Detecting subject-specific fatigue related changes in lifting kinematics using a machine learning approach

by

Sheldon Hawley

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Kinesiology

Waterloo, Ontario, Canada, 2020

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Background.

Functional capacity evaluations (FCEs) are used to determine a worker's capacity for return to work (RTW) or for job matching purposes. FCEs are completed using a subjective approach, where a trained evaluator will determine the capacity of a worker by monitoring for changes in movement patterns as the worker completes manual material handling tasks. It is well established that movement patterns change as workers become fatigued, explaining why evaluators are trained to watch for such changes; however, the current subjective approach used in FCEs assumes that everyone changes in the same way. In part due to the subjectivity of capacity determinations, the predictive validity and reliability of FCEs to produce accurate RTW outcomes has been questioned (Reneman, 2003). Therefore, an objective and personalized approach to detecting the onset of fatigue is needed. Machine learning may provide such an approach, specifically using an outlier detector algorithm.

Objective.

To determine if one-class support vector machines (OCSVM), an outlier detection machine learning algorithm, can be utilized to objectively identify fatigue during repetitive lifting on a subject-specific basis.

Methods.

Fourteen participants completed a repetitive lifting protocol for 60 minutes or until volitional fatigue. Whole-body kinematics were recorded using a 3D motion capture system (Vicon, Oxford, UK). Ratings of perceived exertion (RPE) and heart rate (HR) were recorded

after every 15 lifts. A whole-body kinematic model of each participant was created in Visual3D, where trajectory data from 15 landmark locations were exported for each lift. For each participant, lifts were separated into a training set and multiple test sets. The training set consisted of approximately the first 35% of lifts, and the test sets were subsequent sets of 15 lifts. Principal component analysis (PCA) was used as a data reduction and feature extraction method and applied to the training set. The PC scores from the training set data were used as features in a OCSVM. Test set data were projected back onto the training set principal component (PC) feature space. Test set PC scores were then classified against the decision boundary defined by the OCSVM. The percentage of PC scores from each test set that were beyond the boundary ("outliers") was calculated. Spearman's rank correlation (ρ), a non-parametric test, was used to assess the association between RPE, HR and the percentage of outliers in each test set.

Results.

Significant positive associations between RPE and the percentage of outliers were detected in seven of the ten participants who were likely fatigued based on their RPE. Only two of eight participants who were likely fatigued based on their HR had significant positive associations. All participants who were not likely to be fatigued had no significant association between either RPE or HR and the percentage of outliers. The OCSVM did however reveal changes in movement patterns from baseline for some participants who did not fatigue.

Conclusion.

The application of OCSVMs identified significant changes in movement patterns from baseline in those who experienced fatigue from a repetitive lifting protocol. Although no significant associations were identified in those who were not fatigued, the OCSVM still identified movement pattern changes. These results show support for use of an outlier detection tool to aid in FCE assessments to potentially reduce subjectivity, supporting improved RTW decision making.

Acknowledgements

I would like to thank everybody who has helped me along my journey. First, thank you to my supervisor Dr. Steven Fischer. Your support, guidance, and dedication to developing my education and experiences were paramount to my successes as a student. You helped me become a better researcher, critical thinker, and person.

Thank you to my committee members, Dr. Andrew Laing and Dr. Ryan Graham. The feedback and guidance you have given me for both my original and current thesis were crucial for strengthening my projects. I would also like to thank Dr. Kaylena Ehgoetz Martens for being on my original proposal committee, your feedback and suggestions truly helped guide my original thesis for the better.

I would like to acknowledge all my OBEL colleagues: Justin Davidson, Sarah Remedios, Daniel Armstrong, Nathalie Oomen, Christopher Moore, Sanjay Veerasammy, and Alexander Malone. Thank you for all your feedback, ideas and criticisms during my presentations, and for all the fun times we shared. Also, a special thank you to Nathalie for the generous use of your data for my thesis.

Thank you to Dr. Andrew Hamilton-Wright for your machine learning guidance which greatly assisted in my understanding and application of my project. I would also like to thank the undergraduate assistants I had during the piloting of my original project: Nigel Majoni and Sanjay, I will always be grateful for your willingness to come into the lab to have markers stuck on you and lift heavy boxes day after day.

Lastly, a massive thank you to all my friends and family. To my parents, your endless support and love throughout school and life will never be forgotten. To my brother, Dylan, we challenged and competed every step of the way to make each other better and we share a bond unlike any others. To my grandparents, without your love and help I would not have been able to go through school to be in this position today. Grandma, even though you are no longer here, your love and support made me who I am today, and I am sure you would be proud. Finally, a special thank you to Dominique Lepage, for everything you have and continue to do for me. I would not be who I am today without you.

Table of Contents

List of Figures

List of Tables

List of Equations

List of Acronyms

RTW – return to work

FCE – functional capacity evaluation

MMH – manual materials handling

HR – heart rate

OCSVM – one-class support vector machine

IMU – inertial measurement unit

RPE- rating of perceived exertion

PCA – principal component analysis

PC – principal component

SVM – support vector machine

BPM – beats per minute

1. Introduction

Injuries affect almost everyone at some point in their lifetime, which may disrupt the ability of a person to do work. In 2017, there was a total of 251,625 lost time claims in Canada due to work-related injury or illness (Association of Worker's Compensation Boards of Canada, 2017). In 2015, the total economic cost of injury in Canada was $26.8 billion, consisting of direct and indirect costs (Parachute, 2015). Indirect costs was totaled at $10.9 billion in 2015 (Parachute, 2015), which consist of productivity losses, employee benefits and legal costs. These indirect costs are related to lost opportunity for the worker, the employer, the workplace and community (Lebeau & Duguay, 2013). The longer a person remains away from work, the more money is spent on indirect costs of the injury (Lebeau & Duguay, 2013). Occupational injury also creates psychological issues for the injured worker, reducing functional and occupational outcomes post-injury (Kendrick et al., 2011). A longer return to work (RTW) time may be compounded by these psychological issues, creating an even larger impact of psychological outcomes and economic costs. Safely reintegrating and returning injured workers into the workplace is of utmost importance to reduce the economic burden of injury and improve the overall well-being of the worker.

