with the inclusion of turn data. Feature based classification models such as random forests and neural networks have been successful for fall-risk classification with older adults and are thus viable options for research on amputee fall risk assessment.

3 Evaluating Random Forests for Fall-Risk Classification in Lower Limb Amputees Using Features Extracted from Smartphone Sensor Data

This chapter addresses Objective 1 by training a random forest model, evaluating model performance, and comparing results to existing clinical assessment methods and previous wearable-sensor fall-risk classification research. Methods for feature extraction and selection are explained, and then feature subsets were used to build random forest models. The best random forest model was trained on lower limb amputee features and achieved better sensitivity and specificity than previous models developed for older adults, and better than existing clinical assessment tools.

The content of this chapter was submitted for publication to Plos One.

Kyle J.F. Daines[1*], Natalie Baddour[2], Helena Burger[3,4], Andrej Bavec[3,4], and Edward D. Lemaire[1]. Evaluating random forests for fall-risk classification in lower limb amputees using features extracted from smartphone sensor data.

3.1 Abstract

Fall-risk classification is a challenging but necessary task to enable the recommendation of preventative programs for individuals identified as a fall-risk. Existing research has primarily focused on older adults, with no predictive fall-risk models for lower limb amputees, despite their greater likelihood of fall-risk than older adults. In this paper, 89 amputees with varying degrees of lower limb amputation performed a 6 minute walk test (6MWT) with an Android smartphone placed in a holder located on the back of the pelvis. A fall-risk classification method was developed using data from sensors within the smartphone. The Ottawa Hospital Rehabilitation Center Walk Test app captured accelerometer and gyroscope data during the 6MWT. From this data, foot strikes were identified, and 248 features were extracted from the collection of steps. Steps were segmented into turn and straight walking, and four different data sets were created: turn steps, straightaway steps, straightaway and turn steps, and all steps. From these, three feature selection techniques (correlation-based feature selection, relief F, and extra trees classifier ensemble) were used to eliminate redundant or ineffective features. Each feature subset was tested with a random forest classifier and optimized for the best number of trees. The best model used turn data, with three features selected by CFS, and used 500 trees in a random forest classifier. The resulting metrics were 81.3% accuracy, 57.2% sensitivity, 94.9% specificity, a Matthews correlation coefficient of 0.587, and an F1 score of 0.83. These metrics make the classifier viable for use in clinical practice and are comparable to the metrics achieved by existing clinical tests.

3.2 Introduction

Falling is the second leading cause of accidental injury death [1]. Injuries related to falling can be debilitating, life-altering, and lead to lack of confidence when walking. Although preventative programs exist to reduce the chance of falling, the initial task of fall-risk identification can be challenging. While more than 26 fall risk assessment tools are available for clinicians [39], evolving wearable sensors systems can provide opportunities to augment common clinical mobility tests for fall risk identification.

Wearable technology has been used to develop fall-risk classifiers that provide accurate and automated fall risk classification, to enable timely intervention with fall-risk mitigation techniques. Previous research has used a variety of sensors and technology, such as inertial sensors, for fall

risk classification [7]–[10], [15], [16]. Features can be extracted from data collected during clinical tests such as the 6-minute walk test (6MWT) [7] or 10 meter walk test [8]. The 6MWT includes straight and turn walking, measures functional capacity, and is widely used to measure the response of therapeutic interventions [23]. Research reports differences in turning strategies between fall-risk and no fall-risk populations, and that viable elderly fall-risk classification can be achieved with turn data [7], [16]. If fall risk classification can be determined from the 6MWT, clinic time could be more efficiently used by reducing the number of tests in a session.

Previous studies have used sensors at multiple locations on the body. This can be problematic in the clinic due to the time required to setup sensors on the person, sensors falling off or sliding, and patients changing how they move due to too many sensors, and accuracy of their position. Therefore, the number of sensors should be minimized, and sensors should have easy and reliable placement on the body. For example, the pelvis was reported as the best single sensor location for fall-risk classification [8], and the most commonly used location [11]. Sensors at the pelvis during a 6MWT could be an effective data source for fall-risk classification.

The fall-risk classification literature has focused primarily on elderly populations, with no predictive fall risk models for lower limb amputees. However, lower limb amputees are at higher risk of falling compared to able-bodied and other clinical populations in all phases of rehabilitation [12], [13], and the risk of fall related injuries requiring medical care can be higher than older adults [14]. Since predictive models developed for able-bodied or elderly populations do not necessarily translate to the lower limb amputee population, research is needed to determine the best features and models to perform fall-risk classification in amputees and ensure that similar results can be achieved when compared to existing assessment tools.

3.3 Methods

3.3.1 Participants

A convenience sample of 129 participants with lower limb amputations were recruited from the University Rehabilitation Institute (Ljubljana, Slovenia). Clinical records provided self-reported number of falls, with falling at least once in the past six months prior to testing considered fall risk. For this study, data from 89 participants (19 female, 70 male, age 62.3 ±12.5) were suitable

for fall risk classification (32 fall-risk, 57 no fall-risk). Participants included 4 bilateral transtibial amputees, 1 bilateral transtibial and transfemoral, 63 transtibial, 18 transfemoral, 2 knee exarticulation, and 1 ankle exarticulation. Reasons for unsuitable data were malfunction exporting data from phone (5 people), no fall incidence data on file (9 people), running instead of walking during the 6MWT (1 person), or had unidentifiable foot strikes due to highly irregular gait (2 people), a single crutch (2 people), double crutches (18 people), or non-rolling walker (3 people).

3.3.2 **Equipment**

An Android smartphone was affixed to the midline of the posterior pelvis using a waist belt (Figure 9). Participant demographics and information were input into a custom designed The Ottawa Hospital Rehabilitation Center (TOHRC) Walk Test app [53] (Figure 10). Each participant performed a 6MWT along a 20m hallway (i.e., walk, turn around a cone, and continue the circuit for 6 minutes). The TOHRC Walk Test app collected smartphone 3D accelerometer and gyroscope raw data, and pelvic rotation, tilt, and obliquity at 50 Hz. Each trial was also video recorded. Once the test was complete, data were exported from the smartphone to a text file for post-processing.

Figure 9 6MWT with smartphone on posterior pelvis

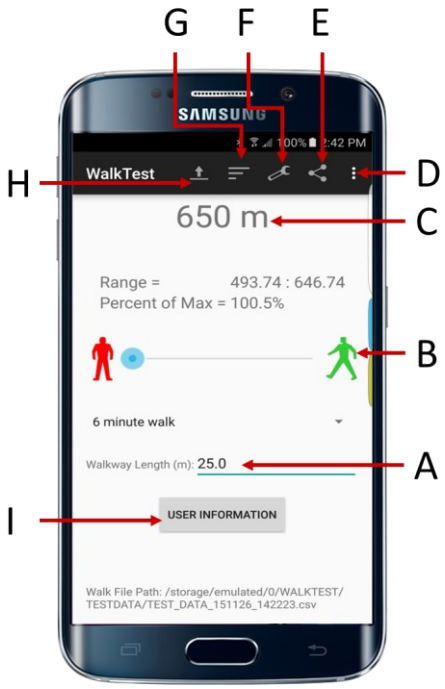

Figure 10 TOHRC Walk Test application. (A) Walkway length; (B) Start trial, (C) Distance walked; (D) About, Help; (E) Share output; (F) Settings; (G) Results tables; (H) Load previous data; (I) Enter patient demographics.

3.3.3 **Step and Turn Segmentation**

Raw accelerometer data, gyroscope data, and smartphone orientation were imported to MATLAB 2013b, along with the time stamps for each recording. Foot strikes were identified from anterior-posterior (AP) linear acceleration (Figure 11 and Figure 12), using the peak value near the estimated next step, based on average step duration. AP acceleration had the least variance when compared to other linear acceleration axes. However, this automated technique for able-bodied gait [53] sometimes failed to select the correct peak with lower limb amputees due to amputee participant's more asymmetric and variable gait (e.g., Figure 11). In these cases, manual step identification was required. The manually cleaned data was used to extract features.

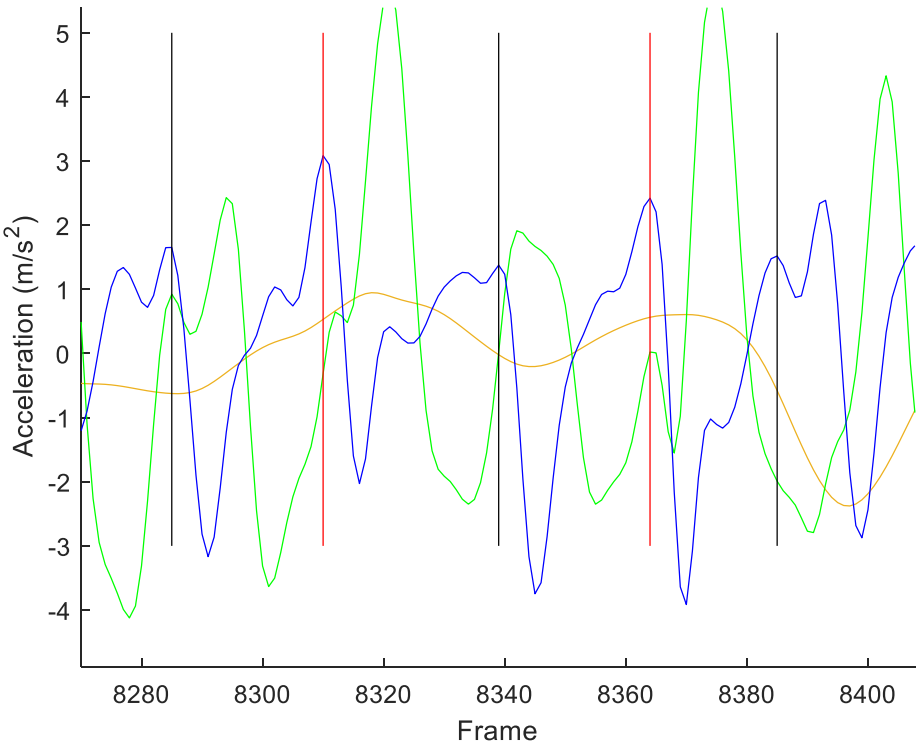

Figure 11 Anterior-posterior (blue), vertical (green), and medial-lateral (yellow) accelerations for straight steps. Black lines indicate right steps, red lines indicate left steps. Manual step identification was required around frame 8340.

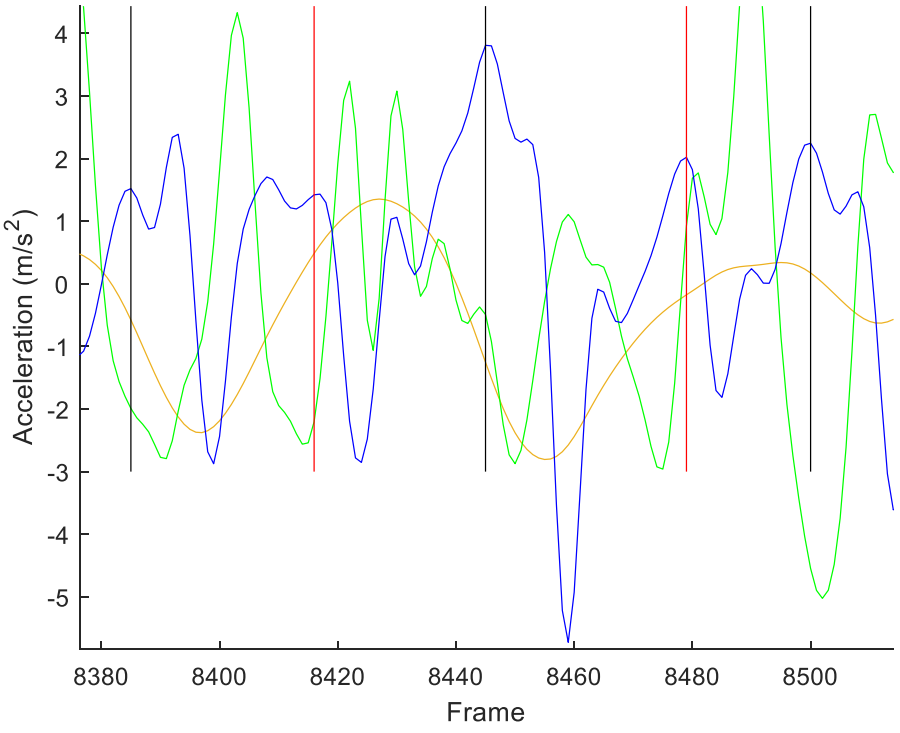

Figure 12 Anterior-posterior (blue), vertical (green), and medial-lateral (yellow) accelerations for turning steps. Black lines indicate right steps, red lines indicate left steps.

In previous research, data from turn walking was better at classifying fall risk than data from straight walking [7]. Therefore, data were segmented into turns and straightaways. Differences between straight and turn steps (Figure 12) include greater ML accelerations and a greater AP acceleration peak in the middle of the turn. Turns were defined as the five steps around the middle of each turn, and straightaways were all other steps. The center of a turn was identified using pelvis rotation, by using the middle frame between the beginning and end of a turn (i.e., when the pelvis started to rotate and when it stopped rotating). Similar to foot strikes, this process was first automated in MATLAB, then verified manually.

Once turn and straightaway steps were identified, four feature sets were created:

1. Only turn step (T) features (248 features)
2. Only straightaway step (S) features (248 features)
3. Straightaway step features and turn step (S&T) features (496 features)
4. All steps (AS) (248 features) (no distinguishing between straight and turn)

3.3.4 **Feature Extraction**

Based on existing literature [7], [11], features were extracted from linear acceleration and angular velocity signals in each step. 62 features were extracted for the four feature sets:

Temporal: cadence, step time (foot strike to foot strike of the opposite foot), stride time (foot strike to foot strike of the same foot), symmetry in right and left limb step times (symmetry index) [101].

Descriptive statistics: minimum, maximum, mean, standard deviation, root mean square in three axes (vertical, medial lateral (ML), AP) for pelvis linear acceleration (Android processed signal, not including gravity) and tilt, rotation, and obliquity angular velocities.

Linear acceleration and angular velocity frequency domain features: from the absolute value of the Fast Fourier transform (FFT) of each step, the first quartile of Fourier transform (FQFFT), ratio of even/odd harmonics (REOH), and peak distinction.

FQFFT: percentage of frequencies within the first quartile of the Nyquist frequency (6.25 Hz was used as the first quartile). Lower FQFFT values indicate more high frequency components, linked to instability [56].

REOH: ratio of frequencies in even harmonics compared to the odd harmonics (using stride time as the fundamental frequency). Lower REOH values have been associated with fall risk [7], [57].

Peak distinction: to determine if the FFT peak frequency was distinct from other frequencies, the percent of frequencies in the FFT with power greater than a threshold (⅓ amplitude of peak signal) was calculated. A lower peak distinction value means a more distinct peak (Figure 13).