Functional capacity evaluations (FCEs) are often requested by employers, insurance companies and physicians to inform RTW decisions for injured employees (Pransky & Dempsey, 2004). As seen in Figure 1, FCEs play an important role within the rehabilitation process, providing information to inform RTW decisions. FCEs provide an assessment of a person's physical and cognitive abilities to accomplish a task, termed their capacity (Gibson & Strong, 2003). Their capacity is then compared to job demands to determine if they are fit to RTW. If their capacity does not meet required job demands, FCE results may highlight opportunities to focus continued therapy to restore capacity as required or to identify if job demands can be reduced to accommodate their capabilities. In select cases, it can also establish that a worker is indeed unable to return and should instead pursue vocational training to prepare for an alternate career path. As a result of the potential uses for an FCE, an FCE can range from two-to-six hours a day, with up to two days of testing depending on the reason for the FCE and the specific tests included in the functional testing battery (King, Tuckwell, & Barrett, 1998).

FCEs are often comprised of a battery of physical tests. Many of the physical tests measure the capacity to complete dynamic manual materials handling (MMH) tasks, such as lifting, lowering and carrying (Allison et al., 2018). Depending on the purpose of the MMH test being administered, capacity can be measured as the maximum safe load a person can safely handle in that specific task, or it can be measured as the endurance to repetitively complete a task (Allison et al., 2018). A common approach for determining capacity during these physical tests is the kinesiophysical approach, which relies on the administrator of the test subjectively determining when the person has reached their capacity (Isernhagen, 1992). The evaluator uses biomechanical (visual appraisal of unsafe movement patterns and coordination) and physiological changes, such as heart rate (HR), to estimate when the person has reached their capacity.

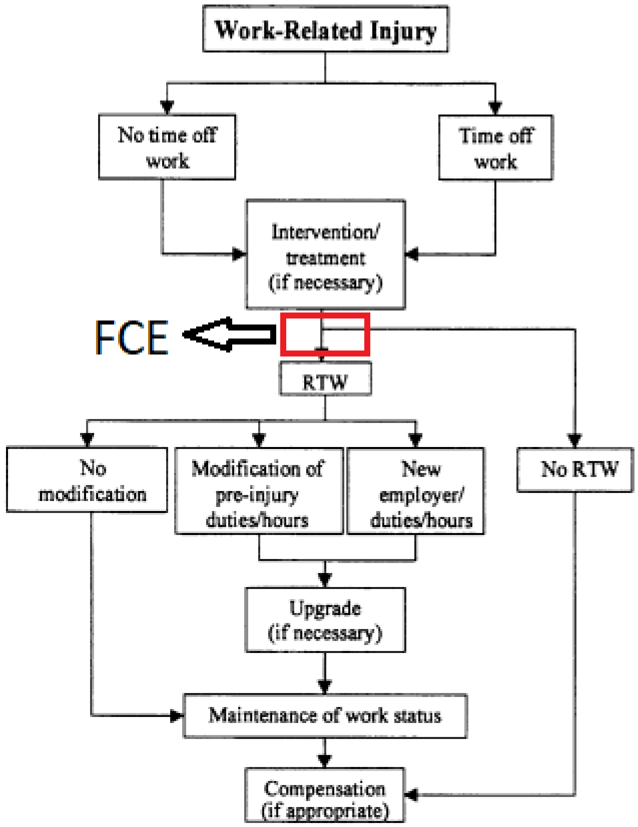

Figure 1. Overview of the occupational rehabilitation process, adapted from Innes and Straker (1998a). The red box indicates where in the rehabilitation timeline an FCE would take place to determine RTW decisions.

The predictive validity and reliability of FCEs to estimate accurate RTW outcomes has been questioned (Gouttebarge et al., 2004; Gross & Battié, 2005; Reneman, 2003). The subjectivity of determining the capacity from pre-determined biomechanical observations may be a reason why FCE reliability continues to be questioned (Sinden et al., 2017). The biomechanical criteria often used to inform evaluator decisions includes muscle recruitment (observation of muscle bulging), movement patterns, coordination and balance (Allison et al., 2018). Accurate use of pre-determined biomechanical observations relies on the experience of the test administrator to correctly identify the biomechanical criteria during the test. For example, an evaluator may use how close to the body a worker is holding a load during a lift as one of the criteria for determining an endpoint. However, this criterion can be vague and might not be consistent across lifts, or people. If there was an objective method to help determine when someone was reaching their capacity, the validity and reliability of FCE outcomes may improve.

Different methods have emerged to identify and classify changes in subject-specific movement patterns based on objective data and criteria (Chan et al., 2020; Kobsar & Ferber, 2018). In fact, different techniques have been demonstrated to detect changes in movement patterns when a person may be exhibiting fatigue during a repetitive spine flexion-extension task (Chan et al., 2020), or after a clinical gait intervention (Kobsar & Ferber, 2018) as examples. Diving deeper, Chan et al. (2020) used ten movement variables, such as angular velocity, acceleration, and repetition time to create an individualized spine motion composite index. The index was calculated from 50 baseline flexion-extension repetitions, and then again in ten subsequent repetitive fatiguing sets of 50 repetitions. Changes in movement patterns were calculated as the number of standard deviations above or below the mean baseline composite index measurement. Alternatively, Kobsar and Ferber (2018) trained a one-class support vector machine (OCSVM) to objectively classify relevant within-individual changes in movement based on their deviation from their normal range of behavior. OCSVM essentially defines a decision boundary from "typical" data and classifies new data as either fitting within or outside of that decision boundary (Kobsar & Ferber, 2018), serving as an outlier detection algorithm. Such an approach may be useful to objectively identify fatigue related changes in movement during the performance of continuous MMH tasks within an FCE paradigm.