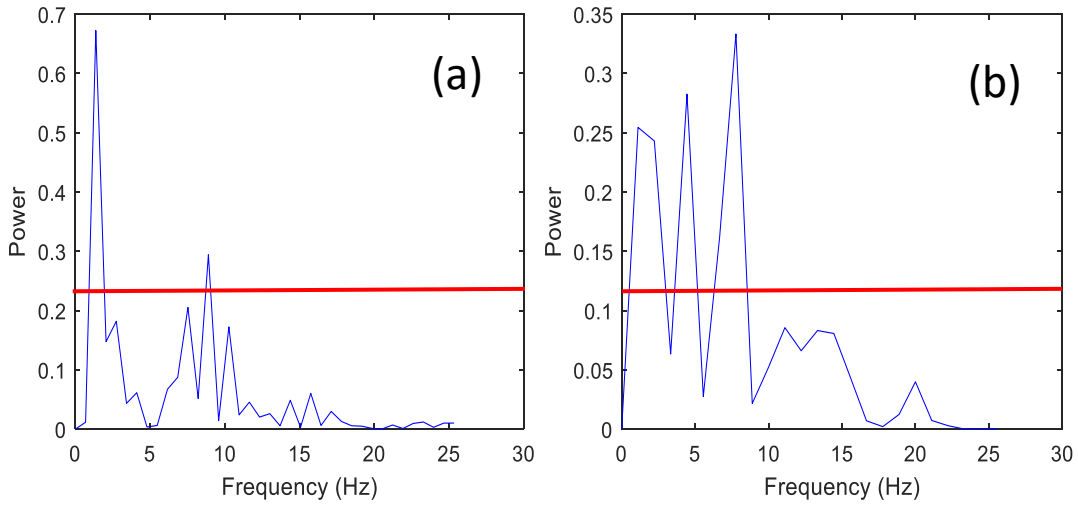

Figure 13 Peak distinction for two steps. (a) A step with a very distinct peak, having a peak distinction of 5.26%. (b) A step with a less distinct peak, having a peak distinction of 20.83%. Horizontal red lines indicate one third of max amplitude.

Once features were extracted for each step, the minimum, maximum, mean, and standard deviation were calculated over all included steps for a total of 248 features (62 multiplied by 4 statistics) per data set (496 for the S&T data set).

The features used in this thesis differed from the elderly fall-risk feature set from Drover et al. [7]. Since Drover only had accelerometers, features were extracted from linear acceleration; however, multiple accelerometer locations were used (i.e., shanks and pelvis). The feature set with the best classification results used minimum of anterior-posterior ratio of even/odd harmonics for right shank, standard deviation (SD) of anterior left shank acceleration SD, SD of mean anterior left shank acceleration, maximum of medial-lateral first quartile of Fourier transform (FQFFT) for lower back, and maximum of anterior-posterior FQFFT for lower back.

3.3.5 **Feature selection**

Feature selection was used to reduce feature space dimensionality, simplifying the problem by removing redundant and irrelevant data [7], [8], [102]. Three feature selection techniques were used, based on previous success in fall risk classification: Correlation-based feature selection (CFS) [59], [64], Relief-F (RelF) [60], [103], and extra trees classifier ensemble (ETC) [7], [69].

CFS is a supervised, filter-based method that identifies a subset of features that are correlated with the class label (i.e., fall-risk or no fall-risk), but also uncorrelated to other parameters by calculating the "merit" based on pair-wise correlations [65], [66]. This allows CFS to develop a subset that has no irrelevant or redundant features, by only adding features that improve the subset's merit.

RelF is a supervised method that ranks features by weighting them based on their relevance and how well instances from different classes and the same class can be distinguished [60], [65]. RelF does not eliminate redundant features, making this method most useful when evaluating parameters with interdependencies.

ETC is an ensemble method that fits a number of randomized decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting [69]. Each feature is then ordered based on their importance and the best features can be selected. An ETC ensemble method was used in this research because, while they are very similar to random forests, ETC computes using randomly selected weightings, making ETC better for feature selection [72].

For RelF and ETC, features sets were created for the top 30, 20, 10, and 5 features. CFS selected less than 5 features for all sets.

3.3.6 **Classification Techniques and Optimization**

A random forest classifier with 100 trees and a leave-one-out strategy was applied to the data set. Random forests consist of many decision trees operating as an ensemble, making them preferable to a single decision tree. Each tree provides a predicted class, and the classifier takes a majority vote to decide which class the model should predict [104]. The belief is that the trees are

uncorrelated so that operating as a committee allows them to outperform any of the individual models.

Five evaluation metrics were used for evaluating the models: accuracy, sensitivity, specificity, Matthews correlation coefficient (MCC), and F1 score. The best five "feature selector – data set combinations" were chosen based on a ranking technique similar to [8] and [105]. Each classifier was ranked in the five evaluation metrics, and the lowest summed rankings were chosen as the top five classifiers, with five being the lowest possible summed ranking since five metrics were used. These five models were then optimized for the number of trees that provided the highest accuracy by testing increments from 5 to 1000 trees. More trees perform better with minimal risk of overfitting, although more trees increase computation times [27]. To test model robustness, each of the five optimized models were built 10 times with different random seeds in a leave-one-out strategy to determine the mean and standard deviation for accuracy, sensitivity, specificity, and MCC score.

3.4 Results

3.4.1 Feature Selection

Each feature selector chose different features, creating a range of subsets (Table 1, Table 2). For turn data, CFS chose only three features but provided the best overall feature set (vertical acceleration maximum standard deviation, AP acceleration minimum peak distinction, and tilt angular velocity minimum peak distinction). Straight walking CFS only chose one feature, standard deviation of vertical acceleration's standard deviation. S&T and AS CFS chose combinations of these 4 features, so that all CFS subsets selected similar features.

AP linear acceleration minimum peak distinction was the only feature selected in the top ten by all three feature selectors, for turn data. More fall-risk participants had lower peak distinction for minimum AP linear acceleration and tilt angular velocity, meaning that more fall-risk participants had distinct FFT peaks.

Table 1 Legend of feature descriptions for feature sets in Table 2.

Variable	Statistic (Feature #)	Variable	Statistic (Feature #)
LR symmetry	Min (1), Mean (3), Std dev (4)	FQFFT rotation angular velocity	Min (121), max (122), mean (123)
Maximum ML acceleration	Max (6)	FQFFT obliquity angular velocity	Max (126), Mean (127), Std Dev (128)
Maximum Vertical acceleration	Min (11), Std Dev (12)	Maximum of ML acceleration FFT	Min (129)
Maximum AP acceleration	Min (13), Max (14), Mean (15), Std Dev (16)	Maximum of Vertical acceleration FFT	Max (134)
Minimum Vertical acceleration	Max (22), Std Dev (24)	Maximum of AP acceleration FFT	Mean (139)
Minimum AP acceleration	Max (26)	Maximum of tilt angular velocity FFT	Std Dev (144)
Mean ML acceleration	Mean (31), Std Dev (32)	Maximum of obliquity angular velocity FFT	Min (149), Mean (151)
Mean Vertical acceleration	Std Dev (36)	Standard deviation of ML acceleration FFT	Min (153)
Mean AP acceleration	Min (37)	Standard deviation of AP acceleration FFT	Min (161), Mean (163)
Standard deviation of ML acceleration	Min (41), Mean (43)	Standard deviation of tilt angular velocity FFT	Min (165)
Standard deviation of vertical acceleration	Max (46), Mean (47), Std dev (48)	Standard deviation of rotation angular velocity FFT	Mean (171)
Standard deviation of AP acceleration 49,50	Min (49), Max (50)	Standard deviation of obliquity angular velocity FFT	Mean (175), Std Dev (176)
Tilt angular velocity range	Min (53), Std Dev (56)	Peak distinction of ML acceleration	Min (177), Max (178), Mean (179)
Rotation angular velocity range	Max (58)	Peak distinction of vertical acceleration	Min (181), Max (182), Mean (183), Std Dev (184)
Tilt angular velocity mean	Min (65), Max (66)	Peak distinction of AP acceleration	Min (185)
Rotation angular velocity mean	Min (69), Std Dev (72)	Peak distinction of tilt angular velocity	Min (189), Max (190)
Obliquity angular velocity mean	Max (74), Mean (75)	Peak distinction of rotation angular velocity	Min (193)
Tilt standard deviation	Std Dev (80)	Peak distinction of obliquity angular velocity	Min (197), Max (198)
Rotation standard deviation	Min (81), Mean (83)	REOH of ML acceleration	Min (201), Max (202)
Step timing	Max (90), Mean (91), Std Dev (92)	REOH of vertical acceleration	Min (205), Max (206), Mean (207), Std Dev (208)
Stride timing	Min (97), Max (98), Mean (99), Std Dev (100)	REOH of tilt angular velocity	Min (213)

Cadence	Min (101), Mean (103), Std Dev (104)	REOH of rotation angular velocity	Min (217), Max (218)
FQFFT ML acceleration	Min (105), Std Dev (108)	REOH of obliquity angular velocity	Max (222), Std Dev (224)
FQFFT vertical acceleration	Std Dev (112)	Root mean squared of ML acceleration	Std Dev (228)
FQFFT AP acceleration	Std Dev (116)	Root mean squared of AP acceleration	Min (233)
FQFFT tilt angular velocity	Min (117), Mean (119)		

Table 2 Top 30 selected features for each feature selector. T=Turn, S=Straight, S&T=Straight and Turn, AS=All steps. Refer to Table 1 for feature descriptions.

T-CFS	**T-RelF**	**T-ETC**	**S-CFS**	**S-RelF**	**S-ETC**	**S&T-CFS**	**S&T-RelF**	**S&T-ETC**	**AS-CFS**	**AS-RelF**	**AS-ETC**
46	100	65	48	90	171	48 (S)	100 (T)	193 (S)	46	98	75
185	90	207		92	103	46 (T)	90 (S)	98 (S)	48	189	101
189	91	98		56	121	185 (T)	90 (T)	183 (S)	185	90	32
	92	15		16	205	189 (T)	56 (S)	179 (T)	189	99	199
	207	176		12	108		91 (T)	121 (S)		92	194
	98	100		99	48		91 (S)	153 (T)		97	10
	72	103		183	182		12 (S)	100 (T)		56	241
	185	119		91	105		185 (T)	32 (S)		91	33
	97	185		97	161		98 (T)	153 (S)		185	143
	206	72		101	16		99 (S)	97 (T)		101	114
	101	228		98	151		92 (T)	56 (S)		183	189
	177	90		4	149		92 (S)	144 (S)		103	108
	32	197		117	189		98 (S)	97 (S)		4	36
	3	41		184	101		16 (S)	6 (T)		69	98
	122	165		103	26		117 (S)	98 (T)		126	116
	201	217		80	181		183 (S)	58 (S)		12	103
	99	189		48	116		207 (T)	128 (T)		201	15
	218	144		202	32		99 (T)	14 (T)		16	92
	190	139		32	6		100 (S)	184 (S)		32	193
	208	213		189	49		72 (T)	98 (S)		100	201
	37	1		193	22		201 (T)	222 (T)		122	44
	98	50		119	75		206 (T)	112 (S)		24	149
	4	134		66	207		99 (T)	99 (S)		217	88
	126	175		190	53		101 (T)	47 (S)		2	159
	128	105		198	178		101 (S)	163 (S)		48	65
	13	43		181	81		97 (T)	112 (T)		190	14
	127	104		197	31		4 (S)	151 (T)		181	83
	1	129		98	202		218 (T)	233 (S)		80	48
	69	83		36	11		177 (S)	24 (S)		192	188
	123	206		127	74		177 (T)	224 (S)		198	225

3.4.2 **Model Optimization**

Table 3 shows the unoptimized results for fall risk classification using each feature selector and a random forest classifier with 100 trees.

Table 3 Unoptimized ranked metrics for top 10 subsets with a random forest classifier. ETC## and RelF## indicate the number of features from that subset. T=Turn, S=Straight, S&T=Straight and Turn, AS=All steps, MCC= Matthews Correlation Coefficient, F1=F1 score, SR=summed ranking.

Feature Selector	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**MCC**	**F1**	**SR**
T-CFS	78.7	53.1	93.0	0.521	0.642	6
S&T-CFS	74.2	53.1	86.0	0.417	0.596	16.5
S-ETC10	70.8	43.8	86.0	0.331	0.519	24.5
S-RelF10	70.8	40.6	87.7	0.326	0.500	27
AS-RelF30	69.7	40.6	86.0	0.301	0.491	37.5
AS-ETC30	69.7	40.6	86.0	0.301	0.491	37.5
T-ETC30	69.7	37.5	87.7	0.295	0.471	45
AS-CFS	68.5	50.0	78.9	0.299	0.533	54.5
S&T-RelF10	67.4	46.9	78.9	0.270	0.508	63.5
StT-RelF30	68.5	34.4	87.7	0.264	0.440	65

The top five models were rebuilt using different numbers of trees, ranging from 5 to 1000. Results for the best model (T-CFS) are shown in Table 4. Table 5 provides a summary for all the best trees.

Table 4 Tree optimization for best model (T-CFS). T=Turn, MCC= Matthews Correlation Coefficient, F1=F1 score

T-CFS	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC	F1
5	76.4	46.9	93.0	0.467	0.588
10	79.8	59.4	91.2	0.547	0.679
25	71.9	46.9	86.0	0.360	0.545
50	75.3	53.1	87.7	0.442	0.607
75	77.5	53.1	91.2	0.493	0.630
100	78.7	53.1	93.0	0.521	0.642
150	79.8	59.4	91.2	0.547	0.679
200	79.8	50.0	96.5	0.555	0.640
250	80.9	59.4	93.0	0.574	0.691
300	79.8	56.3	93.0	0.548	0.667
400	82.0	56.3	96.5	0.606	0.692
500	82.0	56.3	96.5	0.606	0.692
1000	80.9	56.3	94.7	0.576	0.679

Table 5 Summary of tree optimizations for the 5 best models. T=Turn, S=Straight, S&T=Straight and Turn, AS=All steps, MCC= Matthews Correlation Coefficient, F1=F1 score

Feature Selector	Trees	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC	F1
T-CFS	500	82.0	56.3	96.5	0.61	0.69
S&T-CFS	500	80.9	56.3	94.7	0.58	0.69
S-ETC10	250	71.9	46.9	86.0	0.36	0.55
S-RelF10	50	71.9	50.0	84.2	0.37	0.56
AS-RelF30	150	70.8	50.0	82.5	0.34	0.55

Once the optimal number of trees was determined, ten models with different random seeds were built for each model to determine robustness (Table 6).

Table 6 Final mean and standard deviation (in brackets) metrics for optimized models based on 10 random seeds. T=Turn, S=Straight, S&T=Straight and Turn, AS=All steps, MCC= Matthews Correlation Coefficient, F1=F1 score

Feature Selector	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC	F1
T-CFS	81.3 (1.09)	57.2 (2.11)	94.9 (1.29)	0.59 (0.027)	0.69 (0.019)
S&T-CFS	78.9 (1.38)	55 (1.61)	92.3 (1.48)	0.53 (0.033)	0.65 (0.021)
S-ETC10	69.3 (1.5)	41.6 (3.31)	84.9 (0.91)	0.30 (0.038)	0.49 (0.032)
S-RelF10	68.7 (2.4)	44.7 (4.43)	82.1 (3.07)	0.29 (0.055)	0.51 (0.039)
AS-RelF30	69 (1.42)	43.4 (3.74)	83.3 (1.49)	0.29 (0.036)	0.50 (0.031)

3.5 Discussion

This research demonstrated that a random forest classifier with smartphone sensor data collected at the posterior pelvis can provide viable fall-risk classification for lower extremity amputees that completed a 6MWT. The best model had 81.3% accuracy, 57.2% sensitivity, and 94.9% specificity. The very high specificity showed that the model had a low chance of false positives, indicating that if the model has a low chance of inappropriately classifying a person as a faller. This almost 95% specificity was higher than other clinical tests that focus on fall risk in amputees [37]. This is important for health and long-term care systems where appropriate resource allocation is essential.

More than half the people with amputations who are at risk of falling would be properly identified (i.e., 57.2% sensitivity), which is an interesting result considering that the 6MWT was not designed as a fall risk measure. However, this sensitivity was lower than other clinical fall risk tests. Two common tools that can be applied in a clinic are the Four Square Step Test (FSST) and Timed Up and Go (TUG). In one study predicting multiple falling (2 or more falls in 6 months), FSST with lower limb amputees had a predictive sensitivity of 92% and specificity of 93% in amputees using a cut-off time of 24 seconds, and TUG had a predictive sensitivity of 85% and specificity of 74% [37]. However, the sensitivity and specificity could have been high due to the 2 or more falls criteria for fallers, since this group may have had consistently poorer TUG performance than people who have only fallen once. Instead, a review of fall risk assessments found that the Time-up and Go (TUG) test has a predictive sensitivity of 76% and specificity of 49% on older adults [39].

The sensitivity and specificity of the best model from this study were comparable to results from the TUG test, and requires only a 6MWT, rather than a separate TUG test to identify fall-risk. Therefore, the 6MWT approach cannot be considered as a surrogate for other fall risk tests. However, people classified as fall-risk individuals can be identified confidently in a clinic using a 6MWT approach without requiring additional testing, due to the test's high specificity. If other clinical indicators of fall risk are present, and the person was identified as a no fall risk individual from the 6MWT, other clinical fall-risk tests can be performed as indicated.

Random Forest classifiers using feature-selectors have been effective in previous studies on older adults. Using these techniques on an amputee population's pelvis sensor data from a 6MWT provided similarly effective outcomes for fall risk classification. A previous study on older adults who completed a 6MWT with accelerometers located at the pelvis and ankles achieved 73.4% accuracy, 60.5% sensitivity, and 82.0% specificity [7]. The outcomes from this study agreed with previous work that turn data was better for fall-risk identification, but the models generated for amputee participants had a higher specificity and accuracy. It was important to examine older adults and people with amputations separately since gait patterns differ between these groups and amputee populations have a higher fall risk than older adult populations.

While previous studies typically used multiple sensor locations (e.g., accelerometers at pelvis and shanks for older adults [7]), this research only use smartphone sensors at the posterior pelvis. A single pelvis location provides an approach that is efficient to apply and easily repeatable in the clinic. The proposed smartphone-based method could have better chance of knowledge translation at the point of patient contact. Fall risk classification results could be provided to the clinician immediately following the test by including the model in the smartphone 6MWT application, thereby supporting clinical decision-making with instant reporting.

3.5.1 **Features**

Feature selection generally improved classification results since sets with no feature selection were in the bottom 20%. CFS provided noticeably better results than other feature selection techniques, except with straight data where only one feature was selected. Both RelF30 and ETC30 achieved accuracies above 65% for all data sets, and these feature sets were ranked in the top half of all models. However, four of the top five feature subsets had ten or fewer features.

While most feature subsets achieved good specificities, smaller feature subsets also had good sensitivities. Smaller subsets may have led to less data overfitting and therefore better fall-risk classification.

The most selected feature for turn data was AP acceleration minimum peak distinction (i.e., most distinct FFT peak). For AP acceleration, if the FFT had one predominant frequency, the peak would have been more distinct since one peak FFT amplitude would have a noticeably greater than the others. Participants who were a fall-risk were more likely to have more distinct peaks.

T-CFS model had the best classification results and included two minimum peak distinctions and one maximum standard deviation. The second-best model (S&T-CFS) used the same features as T-CFS, but also included straight walking vertical acceleration's standard deviation of the standard deviation. Interestingly, including this single feature decreased all outcome metrics by over 2%.

Turning while walking can be more challenging for people with mobility disabilities, so it is intuitive that a model using turn data provided the most successful classifier. This is consistent with results from an elderly population [7]. Therefore, turn steps should be used for 6MWT-based faller classification.

3.5.2 **Models**

The best classification model was T-CFS, closely followed by S&T-CFS. Model performances were also similar to clinical functional assessment tools [39], making the 6MWT smartphone approach a good tool for clinical fall-risk identification. Since most models had relatively high specificity, the sensitivity results contributed most to the overall ranking. This demonstrates the importance of having multiple types of evaluation metrics. Metrics such as accuracy can be inflated due to class imbalance (this data set had 36% of the participants identified as fall-risk). Class imbalance is an unavoidable problem with fall-risk classification since less of the population is at risk of falling. Therefore, classifiers that are better at dealing with slight class imbalances, such as a random forest, should be considered for fall risk classification.

Initial testing with 100 trees resulted in the T-CFS and S&T-CFS performing better than other classifiers. Optimization by adding more trees improved results up to a plateau in

effectiveness around 100 or 200 trees. No additional improvements occurred after 500 trees. After the optimal number of trees was selected for each classifier, this optimal number of trees was tested ten times each to verify that the results were robust. Mean results for both CFS models by the end of optimization were better than unoptimized models.

A limitation of this study was that all participants were recruited from only one rehabilitation institute. Future research could consider participants from a variety of countries and clinics. Additionally, increasing the number of participants would help improve the model's generalizability and possibly improve model effectiveness.

Since the best model (T-CFS) had better specificity and sensitivity than TUG in older adults [39], it is reasonable to use T-CFS as a preliminary indicator for fall risk. This approach could help reduce the number of tests required for a complete functional assessment, since 6MWT are often performed during clinical evaluations and TUG may not be collected for people who can walk for 6 minutes. As more participants are added to the training set, this Random Forest Classifier approach should continue to improve and complement existing functional assessment tools to assist with fall-risk classification.

3.6 Conclusions

A novel smartphone sensor-based fall-risk classification method was developed to provide a sensor-based fall-risk classification for lower limb amputees. The best classification model used CFS on turn step features in combination with a random forest classifier. This model had very high specificity, leading to few false negatives. This is important so that patients are not mistakenly suggested into preventative programs. While 57% sensitivity indicated that more than half the people at risk of falling were appropriately classified, future research should aim to improve model sensitivity to identify more people who are at risk of falling. Turn steps have been found to be the best indicator of fall-risk in both lower limb amputees and older adults, making them consistently the best choice for fall-risk identification. Addition of a single straightaway step feature negatively affected the turn step classifier's results. The methods developed here for collecting data and classifying individuals can be easily implemented into clinical practice, making it a potential method to indicate if there is need for fall risk assessment tools such as a TUG test. By achieving fall-risk assessment during a 6MWT, the number of required functional mobility tests can be

reduced, thereby reducing patient time in clinics. Future work should continue to add more participants to the dataset, improving the classification metrics to ensure success in clinical implementation.

4 Evaluating and Optimizing Artificial Neural Networks for Fall-Risk Classification in Lower Limb Amputees

This chapter addresses Objective 1 by training ANN and optimizing several hyperparameters to achieve the best results. The best models are evaluated and compared to existing clinical assessment methods and previous research performing fall-risk classification with ANN on older adults. Many ANN models were trained, but none achieved better sensitivity and specificity than previous models developed for older adults, and after using five-fold cross validation, none were better than existing clinical assessment tools.

ANN models were investigated in depth by optimizing multiple hyperparameters to ensure the best model was achieved. Chapter 4 includes results and discussion for the feature subset that achieved the best results after optimizing for feature selection, layers, nodes, dropout, learning rate, and batch size. However, additional optimization was performed on smaller feature subsets to determine if overfitting was leading to poorer results. This research did not result in any improvements, and therefore was not included in Chapter 4. For completeness, these results are included and discussed in Appendix A: Further Optimization of Neural Network Results.

The content of this chapter will be submitted for publication:

Kyle J.F. Daines[1*], Natalie Baddour[2], Helena Burger[3,4], Andrej Bavec[3,4], and Edward D. Lemaire[1]. Evaluating artificial neural networks for fall-risk classification in lower limb amputees using features extracted from smartphone sensor data.

4.1 Abstract

Research has advanced fall-risk identification in older adults. However, the amputee population remains largely unaddressed, despite having a greater chance of falling than older able-bodied populations. This research investigated fall-risk classification for people with lower limb amputations. 89 people with varying levels of lower limb amputation performed a 6-minute walk test (6MWT) with an Android smartphone placed in a holder on the back of the pelvis. Sensor data were collected using TOHRC Walk Test application. Steps and turns were segmented, and 248 features were extracted from the steps. Feature selection techniques were applied, and each feature subset was tested with an artificial neural network (ANN) to achieve fall-risk classification. The best model, evaluated with 5 fold cross-validation, used the best 50 features from an Extra Trees Classifier (ETC), 2 layers of 100 nodes, 10% dropout per layer, a batch size of 25, and a learning rate of 0.005; achieving 69.7% accuracy, 53.1% sensitivity, and 78.9% specificity. When tested with different data stratifications, mean evaluation metrics dropped to 64.3±3.1% accuracy, 47.5±12.6% sensitivity, and 73.7±6.9% specificity, indicating that data stratification can impact model success when training ANNs. Although ANN results in this research were comparable to previous studies using ANN, classification results were lower than existing clinical assessment tools, indicating that this approach may only be useful in combination with existing tools, and not as an independent assessment.

4.2 Introduction

According to the World Health Organization, falling is the second leading cause of accidental injury death [1]. People with limb amputations are at greater fall risk than the general population [12]. Falling can lead to irreversible health, social, and psychological consequences [107]. Preventative programs can reduce the chance of falling, but the initial task of fall-risk identification can be challenging. Clinicians have access to many fall risk assessment tools for this purpose [39]. However, evolving wearable sensors systems could augment common clinical mobility tests for fall risk identification.

Fall risk classification research using wearable sensors, such as inertial sensors, has reported accurate and automated fall risk classification, enabling timely intervention with fall-risk mitigation techniques [7]–[10], [15], [16]. Previous research has extracted features from data

collected during clinical tests such as the 6-minute walk test (6MWT) [7] or 10 meter walk test [8]. The 6MWT measures functional capacity, and is widely used to measure the response to therapeutic interventions [23]. Participants walk straightaways and turns during a 6MWT, which is beneficial for automated classification because turning strategies can differ between fall-risk and no fall-risk populations. Elderly fall-risk classification has been achieved using turn data [7], [16]. Classifying fall risk using a 6MWT could make clinic time more efficient by reducing the number of functional tests required in a session.

Previous studies have used sensors at multiple locations on the body so that data from multiple body segments can assist with fall-risk classification [7], [8]. Unfortunately, multiple sensors can increase setup time, fall off or slide on the body, alter patient movement, and reduce position accuracy when too many sensors are worn. To facilitate use in clinic and at the point of patient contact, the number of sensors should be minimized, and sensors should have easy and reliable placement on the body. Although fewer sensors and less data could be detrimental to deep learning, Howcroft et al. reported the pelvis as the best single sensor location for fall-risk classification, specifically using a fully connected feedforward artificial neural network (ANN) [8], and the pelvis is the most commonly used sensor location for fall risk assessment [11]. Data from an accelerometer and gyroscope at the pelvis during a 6MWT could provide effective features for fall-risk classification.

Deep learning techniques such as ANN have the advantage of being able to learn complex, non-linear relationships [108], which is important because gait varies within and between individuals, especially in populations at risk of falling. These complex gait differences can lead to model generalization difficulties, especially between different at-risk populations. ANN are able to generalize a population, by inferring unseen relationships between samples [108].

Fall-risk classification using techniques such as ANN have focused primarily on elderly populations, with no predictive fall risk models for amputees. However, amputees are at higher risk of falling compared to able-bodied [12] and the risk of fall related injuries requiring medical care can be higher than older adults [14]. Howcroft et al. found that ANN were able to outperform the predictive ability of clinical assessment-based models when classifying fall risk in older adults using accelerometer data [8]. Since predictive models developed for able-bodied or older adult populations do not necessarily translate to the lower extremity amputee population, research is

needed to determine the best features and models for fall-risk classification in lower limb amputees.

The goal of this research was to implement deep learning techniques to predict fall-risk in lower limb amputees, from smartphone data collected during the 6MWT. ANN models were used for this classification task and compared with existing studies that performed similar research with able-bodied older adults. A successful ANN model would have similar, or better, results than existing clinical assessment tools.

4.3 Methods

4.3.1 Participants

A convenience sample of 129 participants with lower limb amputations were recruited from the University Rehabilitation Institute (Ljubljana, Slovenia). Clinical records provided self-reported number of falls, with falling at least once in the past six months prior to testing considered fall risk. Data from 89 participants (19 females, 70 males, 62.3 ±12.5 years old) were suitable for fall risk classification (32 fall-risk, 57 no fall-risk), with varying levels of amputations (4 bilateral transtibial amputees, 1 bilateral transtibial and transfemoral, 63 transtibial, 18 transfemoral, 2 knee exarticulation, 1 ankle exarticulation). Reasons for unsuitable data were malfunction exporting data from phone (5 people), no fall incidence data on file (9 people), running instead of walking during the 6MWT (1 person), or had unidentifiable foot strikes due to highly irregular gait (2 people), a single crutch (2 people), double crutches (18 people), or non-rolling walker (3 people).

4.3.2 Equipment

Participants had an Android smartphone affixed to the center of the posterior pelvis using a waist belt (Figure 14). Participant demographics were input to a custom designed The Ottawa Hospital Rehabilitation Center (TOHRC) Walk Test app [53] (Figure 15). Each participant performed a 6MWT along a 20m hallway (i.e., walk, turn around a cone, and continue the circuit for 6 minutes). The TOHRC Walk Test app collected smartphone 3D accelerometer and gyroscope raw data; and pelvic rotation, tilt, and obliquity at 50 Hz. Each trial was video recorded using a second smartphone. Once the test was complete, data were exported from the smartphone to a text file for post-processing.

Figure 14 Participant performing a 6MWT with the smartphone on their posterior pelvis

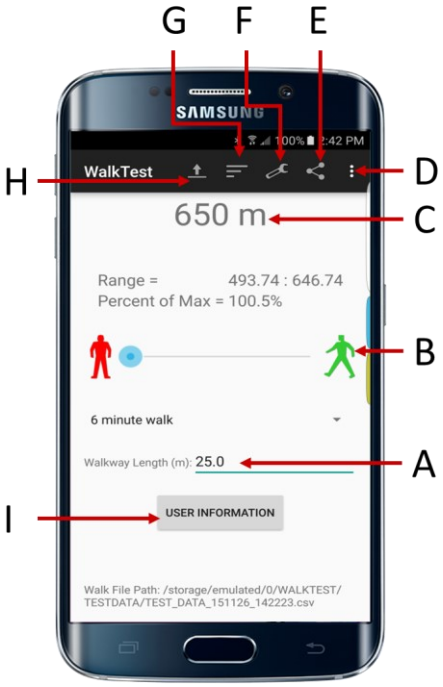

Figure 15 TOHRC Walk Test application, (A) Walkway length; (B) Start trial; (C) Distance walked; (D) About, Help; (E) Share output; (F) Settings; (G) Results tables; (H) Load previous data; (I) Enter patient demographics.

4.3.3 **Step and Turn Segmentation**

Raw accelerometer, gyroscope, and smartphone orientation data were imported to MATLAB 2013b, along with the time stamps for each recording. Foot strikes were identified from anterior-posterior (AP) linear acceleration, using the peak value near the estimated next step, where the estimated next step was based on average step duration. AP acceleration was chosen for this task because it had the least variance when compared to other linear acceleration axes. However, this automated technique (initially developed for able-bodied gait [53]) sometimes selected incorrect peaks with amputees due to amputees more asymmetric and variable gait. In these cases, manual step identification was required. Features were extracted from the manually cleaned data.