In the context of lifting, a common FCE task, a person's "typical" lifting behavior could be identified, and then as the lifter begins to adapt their lifting pattern as a result of fatigue, subsequent lifts could be classified as differing based on the decision boundary. This is illustrated in Figure 2 where the blue circles represent exemplar baseline lifts, and the teal outline represents the OCSVM determined decision boundary. The red 'X's represent new exemplar lifts that were classified as either inliers or outliers based on whether the data fell within or beyond the decision boundary. Outliers that fall outside of the decision boundary can be considered different than the baseline lifts. Note that it is possible for the machine learning algorithm to classify baseline data as outliers as well, where OCSVM performance is dependent on hyperparameter selection (Wang et al., 2018). The ability to objectively identify when an individual lifter adapts their movement patterns as a result of fatigue may help strengthen the ability to predict an individuals prolonged workability.

Lifting kinematics have been shown to change because of fatigue during repetitive lifting. When compared to the baseline lifting movement pattern, shoulder, elbow and trunk kinematics have differed when fatigued during repetitive lifting (Fischer et al., 2015; Mehta et al., 2014). Using full-body motion capture to drive the model, a OCSVM may be able to objectively detect fatigue related kinematic changes. In the FCE context, an objective method for identifying biomechanical related changes from factors such as fatigue may lead to more accurate FCE outcomes and capacity determinations. Instead of relying on the subjective evaluation of the administrator, machine learning could automatically detect these changes based solely on motion data. This study will help identify the applicability and feasibility of such an approach.

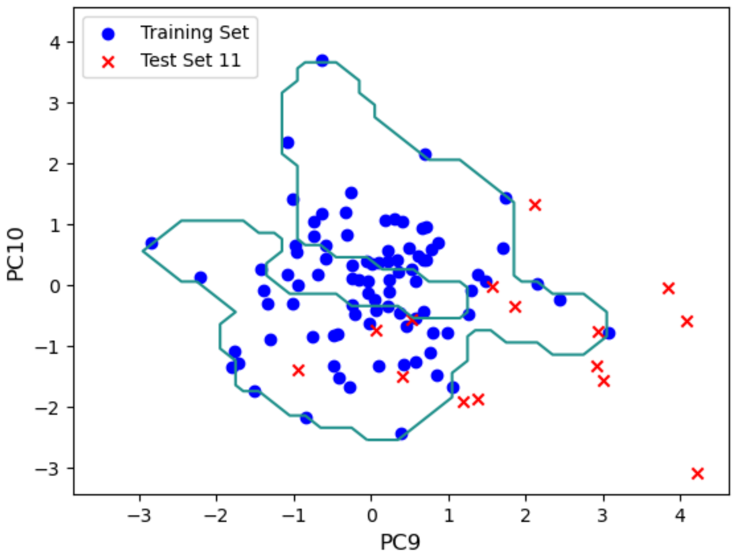

Figure 2. Example of how one-class support vector machines create a decision boundary (teal outline) based on the features of the data. Data points are then classified as inliers or outliers. Baseline training data (used to establish the decision boundary) are shown as blue circles, and test data (classified based on the decision boundary) are shown as red 'X's. This example includes a feature space with two relevant movement features (i.e., PC 9 and PC 10) for example purposes, however a OCSVM may have a multi-dimensional decision boundary.

2. Literature Review

2.1 Functional Capacity Evaluations

FCEs are used in the rehabilitation process to help make RTW decisions for employees recovering from an injury. FCEs produce an estimate of a person's capacity to do work, which is then compared to the job-specific demands to determine their RTW level (Pransky & Dempsey, 2004). A worker may be prescribed to RTW at a normal level, RTW at the same job with modified duties, RTW at a different job, or not to RTW. Results of the FCE may also be used by medical and legal teams to make decisions about further rehabilitation options and worker's compensation claims (Innes & Straker, 1998b).

Isernhagen (1992) describes and contrasts two styles of FCEs commonly used that differ in their approaches to generate capacity determinations. First, Isernhagen (1992) describes the kinesiophysical approach. This approach relies on an evaluator to determine when the person has reached their maximum safe load. The evaluator uses biomechanical (unsafe movement patterns and coordination) and physiological changes, such as HR, to decide when the person has reached their max capacity. Other signs of fatigue are also used, such as recruitment of secondary muscles to complete a movement when the primary muscles are fatigued (Isernhagen, 1992). This method emphasizes safe movement patterns throughout testing. The other approach that Isernhagen (1992) describes is known as the psychophysical approach. In contrast to the kinesiophysical approach, the psychophysical approach relies on the person completing the FCE to determine when they have reached their self-selected maximum safe load. This approach is sensitive to factors such as pain, fear of reinjury and willingness to RTW. If a person is influenced by one of these factors, their capacity determination may be lower than what they are truly capable of (Isernhagen, 1992).

Extensive research has aimed to determine the extent of FCE usefulness and rigor. Many different FCE systems have been developed, so it is important that these systems are validated with evidence of their practicality. Psychometric properties are crucial for FCEs to exhibit for validation that they are tools that can be safely and reliably used for accurate RTW decisions (King et al., 1998; Reneman, 2003). Two key psychometric properties important for FCEs are validity and reliability (Gross, 2004). Subsequent sections will discuss research completed on these properties on various FCE systems, with a focus on predictive validity, interrater and intrarater reliability.

2.1.1 FCE Predictive Validity

Predictive validity is a crucial psychometric property that FCEs must exhibit in order to be valuable (Gross, 2004). Predictive validity within an FCE context means that the FCE results correctly predicts future RTW level and performance (King et al., 1998). The determination made by the evaluator needs to be compared to what occurs in the future for that individual's RTW status. For example, if the FCE outcome concludes a worker can RTW in two weeks with minimal modifications to their job and the worker does not RTW for another 3 months, the FCE would have low predictive validity, especially if incorrect predictions are a consistent trend with the FCE system. Issues that may contribute to a low predictive validity include poor characterization of job demands and the incorrect measurement of the capacity of an individual, through an FCE, in relation to those job demands (Pransky & Dempsey, 2004). The subjective means of an evaluator determining the capacity of an individual may also contribute to poor predictive validity.

Several studies and comprehensive reviews conducted to determine the predictive validity of RTW decisions made by various FCE systems have shown that FCEs have a generally low predictive validity. In a review looking to determine the extent and quality of evidence for validity of 28 commercially available work-related assessments being used in Australia, Innes and Straker (1999) determined that most systems had limited evidence of validity. Additionally, none of the systems had moderate to good validity in each area of validity that was measured. Of the 28 FCE systems, only 2 were shown to have good evidence of predictive validity. Similarly, a systematic review examining reliability and validity of four different FCE systems found that only one of them, the Isernhagen Work Systems FCE, had evidence of good predictive validity. Two of the remaining three systems had not demonstrated good validity in their studies, while the remaining FCE system had no studies examining validity (Gouttebarge et al., 2004). These results highlight why it is important to probe and understand how new approaches to determining capacity might improve these outcomes.

In a study measuring the predictive validity of the Isernhagen Work Systems FCE in chronic low back pain clients, Gross and Battié (2005) found that the number of failed FCE tasks and a higher maximum weight on floor-to-waist lifting were only mildly associated with faster RTW. Also, four of the six claimants that met or exceeded physical job requirements on all FCE tasks experienced a recurrent injury event or were not working at one-year follow up. The authors suggest that FCEs should not be relied upon for patients with chronic low back pain to correctly predict safe RTW. Similarly, Gross et al. (2014) found no significant difference in RTW or functional work levels at any of the follow-up periods between claimants who underwent an FCE or a functional interview as a part of their rehabilitation programs. The findings of this study showed that a functional interview, which would be safer and easier to implement, may be just as useful as an FCE. In contrast to other studies, Matheson, Isernhagen, and Hart (2002) showed that RTW level was predicted by various lifting subtests in the ISW FCE. However, they showed no evidence of the full FCE outcome predicting RTW level, using only five specific subtests of the 29 tests in the Isernhagen Work Systems FCE. All of the above-mentioned results show that the generic tests done in an FCE may not represent a worker's ability to do work in the workplace, or their ability to stay injury free.

Overall, there is lack of evidence of the predictive validity of FCEs. Of the available evidence, the quality is either poor or there is minimal evidence supporting good predictive validity. Reneman (2003) questioned the ability of FCEs to accurately predict RTW and suggests that successful RTW depends on many other factors than just functional capacity itself. He concludes that workability is multidimensional, and the FCE test is only measuring a single dimension. FCE results should be combined with other tests, and if not, the results should be used to confirm or deny the patient's beliefs about their abilities to do work (Reneman, 2003). If an objective approach using machine learning was introduced, the predictive validity of FCEs may be improved by more accurately determining when a person has reached their endpoint.

2.1.2 FCE Reliability

Another important psychometric property that FCEs should exhibit is reliability, defined as the consistency of a measurement. Specifically, the interrater and intrarater reliability are the most important reliability measures for FCEs (King et al., 1998). Good interrater reliability means that if two different evaluators administer an FCE to the same person, the outcomes should be the same. This situation may occur if one evaluator administers an FCE prior to a treatment, and a different evaluator administers one after the treatment. Any difference in scores should not be attributed to the administration of the test by the different evaluators. High intrarater reliability is when the scores of an FCE test administered twice by the same evaluator on the same person is consistent. If the evaluator gave a different score, when other factors should be unchanged, there may be an issue with the test or the evaluator (King et al., 1998).

Test-retest reliability can also refer to the consistency of the capacity determinations of an individual (Bieniek & Bethge, 2014). For example, if a person did a lifting protocol twice, they would reach the same lifting capacity both times. Although reliability measures may be influenced by the test protocol or the evaluators, reliability may also be susceptible to participant's behavior during the testing. For example, participant attitudes, beliefs and motivations for the testing may provide inconsistent results (Gross, 2004). Learning effects may also play a role, as Reneman et al. (2002) saw improved performance of lifting and carrying capacity on the second day in participants with chronic low back pain, possibly due to familiarity with the testing. Since many factors have been shown to influence reliability of FCEs, it is important that FCEs are designed to limit the possibility of inaccurate results.

Similar to validity, studies have been conducted on various FCE systems to examine the reliability of the tests. In a systematic review of 11 studies examining the reliability of the WorkWell Systems FCE, Bieniek and Bethge (2014) found that 82% of the overall reliability statistics measured were acceptable. However, 85% of the measures had poor methodological quality, and 76% of the measures were for test-retest reliability. Also, all the inter- and intra-rater reliability measures were for weight-handling and strength portions of the FCE. They concluded that all reliability measures for the weight-handling and strength portions were acceptable, but results for the other portions were inconsistent or were lacking evidence. Reneman et al. (2002) also provided evidence that the test-retest reliability of lifting and carrying in the Isernhagen Work Systems FCE was acceptable in 50 patients with chronic low back pain. There was an average difference of 1.7kg lifted and 3kg carried between the two days, which they argue is not clinically relevant. These two specific FCE systems show that their reliability measures are fairly limited and inconsistent, except in portions that may be more standardized than others, such as the weight-handling portions.

The above studies only examined reliability of two different FCE systems. Another systematic review examined evidence for reliability of four FCE systems, one of them being the Isernhagen Work Systems FCE (Gouttebarge et al., 2004). Their results showed that only the Isernhagen Work Systems FCE had studies assessing reliability, which had good evidence for interrater reliability but not intrarater reliability. None of the other FCE systems had studies examining their reliability. The same authors conducted a study two years later to assess the interrater reliability of one of the FCE systems that had no reliability studies during the time of their systematic review. They assessed five different Ergo-Kit FCE lifting tests in subjects with chronic low back pain, showing good interrater reliability (Gouttebarge et al., 2006). However, there were only two different evaluators used, and only the lifting portion of the FCE was evaluated. There is a trend showing that the available reliability studies only truly show reliability in portions of the FCE that can be easily standardized, however, evidence is very limited or non-existent.