Since data from turn walking is better at classifying fall risk than data from straight walking [7], data were segmented into turns and straightaways. Turns were defined as the five steps around the middle of each turn. The center of a turn was identified using pelvis rotation (change of angle in the transverse plane as measured by the smartphone azimuth [53]), the middle frame between the beginning and end of a turn (i.e., when the azimuth started and stopped rotating). This process was first automated in MATLAB, then verified manually.

4.3.4 **Feature Extraction**

Once turn steps were identified, a set was created with 62 features from each step [7], [11]. These features were:

Temporal: cadence, step time (foot strike to foot strike of the opposite foot), stride time (foot strike to foot strike of the same foot), symmetry in right and left limb step times (symmetry index) [101].

Descriptive statistics: minimum, maximum, mean, standard deviation, root mean square in three axes (vertical, medial lateral (ML), AP) for pelvis linear acceleration (Android processed signal, not including gravity); and tilt, rotation, and obliquity angular velocities.

Linear acceleration and angular velocity frequency domain features: from the absolute value of the Fast Fourier transform (FFT) of each step, the first quartile of Fourier transform (FQFFT), ratio of even/odd harmonics (REOH), and peak distinction.

FQFFT: percentage of frequencies within the first quartile of the Nyquist frequency (6.25 Hz is the first quartile based on a Nyquist frequency of 25Hz). Lower FQFFT values indicate more high frequency components, linked to instability [56].

REOH: ratio of the frequencies in the even harmonics compared to the odd harmonics, using stride time as the fundamental frequency. Lower REOH values have been associated with fall risk [7], [57].

Peak distinction: determines if the FFT peak frequency is distinct from other frequencies, the percent of frequencies in the FFT with power greater than a threshold (⅓ amplitude of peak signal) was calculated. A lower peak distinction value means a more distinct peak (Figure 16).

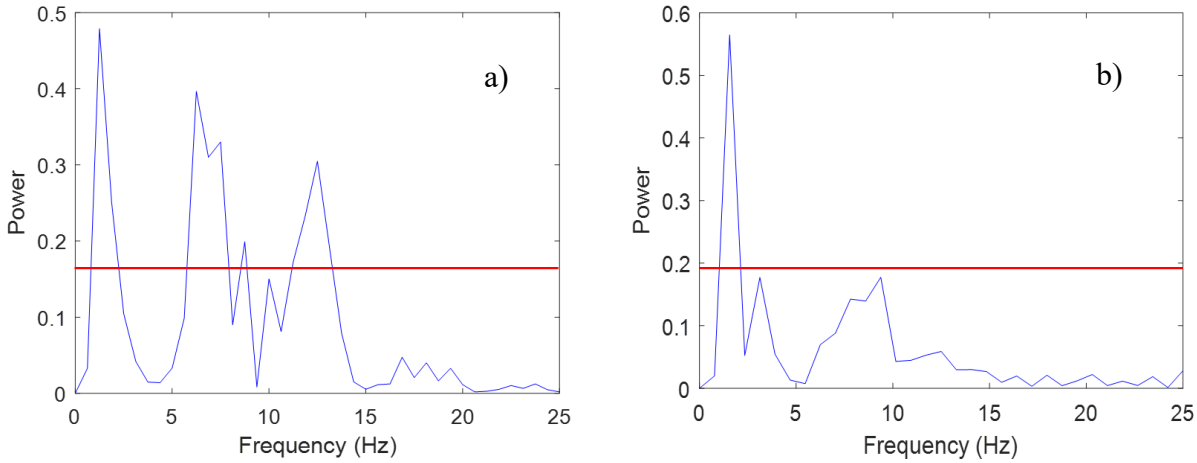

Figure 16 Peak distinction for two steps. (a) A step with multiple peaks, resulting in a greater percentage of frequencies above the threshold. (b) A step with a single peak over the threshold, resulting in fewer frequencies above the threshold. Horizontal red lines indicate one third of max amplitude.

Once features were extracted for each step, the minimum, maximum, mean, and standard deviation were calculated over all turn steps for a total of 248 features (62 features multiplied by 4 statistics).

4.3.5 Feature Selection

Feature selection was used to reduce feature space dimensionality, by removing redundant and irrelevant data [7], [8], [102]. Three feature selection techniques were used, based on previous success in fall risk classification: an extra trees classifier ensemble method (ETC) [7], [69], Relief-F (RelF) [60], [103], and Correlation-based feature selection (CFS) [59], [64].

ETC is an ensemble method that fits randomized decision trees on various dataset sub-samples and uses averaging to improve the predictive accuracy and control over-fitting [69]. Features are then ordered based on their importance so that features subsets of desired sizes can be selected from the best ranked features. ETC uses randomly selected weightings when building trees, making ETC better for feature selection than random forests [72].

RelF is a filter method that ranks features by weighting them based on their relevance and how well instances from different classes and the same class can be distinguished [60], [65]. RelF does not eliminate redundant features, making this method most useful when evaluating parameters with interdependencies.

CFS filter-based method that identifies a subset of features correlated with the class label (i.e., fall-risk or no fall-risk), but also uncorrelated to other parameters. CFS calculates feature subset "merit" based on pair-wise correlations by adding one feature at a time, and only keeping features which improve the merit [65], [66]. This allows CFS to develop a subset without irrelevant or redundant features.

For RelF and ETC, all features were ranked, and the user selected the desired subset size. Feature subsets were created for the top 100, 50, 30, 20, 10, and 5 features for RelF and ETC. CFS selects features based on their merit, and only includes features that improve subset merit. Only 3 features were selected by the CFS feature selector, so only one CFS feature subset was used.

4.3.6 **Classification Techniques**

Neural network models were developed using 13 feature sets (6 RelF, 6 ETC, 1 CFS) and with all features. Models were evaluated using five metrics: accuracy, sensitivity, specificity, Matthews correlation coefficient, and F1 score. A ranking technique similar to [8] and [105] was used to determine the best models. Each model was ranked in the five evaluation metrics, and the models with the lowest summed rankings were chosen as the top classifiers. ANN models were built using the Keras library [109] in Python, and preprocessing was done using Scikit Learn [110].

The following tasks were completed to select the best feature subset and optimize the networks. Unless stated otherwise, learning rate=0.001, optimizer is RMSProp, batch size is 25, and activation for the hidden layers is relu.

4.3.6.1 **Task 1: Determine the Best Feature Set**

To determine the best feature set, ANN models for each feature set were built with one hidden layer and a batch size of 25. Hidden layers were activated using a relu activation, with the RMSProp optimizer and learning rate of 0.001. Each feature set was tested with 5, 10, 15, 25, 50, 100, 250, and 500 nodes, with 500 epochs (total of 112 models). A stratified train-test split (66% train and 33% test) was used to generate evaluation metrics. The best result, using the ranking method, was used to determine the number of nodes for each of the 14 feature sets.

Since accuracy for some smaller feature sets did not plateau at 500 epochs, the 14 feature sets were rerun with 1000 epochs and the determined number of nodes. The highest ranked feature set was ETC50 with 400 epochs (416 rounded down to 400 epochs), and was used for further model development.

4.3.6.2 **Task 2: Optimize Model**

Three phases were performed to determine the optimal hyperparameters for the ANN model, with the ETC50 feature set and 400 epochs. 5-fold cross validation was implemented for all models [95].

1. **Phase 1**: Eighty-four models were built with 3, 5, 10, 50, 100, 250 and 500 nodes; 1, 2, and 3 layers; and 0%, 10%, 25%, and 50% dropout. The best model was determined using the summed ranking method.
2. **Phase 2**: The best model from phase 1 was built with 0.005, 0.002, 0.001, 0.0005, 0.00025, and 0.0001 learning rates; and 25, 15, and 5 sample batch sizes (18 models).
3. **Phase 3**: The best model from phase 2 was built using five different random seeds, that produced different stratifications. This tested the model's response to different data segmentations for each of the 5 folds.

4.4 Results

4.4.1 Task 1: Determine the Best Feature Set

As shown in Table 7, the ETC50 model had the best summed ranking, with 80% accuracy, 73% sensitivity, and 84% specificity. The greatest accuracy was achieved using between 238 and 594 epochs (mean = 416 epochs). 500 nodes were used in the top ranked model.

Table 7 Results and summed rankings for each feature set in Task 1. MCC= Matthew's Correlation Coefficient, F1=F1 Score, SR=summed ranking (lower scores indicated better rank)

Feature Subset	Nodes	Accuracy	Specificity	Sensitivity	MCC	F1	SR
ETC50	500	80.0	84.2	72.7	0.569	0.780	15
RelF100	50	73.3	89.5	45.5	0.398	0.563	22.5
RelF30	250	73.3	89.5	45.5	0.398	0.563	22.5
ETC20	250	73.3	89.5	45.5	0.398	0.563	22.5
ETC5	100	73.3	94.7	36.4	0.402	0.483	23.5
RelF5	100	73.3	100.0	27.3	0.438	0.393	28.5
None	100	70.0	89.5	36.4	0.311	0.481	38.5
ETC30	15	70.0	89.5	36.4	0.311	0.481	38.5
ETC10	100	70.0	100.0	18.2	0.351	0.287	42.5
ETC100	100	63.3	73.7	45.5	0.196	0.551	46
RelF20	5	66.7	89.5	27.3	0.217	0.389	50.5
RelF10	50	66.7	89.5	27.3	0.217	0.389	50.5
RelF50	50	63.3	78.9	36.4	0.167	0.474	54
CFS	500	63.3	100.0	0.0	N/a	0.000	N/a

The ETC50 feature set is shown in Table 8.

Table 8 Features in the top feature set (ETC50), in ranked order according to ETC feature selection.

Rank	Definition	Statistic	Rank	Definition	Statistic
1	Tilt angular velocity mean	Min	26	Standard deviation of ML acceleration	Mean
2	REOH of vertical acceleration	Mean	27	Cadence	Std Dev.
3	Stride timing	Max	28	Maximum of ML acceleration FFT	Min

4	Maximum AP acceleration	Mean	29	Rotation standard deviation	Mean
5	Standard deviation of obliquity angular velocity FFT	Std Dev.	30	REOH of vertical acceleration	Max
6	Stride timing	Std Dev.	31	Rotation angular velocity mean	Min
7	Cadence	Mean	32	Obliquity angular velocity range	Std Dev.
8	FQFFT tilt angular velocity	Mean	33	REOH of rotation angular velocity	Max
9	Peak distinction of AP acceleration	Min	34	LR symmetry	Mean
10	Rotation angular velocity mean	Mean	35	Peak distinction of vertical acceleration	Std Dev.
11	Root mean squared of ML acceleration	Std Dev.	36	Standard deviation of vertical acceleration FFT	Max
12	Step Timing	Max	37	Peak distinction of tilt angular velocity	Max
13	Peak distinction of obliquity angular velocity	Min	38	REOH of obliquity angular velocity	Mean
14	Standard deviation of ML acceleration	Min	39	FQFFT obliquity angular velocity	Min
15	Standard deviation of tilt angular velocity	Min	40	Tilt angular velocity mean	Mean
16	REOH of rotation angular velocity	Min	41	Maximum of obliquity angular velocity FFT	Min
17	Peak distinction of tilt angular velocity	Min	42	Minimum AP acceleration	Min
18	Maximum of tilt angular velocity FFT	Std Dev.	43	Maximum ML acceleration	Std Dev.
19	Maximum of AP acceleration FFT	Mean	44	Standard deviation of AP acceleration FFT	Std Dev.
20	REOH of tilt angular velocity	Min	45	Standard deviation of AP acceleration FFT	Max
21	LR symmetry	Min	46	Peak distinction of rotation angular velocity	Mean
22	Standard deviation of AP acceleration	Max	47	Maximum of Vertical acceleration FFT	Min
23	Maximum of Vertical acceleration FFT	Max	48	Peak distinction of rotation angular velocity	Max
24	Standard deviation of obliquity angular velocity FFT	Mean	49	FQFFT tilt angular velocity	Std Dev.
25	FQFFT ML acceleration	Min	50	Peak distinction of vertical acceleration	Min

4.4.2 **Task 2: Optimize the Best Feature Set**

Using 5-fold cross-validation, the best ranked model in Phase 1 had two layers of 100 nodes and a 10% dropout per hidden layer. This model achieved 68.5% accuracy, 46.9% sensitivity, and 80.7% specificity (Table 9).

Table 9 Top ten models when optimizing for layers, nodes, and dropout (Phase 1). MCC= Matthew's Correlation Coefficient, F1=F1 Score, SR= Summed Ranking (lower scores indicated better rank). Standard deviation in brackets.

Nodes	Layers	Dropout (%)	Accuracy (%)	Specificity (%)	Sensitivity (%)	MCC	F1	SR
100	2	10	69 (11.6)	81.2 (15.9)	47.1 (12.4)	0.326 (0.237)	0.729 (0.073)	65.5
250	1	10	69.8 (9.2)	86.4 (12.5)	40 (22.9)	0.314 (0.215)	0.725 (0.074)	66.5
250	1	25	69.7 (5.7)	86.2 (9.6)	40 (22.9)	0.308 (0.157)	0.727 (0.064)	66.5
250	3	50	68.8 (10.1)	82.7 (12.1)	43.8 (6.9)	0.304 (0.232)	0.721 (0.053)	67.0
100	1	25	68.5 (4.4)	84.4 (6.7)	40 (17.7)	0.275 (0.125)	0.721 (0.058)	81.0
250	2	10	67.4 (12.6)	80.8 (17.3)	43.3 (10.4)	0.289 (0.257)	0.713 (0.071)	85.5
100	2	50	67.5 (3.4)	81.1 (10.5)	42.9 (17.6)	0.261 (0.092)	0.722 (0.034)	85.5
100	3	0	66.2 (4.6)	73.6 (6.5)	52.9 (12.4)	0.266 (0.114)	0.739 (0.042)	87.0
500	1	10	67.5 (3.4)	82.6 (5.7)	40.5 (7.1)	0.256 (0.073)	0.713 (0.024)	96.5
500	2	50	66.6 (16.9)	75.9 (23.4)	49.5 (21.8)	0.285 (0.358)	0.725 (0.096)	96.5

The Phase 1 model was then optimized for learning rate and batch size. The best summed-ranked model used a 0.005 learning rate with a batch size of 25, achieving 69.8±9.6% accuracy, 53.3±9.0% sensitivity, and 79.1±13.4% specificity across the 5 folds (Table 10).

Table 10 Top ten models optimized for learning rate and batch size (Phase 2), and ordered by summed ranking from lowest to highest. MCC= Matthew's Correlation Coefficient, F1=F1 Score, SR= Summed Ranking (lower scores indicated better rank). Standard deviation in brackets.