Although there is good evidence available that some FCE systems provide sufficient reliability, the majority of researchers have examined test-retest reliability, and only test specific sections of an entire FCE protocol. There is an overall lack of reliability studies available for the majority of FCE systems. At the time of the King et al. (1998) paper, they only found evidence of inter- and intrarater reliability in peer-reviewed journals for two FCE systems when many more FCE systems are available. Experts suggest one way to increase reliability is to increase standardization of the tests (King et al., 1998). Standardization would involve aspects such as task demonstrations, data collection and analysis, and instructions given. These should not change with each different evaluator conducting the test. Introduction of an objective machine learning approach for assessing fatigue related changes to determine endpoints would help to improve FCE reliability. A machine learning approach could eliminate the vague endpoints used by evaluators. Also, opportunity for poor intra- and interrater reliability would essentially be negated since the machine learning model would be objective and work the same way every time. Overall, better reliability of FCEs would lead to better predictions of RTW leading to improved outcomes for employees and employers.

2.1.3 Other FCE Limitations

As discussed in the previous sections, FCEs are not short of debate and potential for research. Although problems with validity and reliability have seen the most research, other issues exist with FCEs. One such issue is the ability of people undergoing the test to provide dishonest effort (Gross, 2006). Evaluators conducting the FCEs are trained to detect insincere efforts and will adjust decisions based on their perceptions, but this is not perfect. There are various reasons why a person may not give full effort, such as fear of reinjury or anxiety towards RTW (Innes & Straker, 1998a), or for secondary gain such as financial compensations (Geisser et al., 2003).

To attempt to measure sincerity of effort, various methods have been studied. Variability of performance is one such method, with low variability suggesting lower sincerity of effort (Geisser et al., 2003). Jay et al. (2000) examined various indicators of sincere effort in a previously injured population while performing an EPIC lift capacity test. These indicators included HR and systolic blood pressure increases, and an evaluator's subjective evaluation of the participant's exertion. Using each of the indicators, the evaluators were able to accurately classify 86.8% of full and insincere efforts. However, the most reliable measure was the evaluators subjective evaluation of the participants. This study shows that various indicators may be valuable to assess sincerity of effort during an FCE but more research needs to be done to validate these methods. Although sincerity of effort determinations can be completed, the reasons for a person giving less than full effort should be examined before any critical RTW decisions are made (Innes & Straker, 1998b). Implementation of an objective machine learning method based solely on the worker's movements during an FCE task may be able to pick up on sincerity of effort. Some of the measures used previously, such as variability of movement, may easily be detected with a machine learning data analysis approach. In combination with an evaluator’s perception and physiological measures such as HR or RPE, a machine learning approach may help estimate sincerity of effort.

Another issue with FCEs is the length of the protocol. The time commitment needed for the evaluator and patient is extensive, especially if it is a two-day protocol. Mentioned in the FCE validity section above, some studies show that just a few factors, such as weight lifted, are just as predictive as the overall protocol itself (Gross, 2006). If a shorter protocol with specific predictive tests is developed, the length of FCEs may be decreased to reduce the time commitment burden on health care providers and employees.

The above sections have discussed issues with FCEs that may lead to the tool not correctly predicting an individual’s RTW status, as well as other debated issues. Implementation of an objective machine learning approach is not a fix to all these issues, nor will it eliminate the need for trained evaluators. However, it has potential to reduce or eliminate some of the FCE limitations. For example, the reliability of capacity determinations with an objective approach would be much better than the subjective, evaluator approach currently used since an algorithm will run consistently each time. The research and knowledge that this study can contribute to this field will provide evidence on whether an outlier detection machine learning algorithm is feasible for use in FCEs.

2.1.4 FCE Future Directions

Much of the research on FCEs was in the 1990's and early 2000's. FCE research slowed down after that period, but there has been occasional research completed as new FCE systems were developed and more updates were needed about the state of FCEs. For example, a 2018 systematic review updated available evidence of psychometric properties of FCEs post-2004 (De Baets et al., 2018). Similar to other studies presented in earlier sections, these authors found that some of the FCE systems available had extensive research but showed reliability and validity of the FCEs were limited in strength. They also found that much of the research done on specific FCE systems were often done by the same authors, which may introduce bias. The importance and use of FCEs is unchanged at present time, as it was in the past when research was extensive. Research needs to continue to develop methods and protocols that strengthen the validity and reliability of FCEs to accurately assess whether an injured person can RTW.

In a recent report on current FCE practices gathered from literature and clinical experts, Allison et al. (2018) suggest that wearable technologies is an emerging trend to be aware of for FCE practitioners. Such technologies include electrogoniometers, strain gauge sensors, accelerometers and surface electromyography. This list could possibly be extended for other advancing technologies and approaches such as machine learning. Data from any of the above sensors, as well as movement data, could be added as inputs in a machine learning model to help improve capacity determinations and RTW decisions. Doing so could enhance the validity and reliability of FCEs, while also aiding evaluators in some other problem areas, such as sincerity of effort. Also, efforts exist to use intelligent robotics incorporating machine learning to improve FCE assessments by simulating workplace tasks (Fong et al., 2020). Studies such as this thesis can help advance FCEs by implementing new technologies, resulting in better RTW decisions that benefit both employers and employees.

2.2 Considering the Use of Machine Learning Methods to Detect Changes in Movement Patterns

As mentioned previously, evaluators in FCEs are typically trained to watch for biomechanical changes that may occur while a person is completing MMH tasks (Allison et al., 2018; Isernhagen, 1992). These biomechanical changes include but are not limited to counterbalancing or leaning in the opposite direction of the load, activating accessory muscles (e.g. the trapezius muscle to aid the biceps during a heavy lift), and using unsafe movement patterns (e.g. not using the knees during a heavy lift) (Isernhagen, 1992). Although evaluators can be trained to notice these biomechanical changes to detect fatigue in an individual, not everybody may exhibit the same changes in the same way. Fatigue related movement changes are not always consistent across each person, with evidence showing that there can be an increase or decrease in movement variability depending on the kinematic variable (N. Cortes et al., 2014). Some people may also change their movement patterns when fatigued in more indiscrete ways that cannot be noticed by a trained eye. Therefore, it may be useful to use subject-specific machine learning approaches to identify and classify changes in movement patterns that may be relevant to the task being completed.