Learning Rate	Batch Size	Accuracy (%)	Specificity (%)	Sensitivity (%)	MCC	F1	SR
0.005	25	69.8 (9.6)	79.1 (13.4)	53.3 (9.0)	0.35 (0.206)	0.749 (0.054)	17.0
0.001	25	69.0 (11.6)	81.2 (15.9)	47.1 (12.4)	0.326 (0.237)	0.729 (0.073)	20.5
0.001	5	68.9 (13.5)	81.1 (13.4)	47.1 (20.8)	0.297 (0.304)	0.732 (0.097)	20.5
0.001	15	68.4 (10.6)	73.6 (6.0)	59 (25.5)	0.316 (0.275)	0.773 (0.106)	22.5
0.0005	15	67.2 (9.2)	79.1 (11.7)	45.7 (21.3)	0.264 (0.217)	0.728 (0.075)	31.0
0.0001	15	66.2 (4.2)	82.6 (5.7)	36.7 (14.8)	0.212 (0.120)	0.703 (0.039)	37.5
0.002	25	66.8 (13.5)	79.5 (21.4)	43.8 (20.7)	0.261 (0.259)	0.712 (0.086)	38.0
0.002	5	64.2 (11.2)	68.3 (10.5)	57.1 (17.6)	0.249 (0.234)	0.741 (0.100)	38.0
0.002	15	65.8 (15.5)	79.7 (23.1)	40.5 (13.8)	0.247 (0.304)	0.694 (0.092)	45.0
0.0005	25	63.8 (8.9)	79.2 (17.0)	35.7 (26.6)	0.164 (0.229)	0.698 (0.066)	52.0

Phase 3 used different stratifications for each of five, 5-fold validations, and the best model from Phase 2 were retested. Mean and standard deviations from all 5-fold results were 64.3±3.1% accuracy, 47.5±12.6% sensitivity, and 73.7±6.9% specificity. These average results were approximately 5% lower than optimized model metrics in Phase 2.

4.5 Discussion

The best neural network model for fall risk prediction from smartphone sensor data during a 6MWT used the ETC50 data set in Task 1. Task 2 optimization used 5-fold cross-evaluation instead of a train-test split to better estimate classifier accuracy on unseen data, which decreased evaluation metrics. The best model after optimization achieved 69.7% accuracy, 53.1% sensitivity, and 78.9% specificity when using 5-fold cross validation, which is similar to results from some clinical assessment tools. Interestingly, when the same parameters were used with different sample stratifications, the mean evaluation metrics were approximately 5% lower, which reduces clinical viability of this neural network-based model.

The ETC feature sets ranked better than the RelF feature sets when using 5, 10, 20 and 50 features, but ETC ranked lower when using 30 and 100 features. ETC could have been better when using fewer features because RelF feature selection is a univariate feature selector and selected

features can be redundant. ETC selected a variety of features with no clear similarities, using a wide range of variables. The best-ranked feature using ETC was the minimum value for mean tilt angular velocity across all turn steps. This feature was not available to previous studies that only used accelerometers. Even with only one sensor location, including both gyroscope and accelerometer features could improve classification.

4.5.1 **Model Optimization**

The ETC50 data set produced good results in Task 1, using 500 nodes. Further optimization in Task 2 determined that 100 was the best number of nodes with two layers. This likely means that the ETC 50 feature set benefits from a complex architecture to achieve the best results, whether through added layers or many nodes. Using more layers and more nodes increased model capacity and allowed the model to better learn the training set.

Using 100 nodes and two layers was found to be optimal. Additionally, including dropout improved the evaluation metrics. Table 9 demonstrates that the top eight models had dropout. Including dropout when training ANN helps to reduce overfitting by increasing the strength of each individual node. Because dropout improved the results, models without dropout may have overfitted. The top two models both used 10% dropout, which indicates that smaller dropout was preferred (larger percent dropout excludes more nodes when training each epoch).

Increasing learning rate to 0.005, from 0.001, improved sensitivity but slightly reduced specificity. The increased learning rate was ranked slightly higher with the summed ranking method, and reducing the learning rate tended to decrease the model rankings. Models with the lower learning rates may have been converging on local minima of loss. A greater learning rate can allow models to converge to a global minimum and therefore perform better. However, the improvement was minimal when optimizing learning rate or batch size.

4.5.2 **Model Evaluation**

The ANN with the optimized ETC50 feature set from Task 2 had lower sensitivity (53.1%) and higher specificity (78.9%) than existing clinical tests that can assess fall risk [39]. The Timed-up and Go (TUG) test has a predictive sensitivity of 76% and specificity of 49% on older adults [39]. One study testing multiple fall rate (2 or more falls) using TUG with amputees had better

prediction results (85% sensitivity, 74% specificity) than with older adults (N=40) [37]. While this study requires replication to verify the evaluation metrics, the higher sensitivity and specificity could also have been due to the 2 or more falls criteria for fallers, since this group may have had consistently poorer TUG performance than people who have only fallen once.

A study using ANN for sensor-based fall-risk detection [8] reported the pelvis accelerometer as the best single sensor location during dual task walking along a 10m straight path; achieving 57% accuracy, 43% sensitivity, and 67% specificity. The results in our study, with the 6MWT and both accelerometer and gyroscope features, were better than [8]. The highest ranked feature using ETC was based on angular velocity, supporting the benefit of using a wearable device with multiple inertial sensors. However, future work with neural networks should consider using a larger data set or collecting data from more locations on the body since more sensors have been shown to improve results in [8]. Machine learning techniques that perform better with less data could also improve results, such as the random forests used by Drover et al. [7].

Drover et al. [7] analyzed turns during a 6MWT using a random forest classifier with accelerometer data collected from the pelvis and shank to achieve better results with older adults (77.3% accuracy, 66.1% sensitivity, 84.7% specificity). The 6MWT includes a longer walking period and multiple turns, which may be a better task for fall risk classification than 10m straight walking. The research by Drover et al. [7] had better results than the ANN model in our study, and the study by Howcroft et al. This may indicate that, although the classification metrics in our study were comparable to clinical assessment tools, ANN is not the best fall risk classifier for lower limb amputees when using 6MWT data.

Model specificity was greater than both sensor-based and clinical approaches in the literature. Therefore, fewer people would be misclassified as fallers with the ANN model. This is important because misclassifying people as fall risk individuals may lead to people being inappropriately entered into fall mitigation programs or treatments, therefore unnecessarily using limited resources and clinic time.

An interesting result from our study is that averaging five different 5-fold stratifications (Phase 3), produced approximately 5% lower evaluation metrics that the 5-fold validation in Phase 2. It is not currently common practice in sensor-based fall risk classification to re-run cross-fold

validations with multiple stratifications to determine the most realistic outcome with unseen data. In future research, retesting the best model with multiple newly collected data sets can help verify if this "multiple stratification" test is beneficial and is recommended as best practice when evaluating fall risk classification models.

Although the best ANN in this study was comparable to previous studies using ANN, this approach may not be the best solution for fall risk classification using inertial sensors and walking tasks for people with lower limb amputation. Other machine learning techniques should be pursued. The model developed in this study did not improve on existing clinical techniques. Although the model in this study is likely insufficient for clinical use, it could be used in addition to existing assessment tools. However, the models in this study did provide better evaluation metrics than similar ANN approaches in the literature.

4.6 Conclusions

A novel smartphone sensor-based classification method using ANN was developed to identify fall-risk in lower limb amputees. Classification was performed using turn data from a 6MWT. When evaluated with a 5 fold cross-validation, it was found that the best model used 50 features from an Extra Trees Classifier (ETC), 2 layers of 100 nodes, 10% dropout per layer, a batch size of 25, and a learning rate of 0.005 to achieve 69.7% accuracy, 53.1% sensitivity, and 78.9% specificity. When tested with different data stratifications, mean evaluation metrics were approximately 5% lower. Methods tested in this research could be implemented in smartphone applications to allow for fall-risk classification when using a 6MWT but should be used in addition to existing clinical techniques since results do not support use for independent classification. Future work should consider other machine learning techniques and add more participants to the data set, with the ultimate goal of automated smartphone classification that is better than or similarly effective to existing clinical assessment tools.

5 Thesis Conclusions and Future Work

Throughout this thesis, novel machine learning models were designed, developed, and evaluated for their ability to classify fall-risk in people with lower limb amputations. Features were extracted from smartphone inertial sensor data collected during a 6MWT and ranked using three different feature selection techniques. It was found that the random forest model had superior results compared to the ANN, and was effective with less optimization than the ANN. The final random forest model was most successful when using data collected from turning, and features selected using a CFS feature selection technique.

The models in this thesis used data collected from a single sensor location during a 6MWT. Model evaluation metrics, specifically sensitivity and specificity, were similarly effective when compared to existing clinical assessment tools. When compared to previous research classifying fall-risk in older adults with machine learning, the random forest on data collected from amputees was similarly effective or better. These techniques could help augment the traditional 6MWT, expanding its clinical use, and reducing clinic time for patients by reducing the number of required functional assessments per visit. The method proposed in this thesis can create opportunity for facilities that already use the 6MWT as an outcome measure to automatically assess fall-risk in their patients, and potentially catch additional FR individuals with little to no extra effort. Knowing that the specificity is high and that misclassifying NFR individuals is very unlikely, utilization of this random forest classifier could identify 57% of FR individuals without the need for any additional testing, which is much better than missing FR individuals who complete a typical 6MWT.

Due to the limited size of the data set, no test sets were held back as unseen data for final evaluation. Furthermore, a single small test set could introduce sampling error, where the results are over or underestimated, which is why cross-fold validation was used. Although the random forest and ANN were compared directly, cross-fold validation used to test the random forest was a leave one out strategy and the ANN used 5-fold cross validation. Ideally, the same cross-validation would have been used for both modelling techniques.

The conclusions for each thesis objective and the corresponding hypotheses are presented below:

5.1 Objective 1: Create and evaluate a viable machine learning model for predicting fall risk in people with lower extremity amputation using smartphone sensor data collected during a 6MWT.

For objective 1, conclusions were made independently for random forests and neural networks developed in this thesis. Comparisons between the two techniques are made in objective 2.

5.1.1 Hypothesis 1: Sensitivity and specificity for fall risk classification will be equivalent or better than metrics from the TUG and FSST clinical assessment tools for identifying fall risk in amputees and older adults.

5.1.1.1 Random Forests

The best classification model developed in this thesis was a random forest that used 500 trees and achieved a mean 57.2% sensitivity and 94.9% specificity across five random seeds. The features used in the best model were chosen using a CFS, which chose only 3 features. The best model had a very high specificity that was greater than the specificity achieved using TUG (74% specificity) or FSST (93% specificity) to identify amputees who have fallen multiple times (2 or more), and far greater than the TUG when used on older adults (49% specificity). Higher specificity means fewer NFR individuals will be misclassified than with existing clinical assessment tools, and resources will not be mistakenly allocated to those who do not need them.

The random forests sensitivity was less than the specificity and was not greater than the sensitivities achieved by the TUG or FSST on either amputees or older adults. However, a sensitivity of 57% is likely sufficient for clinical use because it allows for confident classification of those who are classified as FR individuals due to the model's very high specificity. Although not all FR individuals will be identified using this model, clinicians who feel a FR individual was misclassified as NFR can continue with existing clinical assessment tools. This means that the random forest model developed in this thesis is a viable option for clinical use that could identify FR individuals who would not necessarily be considered for a fall prevention intervention, as long as existing clinical assessment tools are used in addition to the model as needed.

5.1.1.2 Neural Networks

The best neural network model developed in this thesis achieved a 78.9% specificity and 53.1% sensitivity when using an ETC50 feature set, and after thorough investigation of many hyperparameters. Although specificity was moderate, sensitivity was quite low. When compared to research using the TUG to identify multiple fall amputees, the specificity is slightly greater, but the sensitivity is far lower. The FSST has much better sensitivity and specificity when identifying multiple falls in amputees when compared to the ANN model developed in this thesis. This is especially true if the decreased evaluation metrics with different data stratifications are considered. It is unlikely that the ANN model developed in this thesis would be sufficient for clinical use, unless used in parallel with existing clinical assessment tools.

5.1.2 Hypothesis 2: Sensitivity and specificity for fall risk classification will be equivalent or better than metrics from studies with an older adult cohort that used wearable inertial sensors.

5.1.2.1 Random Forests

The best random forest model developed in this thesis used a similar data collection technique to a previous study by Drover et al. [7], which was to collect walking data during a 6MWT using accelerometers on the lower back and shanks. Even though the pelvis was the only sensor location used in this thesis, the results were similar or better. The best model achieved by Drover et al. had a 73.4% accuracy, 60.5% sensitivity, 82.0% specificity, and 0.44 MCC. The random forest model from this thesis achieved better results in each of these metrics except for sensitivity, where it was only 3.3% less. Having achieved a better accuracy, specificity, and MCC, indicates that using inertial sensor data with random forests is as effective with amputees as it is with older adults. Additionally, Since Drover et al. had two additional sensor locations on the left and right shank, but no gyroscope, it can be seen that the data from a single smartphone's sensors at the lower back is similarly effective for fall-risk identification as using three sensor locations. Additionally, a single smartphone is far more efficient and simpler to apply in clinics.

Another interesting commonality between this thesis and the research by Drover et al. is that the best results were achieved by using data from turns. This shows that both older adult and

amputee populations have better fall-risk classification when using data collected from turns.

5.1.2.2 Neural Networks

ANN developed by Howcroft et al. [8] achieved a mean 57% accuracy, 43% sensitivity, and 65% specificity when using multiple sensors for the 25 foot walk test. Their results were slightly worse when using a single sensor on the pelvis, achieving a mean 54% accuracy, 35% sensitivity, and 67% specificity. The best ANN developed in this thesis had better results (69.7% accuracy, 53.1% sensitivity, and 78.9% specificity) than Howcroft et al., especially when compared to the single sensor location results. Mean evaluation metrics from this thesis for multiple data stratifications were approximately 5% lower, but this result was still greater than the results achieved by Howcroft et al.

Because results were slightly better with this research than in previous work with ANN, the 6MWT task might be an improvement on other data collection tests such as 10-meter walk tests or dual task straight walking. Possible reason is that a 6MWT includes turn data, which has been found as important for fall-risk classification in both previous work on older adults and this thesis, and the longer walking duration that could include mild fatigue effects on gait.

5.2 Objective 2: Compare random forests and ANN to determine which technique provides better results for fall-risk classification in lower limb amputees, and which feature selection techniques are most effective.

5.2.1 Hypothesis: Random forests will have greater sensitivity and specificity than ANN when classifying fall-risk in amputees, similar to previous research identifying fall risk in older adults.

The best fall-risk classification results from Drover et al. used a random forest, and provided better results than the ANN used by Howcroft et al. [7], [8]. This thesis found similar results, with the random forest performing much better than the ANN despite using the same features and feature selection techniques. The random forest achieved a mean 57.2% sensitivity and 94.9% specificity, while the ANN achieved a peak 53.1% sensitivity and 78.9% specificity, which decreased by around 5% when data stratifications changed. From these results, random forest achieved 4.1% better sensitivity and 16.0% better specificity. Additionally, the random forest results were

achieved using leave-one-out cross fold validation, while the ANN was built using 5-fold cross validation due to long training times to build models. When considering the more thorough cross-fold evaluation used for the random forest, and the noticeable decrease in ANN results after building the model with multiple data stratifications, the random forest results are more likely to reflect how the model will behave on future data.

An advantage of the random forest over an ANN is a shorter training time, facilitating a leave-one-out strategy that improves model validity. In application, this also makes the random forest more efficient to train and therefore easier to rebuild with new data or different feature subsets. Another aspect of random forests that make them more efficient is that fewer hyperparameters are needed to optimize the model. With ANN, hyperparameter optimization takes up large amounts of time, and does not always improve model results, as was seen in this thesis with learning rate and batch size. With the random forest, the only optimization needed to achieve acceptable results was increasing the number of trees.

5.2.2 Hypothesis: Both random forests and ANN will have better evaluation metrics when using subsets developed from feature selection strategies rather than all features.

Both the random forest and ANN models improved with feature selection, although the impact of feature selection on random forests was far greater. Using no feature selection with a random forest produced a model that was at least 17% worse in all evaluation metrics. However, with ANN, using all features in initial train-test validation models resulted in the third highest ranked model.

One benefit of filter-based feature selection is that the feature sets remained the same whether using a random forest or neural network, making comparisons easier. All models in this thesis were tested with feature subsets selected by RelF, ETC, and CFS feature selectors. The only difference was for ANN, where 50 and 100 features were kept for the ETC and RelF feature subsets because ANN performed better with more data. This was seen when comparing the best feature subset for each of the two models, where random forest worked better with the smallest feature subset and ANN models performed better using larger feature subsets or all features.

Another interesting comparison between the random forest and ANN was that the CFS feature selector was the best feature selector for the random forest model, but was unable to converge for the ANN (i.e., worst feature selector for the ANN). Similarly, the ETC50 feature set was the best for the ANN models, but the best ETC model for the random forest had 15.6% less sensitivity and 10% less specificity than the CFS model.

The CFS feature subset provided a very small feature subset (only 3 features), which is often important for reducing overfitting and redundancy in features. Although reducing features is important, ANN may have benefited from having more features to choose from, since ANN inherently emphasize some features over others during training. Having too few features greatly reduced ANN model capacity but improved random forests capacity.

The three features in the CFS subset were not used by Drover et al. [7], because most of the features selected by Drover were shank related. However, similar to Drover, about half were frequency domain features. Two of the three CFS features, five of the top ten ETC features, and three of the top ten RelF features were from the frequency domain. This demonstrates the importance of frequency domain features in fall-risk assessment.

5.3 Future Work

This research presented machine learning techniques, including feature extraction, feature selection, model optimization, and model evaluation for fall-risk identification in people with lower limb amputations. Existing clinical assessment tools could benefit from the techniques presented in this thesis to achieve fall-risk classification and save on clinic time by performing fall-risk classification during a 6MWT. Because data was collected using a smartphone Walk Test application, which can be downloaded and used by any Android smartphone, this research has direct and feasible use in almost any clinic. The models built in this thesis have confirmed that existing machine learning techniques that were feasible with older adults are applicable to amputee populations, and also achievable using one sensor location.

5.3.1 Sensor and Smartphone Development

As previously mentioned, future data collection can be easily achieved using accessible smartphone technologies. With current smartphone computing power and software development,

feature extraction and machine learning can be done directly on the phone, or data can be quickly exported to the cloud. Additionally, smartphone manufacturers continue to add and improve on sensors available in the phone. While this thesis used inertial measurements, future smartphones could use camera sensors to apply computer vision technology and refine existing inertial sensor data through indoor mapping [111]. With newer inertial sensors and the addition of other smartphone data, more precise measurements could be made, data would be less noisy, and machine learning would have additional training data to achieve better results.

5.3.2 Feature Selection Techniques

Feature selection was effective for improving fall-risk classification, specifically when using random forests. Although a variety of feature selectors were tried, future work should investigate other types of feature selection, specifically those proven effective by similar studies such as select-*k*-best using ANOVA statistics [7]. While the best feature set from this thesis achieved results similar or better than existing studies and clinical assessment tools, other feature selectors could further improve results, especially with more participants in the data set.

Other techniques for reducing feature set complexity should also be considered, such as wrapper techniques or feature extraction. For this thesis, wrappers were not chosen because ANN wrapper are uncommon and furthermore make comparison between random forests and ANN difficult. Given that the random forest was the more effective method of classification, future work could investigate wrapper methods or additional embedded methods such as the ETC feature selector that use decision trees. The ETC feature selection was chosen for this thesis rather than a wrapper method for reduced computational times and easier comparison between classifier types, but future work could consider both approaches.

Feature extraction, a technique that projects the original features onto a new, lower dimensionality feature space, was excluded from this thesis. Feature extraction causes features to lose their real-world meaning, making it difficult or impossible to know which features are important to classification. Understanding which features are being used for clinical applications such as fall-risk identification is important since clinicians may benefit by knowing which data are relevant, which in turn encourages them to trust the results and improve outcomes in the future. However, future work in fall-risk classification could use feature extraction rather than feature

selection if better results can be achieved.

5.3.3 Optimize Additional Random Forest Hyperparameters

In this thesis, the only hyperparameter adjusted for the random forest was the number of trees. Future optimization on random forests could consider additional hyperparameters such as maximum tree depth and minimum samples in a split. These parameters tend to reduce overfitting and should be considered when more data is available for training.

5.3.4 Classification Techniques

Although previous research showed that decision trees and forests were the best models for fall-risk classification, other models may produce better results. In the preliminary work for this thesis, brief tests were performed in WEKA with default parameters for SVM, KNN, Naïve Bayes, and J48 classifiers to compare with random forests. While random forest with default parameters was found to be the best classifier, these classifiers were not optimized, and others were not investigated. Future work should investigate machine learning techniques besides random forests and ANN in more depth determine their performance on amputee data.

One type of machine learning that could be a benefit to fall-risk identification is recurrent neural networks (RNN). Although ANN were not successful in this thesis, fall-risk identification may improve with RNN that can learn changes in features over time. Features could also be retrieved from specific gait phases that could be identified through RNN models, and classification could be done while considering temporal aspects of gait. Models that can continuously learn from new input data should also be considered, since they could allow for model improvement within clinics, without requiring updates to any software for implementing new data.

5.3.5 Continuous Monitoring and Fall Detection in the Phone App

One benefit of using a smartphone application for fall-risk detection is that smartphones can be used outside clinics. From this research, a 6MWT was beneficial for fall-risk identification, and turning features were better than straight walking features. Continuous monitoring would provide more turning data and could provide data from activities of daily living that are more indicative of fall risk than turning. Existing devices that use inertial measurements, such as medical

alert bracelets, are already used to detect falls in everyday life. This has led to research on smartphone fall detection to replace existing devices [112], [113], and even applications that combine fall detection with activity recognition [114]. Using similar technology, smartphones could be used for both fall-risk detection and, if people are deemed a FR, used to detect falls. These techniques would further benefit from machine learning models that constantly learn, by classifying someone based on whether or not they fall while using the smartphone, and then using their data to further train fall-risk identification models.

5.3.6 **Participant Diversity and Separation**

One of the advantages of the data from this thesis was the diversity in the amputee population. Generally, research involving amputees is limited to one type of amputation, since people with different amputation levels walk differently and have different functional capacities. The most common type of amputation researched is transtibial because it is the most common type of lower limb amputation (excluding toe amputation which is slightly higher) [115]. Due to the differences in gait, building separate models for different levels and types of amputation could improve classification ability.

For this thesis, all levels of amputation were included in training the models for two reasons: generalizability to the entire amputee population would be ideal, to not require multiple models, and only 89 participants available in the dataset (63 of which were transtibial). Future research with larger data sets should be considered to develop models specifically for different types of amputations, to determine if classification results can be improved.

5.3.7 **Weighted Sum Ranking**

A summed ranking technique was used in this thesis by assuming that each evaluation metric is equally important to the end user. However, this is often not the case. Some users would prefer a model that has a higher sensitivity but lower specificity. Additionally, as was mentioned in this thesis, some metrics give a better overview of performance, such as MCC. Future work could consider adding metric weighting to the model ranking method, based on end user preferences.

5.3.8 **Clinical Implementation**

Clinical application of the research presented in this thesis could help mitigate the risk of falls that lead to debilitating injury for amputees. Even without some of the future work suggested in this chapter, implementing the best model into a 6MWT smartphone application could improve the lives of amputees and make clinic time more efficient for both patients and clinicians, by automatically assessing fall risk upon completion of the 6MWT. Although research is still required to validate the machine learning techniques proposed, random forest classifiers could be immediately implemented into smartphone applications and used to augment existing clinical assessment tools.

References

[1] "Falls." https://www.who.int/news-room/fact-sheets/detail/falls (accessed Jan. 14, 2020).

[2] F. El-Khoury, B. Cassou, M.-A. Charles, and P. Dargent-Molina, "The effect of fall prevention exercise programmes on fall induced injuries in community dwelling older adults: systematic review and meta-analysis of randomised controlled trials," BMJ, vol. 347, Oct. 2013, doi: 10.1136/bmj.f6234.

[3] S. Gates, L. A. Smith, J. D. Fisher, and S. E. Lamb, "Systematic review of accuracy of screening instruments for predicting fall risk among independently living older adults," J. Rehabil. Res. Dev., vol. 45, no. 8, pp. 1105–1116, 2008.

[4] S.-H. Park, "Tools for assessing fall risk in the elderly: a systematic review and meta-analysis," Aging Clin. Exp. Res., vol. 30, no. 1, pp. 1–16, Jan. 2018, doi: 10.1007/s40520-017-0749-0.

[5] M. Moore and K. Barker, "The validity and reliability of the four square step test in different adult populations: a systematic review," Syst. Rev., vol. 6, no. 1, p. 187, Sep. 2017, doi: 10.1186/s13643-017-0577-5.

[6] A. T. Özdemir and B. Barshan, "Detecting Falls with Wearable Sensors Using Machine Learning Techniques," Sensors, vol. 14, no. 6, pp. 10691–10708, Jun. 2014, doi: 10.3390/s140610691.

[7] D. Drover, J. Howcroft, J. Kofman, and E. D. Lemaire, "Faller Classification in Older Adults Using Wearable Sensors Based on Turn and Straight-Walking Accelerometer-Based Features," Sensors, vol. 17, no. 6, 2017, doi: 10.3390/s17061321.

[8] J. Howcroft, J. Kofman, and E. D. Lemaire, "Prospective Fall-Risk Prediction Models for Older Adults Based on Wearable Sensors," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 25, no. 10, pp. 1812–1820, Oct. 2017, doi: 10.1109/TNSRE.2017.2687100.

[9] A. Nait Aicha, G. Englebienne, K. S. Van Schooten, M. Pijnappels, and B. Kröse, "Deep Learning to Predict Falls in Older Adults Based on Daily-Life Trunk Accelerometry," Sensors, vol. 18, no. 5, p. 1654, May 2018, doi: 10.3390/s18051654.

[10] A. Ejupi, S. R. Lord, and K. Delbaere, "New methods for fall risk prediction," Curr. Opin. Clin. Nutr. Metab. Care, vol. 17, no. 5, pp. 407–411, Sep. 2014, doi: 10.1097/MCO.0000000000000081.

[11] J. Howcroft, J. Kofman, and E. D. Lemaire, "Review of fall risk assessment in geriatric populations using inertial sensors," J. NeuroEngineering Rehabil., vol. 10, no. 1, p. 91, Aug. 2013, doi: 10.1186/1743-0003-10-91.

[12] N. Vanicek, S. Strike, L. McNaughton, and R. Polman, “Gait patterns in transtibial amputee fallers vs. non-fallers: Biomechanical differences during level walking,” Gait Posture, vol. 29, no. 3, pp. 415–420, Apr. 2009, doi: 10.1016/j.gaitpost.2008.10.062.

[13] N. Steinberg, A. Gottlieb, I. Siev-Ner, and M. Plotnik, “Fall incidence and associated risk factors among people with a lower limb amputation during various stages of recovery - a systematic review,” Disabil. Rehabil., vol. 41, no. 15, pp. 1778–1787, 2019, doi: 10.1080/09638288.2018.1449258.

[14] C. K. Wong, S. T. Chihuri, and G. Li, “Risk of fall-related injury in people with lower limb amputations: A prospective cohort study,” J. Rehabil. Med., vol. 48, no. 1, pp. 80–85, Jan. 2016, doi: 10.2340/16501977-2042.

[15] J. Silva et al., “Comparing Machine Learning Approaches for Fall Risk Assessment:,” in Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies, Porto, Portugal, 2017, pp. 223–230, doi: 10.5220/0006227802230230.

[16] “Quantitative Falls Risk Assessment Using the Timed Up and Go Test - IEEE Journals & Magazine.” https://ieeexplore.ieee.org/abstract/document/5594623 (accessed Mar. 06, 2020).

[17] A. Hua et al., “Accelerometer-based predictive models of fall risk in older women: a pilot study,” Npj Digit. Med., vol. 1, no. 1, pp. 1–8, Jul. 2018, doi: 10.1038/s41746-018-0033-5.

[18] M. Marschollek et al., “Sensor-based Fall Risk Assessment – an Expert ‘to go,’” Methods Inf. Med., vol. 50, no. 5, pp. 420–426, 2011, doi: 10.3414/ME10-01-0040.

[19] D. Giansanti, G. Maccioni, S. Cesinaro, F. Benvenuti, and V. Macellari, “Assessment of fall-risk by means of a neural network based on parameters assessed by a wearable device during posturography,” Med. Eng. Phys., vol. 30, no. 3, pp. 367–372, Apr. 2008, doi: 10.1016/j.medengphy.2007.04.006.

[20] M. Mancini et al., “Continuous Monitoring of Turning Mobility and Its Association to Falls and Cognitive Function: A Pilot Study,” J. Gerontol. Ser. A, vol. 71, no. 8, pp. 1102–1108, Aug. 2016, doi: 10.1093/gerona/glw019.

[21] A. D. Segal, M. S. Orendurff, J. M. Czerniecki, J. B. Shofer, and G. K. Klute, “Local dynamic stability in turning and straight-line gait,” J. Biomech., vol. 41, no. 7, pp. 1486–1493, Jan. 2008, doi: 10.1016/j.jbiomech.2008.02.012.

[22] M. Justine, H. Manaf, A. Sulaiman, S. Razi, and H. A. Alias, “Sharp Turning and Corner Turning: Comparison of Energy Expenditure, Gait Parameters, and Level of Fatigue among Community-Dwelling Elderly,” BioMed Research International, 2014. https://www.hindawi.com/journals/bmri/2014/640321/ (accessed May 04, 2020).

[23] P. L. Enright, “The Six-Minute Walk Test,” Respir. Care, vol. 48, no. 8, pp. 783–785, Aug. 2003.

[24] D. A. Winter, “Biomechanics and Motor Control of Human Movement / D.A. Winter.,” Book Ser. Process., Jan. 1990, doi: 10.1002/9780470549148.

[25] J. Perry, “Gait Analysis: Normal and Pathological Function,” J. Sports Sci. Med., vol. 9, no. 2, p. 353, Jun. 2010.

[26] Epomedicine, “Physical Examination: Gait,” Epomedicine, May 26, 2014. https://epomedicine.com/clinical-medicine/physical-examination-gait (accessed Aug. 15, 2019).

[27] C. A. Pfortmueller, M. Kunz, G. Lindner, A. Zisakis, S. Puig, and A. K. Exadaktylos, “Fall-Related Emergency Department Admission: Fall Environment and Settings and Related Injury Patterns in 6357 Patients with Special Emphasis on the Elderly,” Sci. World J., vol. 2014, pp. 1–6, 2014, doi: 10.1155/2014/256519.

[28] H. M. Gooday and J. Hunter, “Preventing falls and stump injuries in lower limb amputees during inpatient rehabilitation: completion of the audit cycle,” Clin. Rehabil., vol. 18, no. 4, pp. 379–390, Jun. 2004, doi: 10.1191/0269215504cr738oa.

[29] T. Pauley, M. Devlin, and K. Heslin, “Falls Sustained During Inpatient Rehabilitation After Lower Limb Amputation: Prevalence and Predictors,” Am. J. Phys. Med. Rehabil., vol. 85, no. 6, pp. 521–532, Jun. 2006, doi: 10.1097/01.phm.0000219119.58965.8c.

[30] S. Chihuri and C. K. Wong, “Factors associated with the likelihood of fall-related injury among people with lower limb loss,” Inj. Epidemiol., vol. 5, Nov. 2018, doi: 10.1186/s40621-018-0171-x.