Emerging research highlights opportunities to develop subject-specific approaches to detect kinematic changes in performance. Kobsar and Ferber (2018) used subject-specific

OCSVM models to evaluate whether knee osteoarthritis patients exhibited changes in their gait after a six-week intervention. Patients' baseline gait movement patterns were measured by inertial measurement units (IMUs) before the exercise intervention and used to define a decision boundary in the OCSVM. Gait movement patterns were measured again after the exercise intervention, then were classified against the decision boundary to detect if the intervention lead to changes in the patients' gait. From this example, OCSVM was successfully used as a subject-specific outlier detection algorithm for clinical intervention purposes. OCSVM may therefore have the potential to detect changes in kinematics due to other mechanisms, such as fatigue, which may prove useful for FCEs. Fatigue related changes during prolonged MMH tasks could be detected and identified by OCSVMs.

The ability to detect subject-specific fatigue related changes in kinematics was demonstrated by Chan et al. (2020). Ten pre-selected kinematic related variables associated with muscle fatigue were measured using an IMU setup. From these measured variables, they calculated a spine motion composite index. Participants performed a baseline set, and multiple fatiguing sets of a spine flexion-extension task. They calculated the composite index on a subject-specific basis, where the composite index at the baseline set was determined to be their "typical" movement pattern for the task. The fatigue sets' composite index then represented the number of standard deviations above or below the mean of each participant's "typical" spine movement. This method shows how pre-selected, known fatigue related kinematic variables can be combined with wearable sensors to detect fatigue related changes in movement (Chan et al., 2020). These above examples demonstrate how subject-specific approaches to identifying changes in movement can be accomplished with different methods.

2.2.1 Data Reduction and Feature Extraction using Principal Component Analysis

As demonstrated by Chan et al. (2020), it is possible to choose pre-selected kinematic variables to include in a subject-specific analysis. Since evaluators in FCEs also use pre-selected variables that they try to observe and quantify (Allison et al., 2018) that may not be homogenous across individuals, it may be useful to use an approach that can automatically detect changes in kinematics to create movement- relevant features. Principal component analysis (PCA) is a feature detection and data reduction technique that detects underlying synergistic or functional patterns in a waveform that explain the greatest proportion of variance (Daffertshofer et al., 2004). Instead of using variables such as mean knee angle or peak shoulder adduction that reduce the amount of information about a movement waveform, PCA will preserve the variability of the dataset when used in an analyses (Halilaj et al., 2018). PCA is therefore useful for biomechanical analyses since human movement is variable, and fatigue related changes may not be homogenous, as previously mentioned. PCA can extract meaningful movement-relevant features on an individual basis to create more personalized analyses. PCA will extract principle components (PCs) that represent a certain amount of variance, with the first PC explaining the most amount of variance and each subsequent PC explaining less than the one before it (Brandon et al., 2013). Each waveform will then receive a PC-score for each PC, which represents the degree that the waveform exhibits the specific variance pattern of the PC (Deluzio & Astephen, 2007). Therefore, PC scores can be used to represent the overall movement pattern that was inputted into the PCA.

PC scores can also be used for any statistical testing techniques typically used (Deluzio & Astephen, 2007).

PCA is also useful as a data reduction technique that reduces the number of features of a dataset. In a review of machine learning literature in human biomechanics, Halilaj et al. (2018) found that PCA is one of the two most common approaches for reducing the number of features in a dataset prior to selecting features for use within machine learning models. Having a smaller number of features than the number of observations in most machine learning applications is crucial for the model to work properly (Phinyomark et al., 2018). Since whole-body time series kinematic data has a large amount of features (i.e. multiple three-dimensional trajectories), data reduction is needed (Halilaj et al., 2018; Phinyomark et al., 2018). Most studies assessed by Halilaj et al. (2018) retain enough PCs to explain at least 90% of the variance in the dataset. PCA continues to provide a reasonable method to support data reduction and feature extraction technique prior to the application of machine learning models where examples include the classification of runners based on running kinematics (Clermont, Phinyomark, et al., 2019; Phinyomark et al., 2014), detection of subject-specific gait movement pattern changes after a clinical intervention (Kobsar & Ferber, 2018), and for identifying movement phenotypes during the performance of deep squat and hurdle step movements (Remedios et al., 2020). PCA is therefore a useful tool to use for data reduction and feature extraction for machine learning models related to biomechanical analyses.

2.2.2 Outlier Detection using One-Class Support Vector Machines

Once a feature set has been determined, it is necessary to apply methods that can classify or group data based on those underlying features. Support vector machines (SVMs) are a machine learning approach commonly used for binary classification (Cortes & Vapnik, 1995). Essentially, the SVM will define an optimal hyperplane (also called decision boundary) that has the maximal margin between the vectors of the two classes (usually denoted as positive and negative classes) (Cortes & Vapnik, 1995). The decision function that defines the hyperplane can be linear or non-linear depending on the kernel function used (Burges, 1998). In either case, support vectors are the input data that lie closest to, or on the decision boundary for both positive and negative classes. Figure 3 shows an example of what a simple linear, 2-dimensional hyperplane and support vectors might look like (Cortes & Vapnik, 1995). More complex problems would require a higher-dimension, non-linear decision function. Examples of SVMs use in biomechanics include detecting age-related changes in running kinematics (Fukuchi et al., 2011), gender and age-related differences in lower extremity running mechanics (Phinyomark et al., 2014), classifying runners as high or low milage based on running kinematics (Clermont, Phinyomark, et al., 2019), and classifying lifting postures as correct (safe) or incorrect (unsafe) from IMU data (Conforti et al., 2020). In combination with wearable sensors and/or motion capture, SVMs are a useful tool for biomechanical analyses.

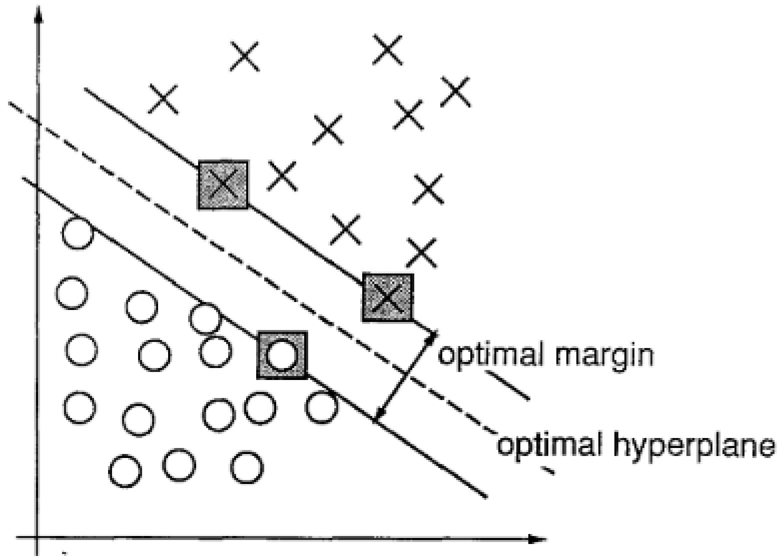

Figure 3. From Cortes & Vapnik (1995). An example of a linear, two-dimensional hyperplane. The points marked with a grey square are the support vectors which define the maximum width of the margin.

Support vector machines can also be multi-class, or one-class. In the case of multi-class models, the goal of the model is to separate each class by a hyperplane, similar to the binary classification example. This can be done by combining several binary classifiers, or by considering all of the data into an optimization problem (Hsu & Lin, 2002). In one-class SVMs, the data are not treated as having positive and negative examples. Instead, a decision boundary is computed around a training set which could be considered the "typical" data. Then, new test examples can be classified as outliers if they fall outside of the decision boundary, or inliers if they fall within (Mourão-Miranda et al., 2011).

OCSVMs can be used for many different applications. For example, Mourão-Miranda et al. (2011) used a OCSVM to define a boundary of "typical" brain activity when viewing a sad facial expression from healthy controls, then tested whether new cases (depressed patients) fit within or outside of this “typical” boundary of brain activity. Another way to use OCSVM is to detect subject-specific changes. As mentioned previously, Kobsar & Ferber (2018) used a subject-specific OCSVM to evaluate whether knee osteoarthritis patients exhibited changes in their gait after a six-week intervention. Figure 4 shows a two-dimensional representation of the decision boundary data from the Kobsar & Ferber (2018) study. From this example, OCSVM was successfully used as a subject-specific outlier detection algorithm for clinical intervention purposes. OCSVMs could also be used to identify subject-specific fatigue related changes in movement, which this study aims to explore.

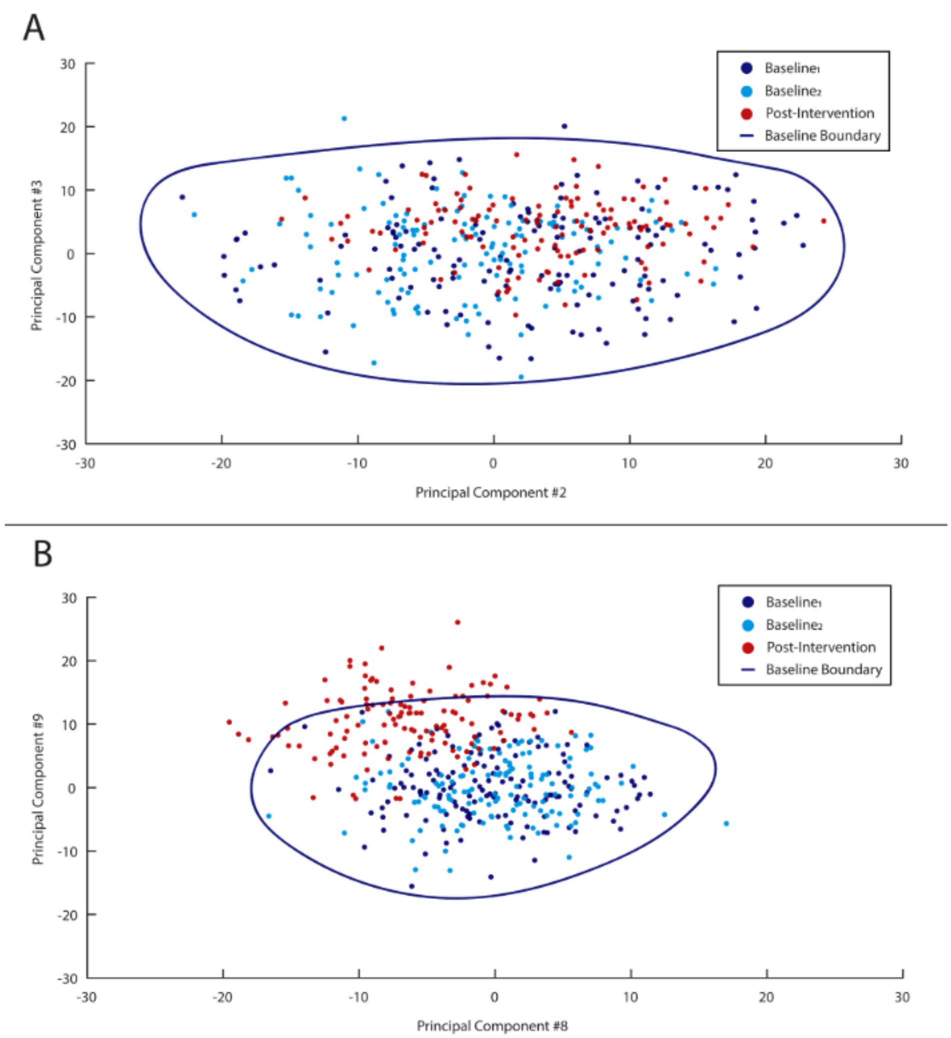

Figure 4. From Kobsar & Ferber (2018). Two-dimensional representation of the decision boundary, baseline and post-intervention data using two features. Example (A) shows no post-intervention outliers, while example (B) showcases a feature space that may better identify post-intervention outliers. In reality, a multi-dimensional feature space is used to capture relevant outliers.