[31] L. Nolan, A. Wit, K. Dudziñski, A. Lees, M. Lake, and M. Wychowañski, “Adjustments in gait symmetry with walking speed in trans-femoral and trans-tibial amputees,” Gait Posture, vol. 17, no. 2, pp. 142–151, Apr. 2003, doi: 10.1016/S0966-6362(02)00066-8.

[32] E. Isakov, H. Burger, J. Krajnik, M. Gregoric, and C. Marincek, “Influence of speed on gait parameters and on symmetry in transtibial amputees,” Prosthet. Orthot. Int., vol. 20, no. 3, pp. 153–158, Dec. 1996, doi: 10.3109/03093649609164437.

[33] “Gait deviations in amputees,” Physiopedia. https://www.physio-pedia.com/Gait_deviations_in_amputees (accessed May 05, 2020).

[34] E. Isakov, J. Mizrahi, H. Ring, Z. Susak, and N. Hakim, “Standing sway and weight-bearing distribution in people with below-knee amputations,” Arch. Phys. Med. Rehabil., vol. 73, no. 2, pp. 174–178, Feb. 1992.

[35] N. Vanicek, S. C. Strike, L. McNaughton, and R. Polman, “Lower Limb Kinematic and Kinetic Differences between Transtibial Amputee Fallers and Non-Fallers,” Prosthet. Orthot. Int., vol. 34, no. 4, pp. 399–410, Dec. 2010, doi: 10.3109/03093646.2010.480964.

[36] S. R. Lord et al., “The Effect of an Individualized Fall Prevention Program on Fall Risk and Falls in Older People: A Randomized, Controlled Trial,” J. Am. Geriatr. Soc., vol. 53, no. 8, pp. 1296–1304, 2005, doi: 10.1111/j.1532-5415.2005.53425.x.

[37] W. Dite, H. J. Connor, and H. C. Curtis, “Clinical Identification of Multiple Fall Risk Early After Unilateral Transtibial Amputation,” Arch. Phys. Med. Rehabil., vol. 88, no. 1, pp. 109–114, Jan. 2007, doi: 10.1016/j.apmr.2006.10.015.

[38] W. Dite and V. A. Temple, “A clinical test of stepping and change of direction to identify multiple falling older adults,” Arch. Phys. Med. Rehabil., vol. 83, no. 11, pp. 1566–1571, Nov. 2002, doi: 10.1053/apmr.2002.35469.

[39] S.-H. Park, “Tools for assessing fall risk in the elderly: a systematic review and meta-analysis,” Aging Clin. Exp. Res., vol. 30, no. 1, pp. 1–16, Jan. 2018, doi: 10.1007/s40520-017-0749-0.

[40] D. J. Drover, “Evaluation of Accelerometer-Based Walking-Turn Features for Fall-Risk Assessment in Older Adults,” May 2017, Accessed: May 19, 2020. [Online]. Available: https://uwspace.uwaterloo.ca/handle/10012/11914.

[41] M. T. Thigpen, K. E. Light, G. L. Creel, and S. M. Flynn, “Turning Difficulty Characteristics of Adults Aged 65 Years or Older,” Phys. Ther., vol. 80, no. 12, pp. 1174–1187, Dec. 2000, doi: 10.1093/ptj/80.12.1174.

[42] “ATS Statement,” Am. J. Respir. Crit. Care Med., vol. 166, no. 1, pp. 111–117, Jul. 2002, doi: 10.1164/ajrccm.166.1.at1102.

[43] Y. Liu, L. Nie, L. Liu, and D. S. Rosenblum, “From action to activity: Sensor-based activity recognition,” Neurocomputing, vol. 181, pp. 108–115, Mar. 2016, doi: 10.1016/j.neucom.2015.08.096.

[44] S. C. Mukhopadhyay, “Wearable Sensors for Human Activity Monitoring: A Review,” IEEE Sens. J., vol. 15, no. 3, pp. 1321–1330, Mar. 2015, doi: 10.1109/JSEN.2014.2370945.

[45] S. Patel, H. Park, P. Bonato, L. Chan, and M. Rodgers, “A review of wearable sensors and systems with application in rehabilitation,” J. NeuroEngineering Rehabil., vol. 9, no. 1, p. 21, Apr. 2012, doi: 10.1186/1743-0003-9-21.

[46] P. C. Fino, F. B. Horak, and C. Curtze, “Inertial sensor-based centripetal acceleration as a correlate for lateral margin of stability during walking and turning,” bioRxiv, p. 768192, Jan. 2020, doi: 10.1101/768192.

[47] A. Danielsen, H. Olofsen, and B. A. Bremdal, “Increasing fall risk awareness using wearables: A fall risk awareness protocol,” J. Biomed. Inform., vol. 63, pp. 184–194, Oct. 2016, doi: 10.1016/j.jbi.2016.08.016.

[48] A. Weiss, A. Mirelman, A. S. Buchman, D. A. Bennett, and J. M. Hausdorff, “Using a Body-Fixed Sensor to Identify Subclinical Gait Difficulties in Older Adults with IADL Disability: Maximizing the Output of the Timed Up and Go,” PLoS ONE, vol. 8, no. 7, 2013, doi: 10.1371/journal.pone.0068885.

[49] F. Riva, M. J. P. Toebes, M. Pijnappels, R. Stagni, and J. H. van Dieën, “Estimating fall risk with inertial sensors using gait stability measures that do not require step detection,” Gait Posture, vol. 38, no. 2, pp. 170–174, Jun. 2013, doi: 10.1016/j.gaitpost.2013.05.002.

[50] P. A. Silva, F. Nunes, A. Vasconcelos, M. Kerwin, R. Moutinho, and P. Teixeira, “Using the Smartphone Accelerometer to Monitor Fall Risk while Playing a Game: The Design and Usability Evaluation of Dance! Don’t Fall,” in Foundations of Augmented Cognition, Berlin, Heidelberg, 2013, pp. 754–763, doi: 10.1007/978-3-642-39454-6_81.

[51] A. J. A. Majumder, I. Zerin, S. I. Ahamed, and R. O. Smith, “A multi-sensor approach for fall risk prediction and prevention in elderly,” ACM SIGAPP Appl. Comput. Rev., vol. 14, no. 1, pp. 41–52, Mar. 2014, doi: 10.1145/2600617.2600621.

[52] T. Isho, H. Tashiro, and S. Usuda, “Accelerometry-Based Gait Characteristics Evaluated Using a Smartphone and Their Association with Fall Risk in People with Chronic Stroke,” J. Stroke Cerebrovasc. Dis., vol. 24, no. 6, pp. 1305–1311, Jun. 2015, doi: 10.1016/j.jstrokecerebrovasdis.2015.02.004.

[53] N. A. Capela, E. D. Lemaire, and N. Baddour, “Novel algorithm for a smartphone-based 6-minute walk test application: algorithm, application development, and evaluation,” J. Neuroengineering Rehabil., vol. 12, p. 19, Feb. 2015, doi: 10.1186/s12984-015-0013-9.

[54] “Does the Evaluation of Gait Quality During Daily Life Provide Insight Into Fall Risk? A Novel Approach Using 3-Day Accelerometer Recordings - Aner Weiss, Marina Brozgol, Moran Dorfman, Talia Herman, Shirley Shema, Nir Giladi, Jeffrey M. Hausdorff, 2013.” https://journals.sagepub.com/doi/full/10.1177/1545968313491004 (accessed May 19, 2020).

[55] B. G. Hordacre, C. Barr, B. L. Patritti, and M. Crotty, “Assessing Gait Variability in Transtibial Amputee Fallers Based on Spatial-Temporal Gait Parameters Normalized for Walking Speed,” Arch. Phys. Med. Rehabil., vol. 96, no. 6, pp. 1162–1165, Jun. 2015, doi: 10.1016/j.apmr.2014.11.015.

[56] J. Howcroft, E. D. Lemaire, and J. Kofman, “Wearable-Sensor-Based Classification Models of Faller Status in Older Adults,” PLOS ONE, vol. 11, no. 4, p. e0153240, Apr. 2016, doi: 10.1371/journal.pone.0153240.

[57] H. J. Yack and R. C. Berger, “Dynamic Stability in the Elderly: Identifying a Possible Measure,” J. Gerontol., vol. 48, no. 5, pp. M225–M230, Sep. 1993, doi: 10.1093/geronj/48.5.M225.

[58] J. Li et al., “Feature Selection: A Data Perspective.” Association for Computing Machinery, Dec. 06, 2017, Accessed: Mar. 27, 2020. [Online]. Available: https://doi.org/10.1145/3136625.

[59] M. A. Hall and L. A. Smith, “Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper,” p. 5.

[60] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, “Relief-based feature selection: Introduction and review,” J. Biomed. Inform., vol. 85, pp. 189–203, Sep. 2018, doi: 10.1016/j.jbi.2018.07.014.

[61] T. Yiu, “The Curse of Dimensionality,” Medium, Jul. 20, 2019. https://towardsdatascience.com/the-curse-of-dimensionality-50dc6e49aa1e (accessed Mar. 27, 2020).

[62] Z. M. Hira and D. F. Gillies, “A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data,” Adv. Bioinforma., vol. 2015, pp. 1–13, 2015, doi: 10.1155/2015/198363.

[63] Y. Charfaoui, “Hands-on with Feature Selection Techniques: Embedded Methods,” Medium, Mar. 05, 2020. https://heartbeat.fritz.ai/hands-on-with-feature-selection-techniques-embedded-methods-84747e814dab (accessed May 05, 2020).

[64] M. A. Hall, “Correlation-based Feature Selection for Machine Learning,” University of Waikato, 1999.

[65] F. Morstatter and H. Liu, “Advancing Feature Selection Research − ASU Feature Selection Repository,” 2010.

[66] J. Howcroft, “Evaluation of Wearable Sensors as an Older Adult Fall Risk Assessment Tool,” Jul. 2016, Accessed: Aug. 27, 2020. [Online]. Available: https://uwspace.uwaterloo.ca/handle/10012/10587.

[67] Md. T. Uddin and Md. A. Uddiny, “Human activity recognition from wearable sensors using extremely randomized trees,” in 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), May 2015, pp. 1–6, doi: 10.1109/ICEEICT.2015.7307384.

[68] S. Ronaghan, “The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark,” Medium, Nov. 01, 2019. https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3 (accessed May 05, 2020).

[69] “1.11. Ensemble methods — scikit-learn 0.22.1 documentation.” https://scikit-learn.org/stable/modules/ensemble.html#forest (accessed Jan. 13, 2020).

[70] F. Ceballos, “An Intuitive Explanation of Random Forest and Extra Trees Classifiers,” Medium, Apr. 06, 2020. https://towardsdatascience.com/an-intuitive-explanation-of-random-forest-and-extra-trees-classifiers-8507ac21d54b (accessed May 06, 2020).

[71] N. Bhandari, “ExtraTreesClassifier,” Medium, Oct. 22, 2018. https://medium.com/@namanbhandari/extratreesclassifier-8e7fc0502c7 (accessed May 06, 2020).

[72] “Random forest vs extra trees,” The Kernel Trip. /statistics/random-forest-vs-extra-tree/ (accessed Mar. 24, 2020).

[73] R. Gandhi, “Support Vector Machine — Introduction to Machine Learning Algorithms,” Towards Data Science, Jun. 07, 2018. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 (accessed Dec. 19, 2018).

[74] S. Luo, “Optimization: Loss Function Under the Hood (Part III),” Medium, Oct. 17, 2018. https://towardsdatascience.com/optimization-loss-function-under-the-hood-part-iii-5dff33fa015d (accessed May 19, 2020).

[75] O. Harrison, “Machine Learning Basics with the K-Nearest Neighbors Algorithm,” Medium, Jul. 14, 2019. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761 (accessed May 19, 2020).

[76] “KNN Algorithm using Python | K Nearest Neighbors Algorithm,” Edureka, Jul. 26, 2018. https://www.edureka.co/blog/k-nearest-neighbors-algorithm/ (accessed May 19, 2020).

[77] R. Gandhi, “Naive Bayes Classifier,” Medium, May 17, 2018. https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c (accessed May 19, 2020).

[78] Richard O. Duda, Peter E. Hart, and David G. Stork, Pattern Classification, 2nd Edition. John Wiley & Sons.

[79] M. A. Nielsen, “Neural Networks and Deep Learning,” 2015, Accessed: Feb. 06, 2020. [Online]. Available: http://neuralnetworksanddeeplearning.com.

[80] G. Ognjanovski, “Everything you need to know about Neural Networks and Backpropagation — Machine Learning Made Easy…,” Medium, Feb. 07, 2019. https://towardsdatascience.com/everything-you-need-to-know-about-neural-networks-and-backpropagation-machine-learning-made-easy-e5285bc2be3a (accessed Feb. 06, 2020).

[81] J. Brownlee, “How to Configure the Number of Layers and Nodes in a Neural Network,” Machine Learning Mastery, Jul. 26, 2018. https://machinelearningmastery.com/how-to-configure-the-number-of-layers-and-nodes-in-a-neural-network/ (accessed May 21, 2020).

[82] J. Brownlee, “Loss and Loss Functions for Training Deep Learning Neural Networks,” Machine Learning Mastery, Jan. 27, 2019. https://machinelearningmastery.com/loss-and-loss-functions-for-training-deep-learning-neural-networks/ (accessed May 18, 2020).

[83] R. Khandelwal, “Overview of different Optimizers for neural networks,” Medium, Feb. 04, 2019. https://medium.com/datadriveninvestor/overview-of-different-optimizers-for-neural-networks-e0ed119440c3 (accessed May 18, 2020).

[84] R. Gandhi, “A Look at Gradient Descent and RMSprop Optimizers,” Medium, Jun. 19, 2018. https://towardsdatascience.com/a-look-at-gradient-descent-and-rmsprop-optimizers-f77d483ef08b (accessed May 18, 2020).

[85] “An overview of gradient descent optimization algorithms,” Sebastian Ruder, Jan. 19, 2016. https://ruder.io/optimizing-gradient-descent/ (accessed May 18, 2020).

[86] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” ArXiv14126980 Cs, Jan. 2017, Accessed: May 18, 2020. [Online]. Available: http://arxiv.org/abs/1412.6980.

[87] J. Brownlee, “How to Choose Loss Functions When Training Deep Learning Neural Networks,” Machine Learning Mastery, Jan. 29, 2019. https://machinelearningmastery.com/how-to-choose-loss-functions-when-training-deep-learning-neural-networks/ (accessed May 18, 2020).

[88] R. Parmar, “Common Loss functions in machine learning,” Medium, Sep. 02, 2018. https://towardsdatascience.com/common-loss-functions-in-machine-learning-46af0ffc4d23 (accessed Jun. 08, 2020).

[89] S. Yadav, “Weight Initialization Techniques in Neural Networks,” Medium, Jan. 17, 2020. https://towardsdatascience.com/weight-initialization-techniques-in-neural-networks-26c649eb3b78 (accessed May 18, 2020).

[90] “Difference Between a Batch and an Epoch in a Neural Network.” https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/ (accessed May 18, 2020).

[91] J. Brownlee, “How to Control the Stability of Training Neural Networks With the Batch Size,” Machine Learning Mastery, Jan. 20, 2019. https://machinelearningmastery.com/how-to-control-the-speed-and-stability-of-training-neural-networks-with-gradient-descent-batch-size/ (accessed Jun. 08, 2020).

[92] J. Brownlee, “A Gentle Introduction to Dropout for Regularizing Deep Neural Networks,” Machine Learning Mastery, Dec. 02, 2018. https://machinelearningmastery.com/dropout-for-regularizing-deep-neural-networks/ (accessed May 18, 2020).

[93] S. Mutuvi, “Introduction to Machine Learning Model Evaluation,” Medium, Feb. 05, 2020. https://heartbeat.fritz.ai/introduction-to-machine-learning-model-evaluation-fa859e1b2d7f (accessed May 12, 2020).

[94] "Overfitting in Machine Learning: What It Is and How to Prevent It," EliteDataScience, Sep. 07, 2017. https://elitedatascience.com/overfitting-in-machine-learning (accessed May 08, 2020).

[95] E. Allibhai, "Holdout vs. Cross-validation in Machine Learning," Medium, Oct. 03, 2018. https://medium.com/@eijaz/holdout-vs-cross-validation-in-machine-learning-7637112d3f8f (accessed May 21, 2020).

[96] Bruce Wallace, "BIOM 5101 Biological Signals, Course Introduction," Carleton University, Sep. 07, 2018, [Online]. Available: https://culearn.carleton.ca/moodle/pluginfile.php/2705545/mod_resource/content/3/1-Introduction.pdf.

[97] "Cross Validation." https://www.cs.cmu.edu/~schneide/tut5/node42.html (accessed May 12, 2020).

[98] Z. Little, "StratifiedKFold v.s KFold v.s StratifiedShuffleSplit," Medium, Jan. 14, 2020. https://medium.com/@xzz201920/stratifiedkfold-v-s-kfold-v-s-stratifiedshufflesplit-ffcae5bfdf (accessed May 12, 2020).

[99] K. Nighania, "Various ways to evaluate a machine learning models performance," Medium, Jan. 30, 2019. https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15 (accessed May 12, 2020).

[100] B. Shmueli, "Matthews Correlation Coefficient is The Best Classification Metric You've Never Heard Of," Medium, Dec. 30, 2019. https://towardsdatascience.com/the-best-classification-metric-youve-never-heard-of-the-matthews-correlation-coefficient-3bf50a2f3e9a (accessed May 12, 2020).

[101] H. Sadeghi, P. Allard, F. Prince, and H. Labelle, "Symmetry and limb dominance in able-bodied gait: a review," Gait Posture, vol. 12, no. 1, pp. 34–45, Sep. 2000, doi: 10.1016/S0966-6362(00)00070-9.

[102] J. Novakovic, P. Strbac, and D. Bulatović, "Toward Optimal Feature Selection Using Ranking Methods And Classification Algorithms," Yugosl. J. Oper. Res., vol. 21, pp. 119–135, Jan. 2011, doi: 10.2298/YJOR1101119N.

[103] "ReliefFAttributeEval." https://weka.sourceforge.io/doc.dev/weka/attributeSelection/ReliefFAttributeEval.html (accessed Apr. 01, 2020).

[104] T. Yiu, "Understanding Random Forest," Medium, Aug. 14, 2019. https://towardsdatascience.com/understanding-random-forest-58381e0602d2 (accessed Mar. 19, 2020).

[105] C. Kendell, E. D. Lemaire, Y. Losier, A. Wilson, A. Chan, and B. Hudgins, "A novel approach to surface electromyography: an exploratory study of electrode-pair selection based on signal characteristics," J. NeuroEngineering Rehabil., vol. 9, no. 1, p. 24, Apr. 2012, doi: 10.1186/1743-0003-9-24.

[106] N. Liberman, "Decision Trees and Random Forests," Medium, Nov. 01, 2019. https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991 (accessed Mar. 20, 2020).

[107] T. Al-Aama, "Falls in the elderly," Can. Fam. Physician, vol. 57, no. 7, pp. 771–776, Jul. 2011.

[108] J. Mahanta, "Introduction to Neural Networks, Advantages and Applications," Medium, Jul. 12, 2017. https://towardsdatascience.com/introduction-to-neural-networks-advantages-and-applications-96851bd1a207 (accessed Apr. 29, 2020).

[109] "Keras: the Python deep learning API." https://keras.io/ (accessed May 08, 2020).

[110] "About us — scikit-learn 0.22.1 documentation." https://scikit-learn.org/stable/about.html (accessed Jan. 13, 2020).

[111] "Crowdsourced Indoor Mapping | Elsevier Enhanced Reader." https://reader.elsevier.com/reader/sd/pii/B9780128131893000058?token=64F311398D5080458177220E85569890BE8E660DA408316EA5C8F8E1ED6CB2390768266577E7DBF0995D0C1F2FF68685 (accessed Jun. 02, 2020).

[112] S. Abbate, M. Avvenuti, F. Bonatesta, G. Cola, P. Corsini, and A. Vecchio, "A smartphone-based fall detection system," Pervasive Mob. Comput., vol. 8, no. 6, pp. 883–899, Dec. 2012, doi: 10.1016/j.pmcj.2012.08.003.

[113] M. A. Habib, M. S. Mohktar, S. B. Kamaruzzaman, K. S. Lim, T. M. Pin, and F. Ibrahim, "Smartphone-Based Solutions for Fall Detection and Prevention: Challenges and Open Issues," Sensors, vol. 14, no. 4, pp. 7181–7208, Apr. 2014, doi: 10.3390/s140407181.

[114] Y. He, Y. Li, and S.-D. Bao, "Fall detection by built-in tri-accelerometer of smartphone," in Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics, Jan. 2012, pp. 184–187, doi: 10.1109/BHI.2012.6211540.

[115] "Lower limb amputations – Epidemiology and assessment – PM&R KnowledgeNow." https://now.aapmr.org/lower-limb-amputations-epidemiology-and-assessment/ (accessed Jun. 02, 2020).

Appendix A: Further Optimization of Neural Network Results

After optimizing the ETC50 feature subset, multiple data stratifications were tested, which resulted in 5% lower evaluation metrics (Chapter 4). One concern was that there may be overfitting when training. Further research was performed to test for overfitting and apply mitigation techniques. Since these investigations did not improve model performance, the information in this Appendix was not included in the main body of the thesis.

1. Testing for Overfitting

Validation loss should decrease each epoch until a minimum. If validation loss increases, the model may be overfitting to the training data (i.e., model converging to better results on the training data, but diverging on validation data). Often the epoch with the minimum validation loss is chosen as the point to stop training the model [94]. Models with no minimum validation loss or that train beyond the minimum validation loss could be overfitting, and often there is a noticeable difference between evaluation metrics in the training and testing data sets. A graph of training and testing accuracies for one fold of the fully optimized ETC50 model can be seen in Figure 17, with a 40% difference in accuracy between training and testing data. From Figure 18, the same fold has an erratic validation loss that increases. While not every fold had these results, validation loss never decreased to a clear minimum with the ETC50 feature set.

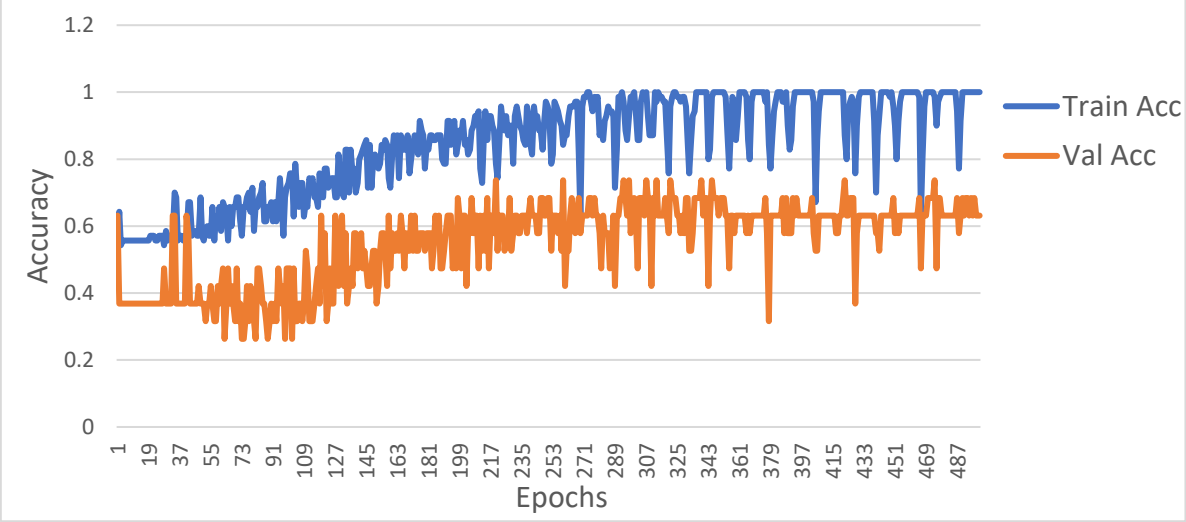

Figure 17 Training accuracy and validation accuracy for a single fold of optimized model (40% difference in training and testing accuracies).

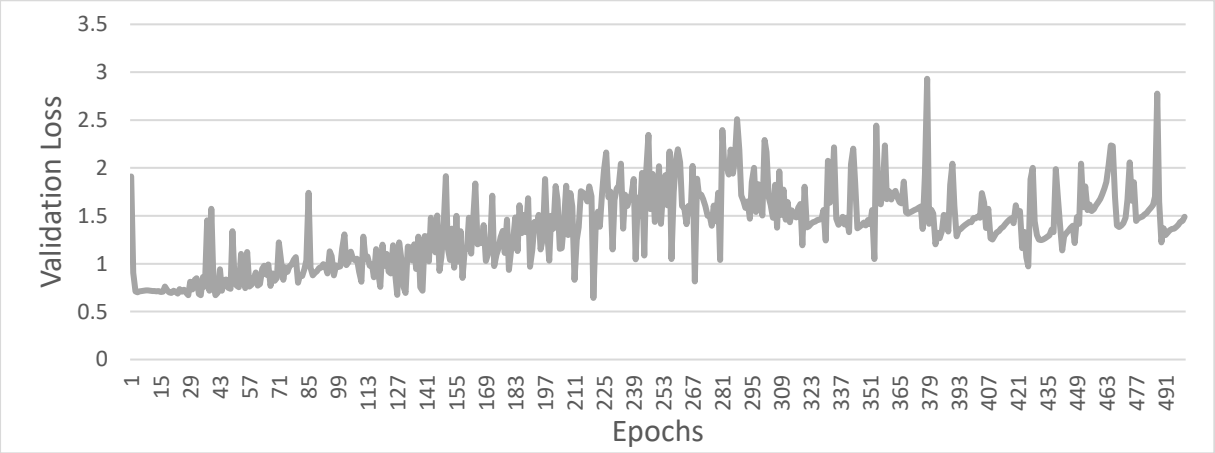

Figure 18 Validation loss for a single fold of optimized ETC50 model. Validation loss is inconsistent.

2. Reducing Overfitting with Smaller Feature Subsets

To reduce overfitting, smaller feature sets were tested. Only four data sets had validation loss decrease to a minimum in at least three of the five folds: ETC10, ETC5, RelF10, and RelF5. From Figure 19, a single fold of the ETC5 feature set had final training and testing accuracies that were only 8% apart, and a validation loss which was steadily decreasing for 1000 epochs. A variety of smaller datasets were tested with 5-fold validation to 1000 epochs, using one layer of 500 nodes (Table 11). The model with the best summed ranking was the RelF3 model.

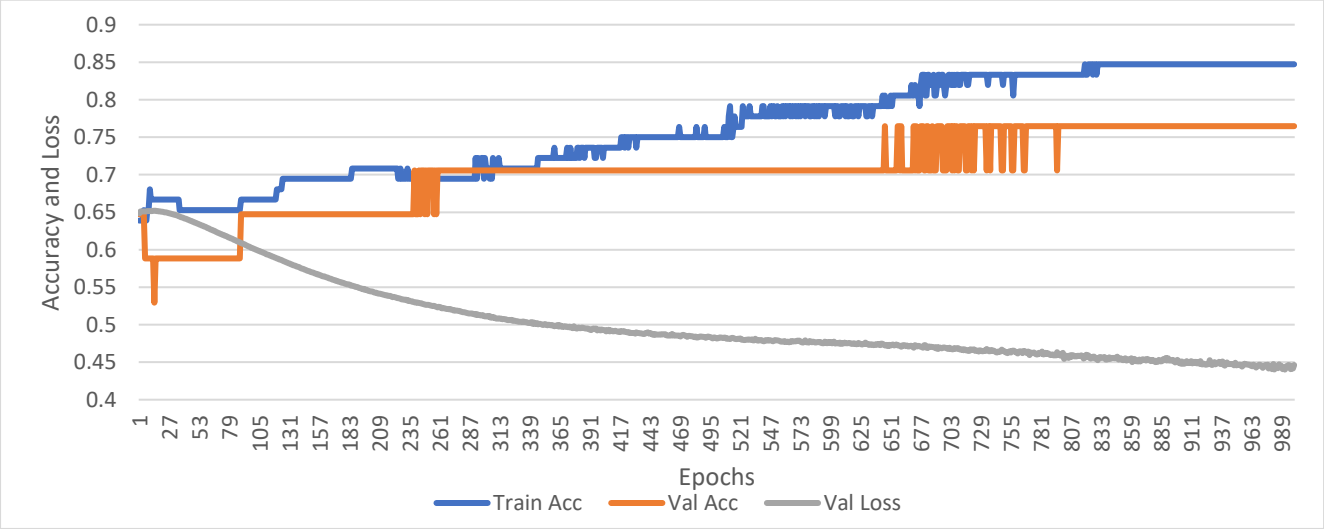

Figure 19 Training accuracy, validation accuracy, and validation loss for single fold of ETC5 feature subset. Training and validation accuracies are much closer than with the ETC50 feature set.

Table 11 Five-fold cross validated results for smaller feature sets with no optimization, using one layer and 500 nodes. SR=summed rank.

Feature Subset	Accuracy	Specificity	Sensitivity	MCC	F1	SR
RelF10	0.64	0.81	0.34	0.17	0.46	32
Relf5	0.63	0.77	0.38	0.16	0.49	35
Relf4	0.70	0.82	0.47	0.31	0.57	19
Relf3	0.72	0.84	0.50	0.37	0.60	8
ETC10	0.66	0.81	0.41	0.23	0.52	25.5
ETC5	0.70	0.79	0.53	0.33	0.62	15.5
ETC4	0.71	0.82	0.50	0.34	0.60	13
ETC3	0.64	0.84	0.28	0.15	0.40	32

Using the same optimization procedures as in Chapter 4, the RelF3 feature subset was optimized. The best results were achieved by a single 500 node layer, 0.001 learning rate, and 25% dropout. The final optimized results were 73.0% accuracy, 46.9% sensitivity, and 87.7% specificity.

While RelF3 had a slightly higher accuracy and specificity than the ETC50 feature subset achieved, RelF3 was ranked lower because it misclassified more than half of FR individuals. Additionally, when re-tested with multiple data stratifications, a large decrease in evaluation metrics was seen, similar to the ETC50 data set. The mean metrics for the RelF3 data set were 64.5±2.6% accuracy, 80.7±5.9% specificity, and 35.6±7.2% sensitivity. The sensitivity achieved by the smaller feature subsets was too low for any practical or clinical use.

3. Testing Early Stopping

Another technique used to reduce overfitting is early stopping. Early stopping was used with the RelF3 feature set along with L1 regularization and a 5-fold cross validation. The epoch at which minimum validation loss occurred was highly erratic and inconsistent at a mean 93.6±97.9 epochs. With early stopping, RelF3 had 68.3±2.6% accuracy, 25±7.0% sensitivity, and 92.6±1.5% specificity, which was not an improvement on previous techniques. Due to the extremely high variability in results and lower sensitivities, early stopping and smaller feature subsets were not pursued further.