2.3 Literature Review Conclusion

The literature demonstrates the importance of predictive validity and reliability of FCEs, but the available evidence for current FCE systems is limited and poor. A need to improve and conduct more research on FCEs is apparent to support improved RTW decision making, and to ultimately improve the safety and productivity of workers. The lack of an objective approach to determine capacity is highlighted as a potential area for improvement. If subjectivity could be eliminated from RTW decisions, outcomes may be improved. Using a machine learning based outlier detection approach may provide a potential solution for this problem. SVMs have been shown to be successful for classifying patterns in gait and running contexts, and in identifying correct versus incorrect lifting postures. However, whole-body kinematics have not been used as inputs into a OCSVM to detect fatigue related changes during lifting. If OCSVMs can be used to identify when a person's lifting movement patterns have changed relative to their baseline due to fatigue, it provides important proof-of-principle data that such methods might be able to overcome limitations associated with the subjectivity of an FCE evaluator's capacity determination.

3. Research Question and Hypotheses

The global objective of this thesis was to develop subject-specific models to objectively identify fatigue related changes in lifting kinematics using one-class support vector machines.

Research Question: Is the percentage of outlier movement patterns as calculated from a OCSVM machine learning model correlated with increases in ratings of perceived exertion (RPE) or HR during a single-session repetitive lifting task in healthy adults?

Hypothesis 1: There will be a significant positive correlation between the changes in lifting patterns, measured as the percentage of outliers determined by the OCSVM, and the RPE exhibited by participants that were likely to have fatigued during the lifting trials. Specifically, participants who exhibited a greater RPE will have more outliers during the final lifting trials.

Hypothesis 2: Similarly, there will be a significant positive correlation between the percentage of outliers determined by the OCSVM, and the HR exhibited by participants that were likely to have fatigued during lifting trials.

4. Methodology

4.1 Study Design

This study was conducted as a secondary analysis of data previously collected. In the original study, a cross-sectional observational-based exploratory research design was used for this study. Participants came to the Occupational Biomechanics and Ergonomics Laboratory at the University of Waterloo for data collection. Participants lifting kinematics were captured during a repetitive lifting protocol, where participants were encouraged to repetitively lift boxes at a self-selected pace for 60 minutes or until volitional fatigue.

The original study was reviewed and approved by the University of Waterloo Ethics Board (ORE #40762) prior to commencement of the study. Informed consent was received by each participant before participating in the study.

4.2 Participants

Data were obtained from fourteen healthy participants (Table 1) that were recruited from a local university student population. Participants were excluded if they had acute and/or chronic pain that interfered with prolonged lifting in the previous seven days, determined using the Nordic Musculoskeletal Disorder Questionnaire. Participants were asked about their repetitive lifting and resistance training experience and length, if any.

Table 1. Participant demographics.

		Males (N=7)	Females(N=7)
Age (years)		25.1 ± 4.1	22.3 ±3.7
Height (cm)		178.7 ± 8.7	160.7 ± 5.3
Weight (kg)		68.5 ± 10.8	60.6 ± 9.9
Repetitive Lifting Experience (# Participants)		3	4
Length of Repetitive Lifting Experience (# Participants)	≤ 1 year	2	3
	> 1 year	1	1
Resistance Training Experience (#Participants)		3	5
Length of Resistance Training Experience (# Participants)	≤ 1 year	1	2
	1 to 2 years	1	1
	> 2 years	1	2

4.3 Instrumentation

4.3.1 Motion Capture

A 12-camera (6 Vero v2.2, 6 Vantage V5) Vicon motion capture system (Vicon, Oxford, UK) was used to track 3D kinematics of the body. Position data were sampled at 100Hz using Vicon Nexus (v2.6, Vicon, Oxford, UK). Reflective markers used for static calibration were placed on anatomic landmarks, shown in Figure 5 (bilaterally on the 2nd and 5th metacarpal head, ulnar and radial styloid, medial and lateral epicondyles, acromia, C7, T8, xyphoid process, suprasternal notch, anterior superior iliac spines, posterior superior iliac spines, lateral iliac crests, greater trochanters, medial and lateral femoral condyles, medial and lateral malleoli, calcaneus tuberosity, 1st and 5th metatarsal head). Rigid clusters with four reflective markers were placed on the subject for calibration and remained on for the experimental protocol. These clusters were placed on the shanks, thighs, pelvis, trunk, upper arms, and forearms. Four markers were also placed on each of the head, both hands and both feet. The anatomical markers, excluding the ulnar and radial styloids, were removed after the static calibration was completed. The three boxes that were used for lifting also had reflective markers attached to them so their position could also be tracked during collections.

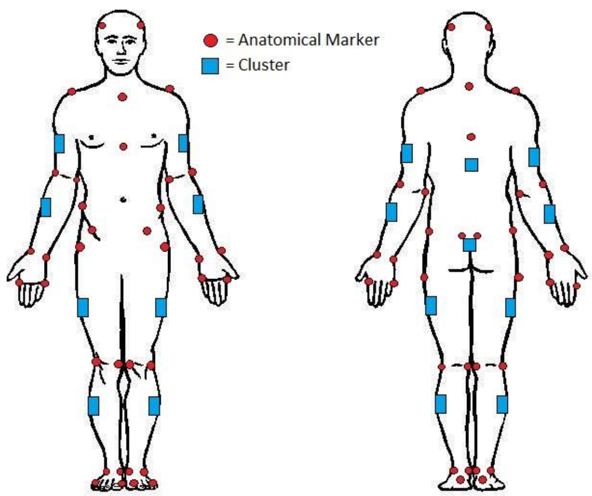

Figure 5. Anterior (left) and posterior (right) views of Vicon reflective marker placement for motion capture collection.

4.3.2 Fatigue Measures

HR was collected using a Polar FT1 Heart Rate Monitor (Polar Electro Oy, Kempele, Finland), displayed on a watch worn by the participant. Borg's 6-20 scale was used to collect RPE data (Borg, 1982). Participants were asked after every 15 lifts (approximately every 2 minutes) what their HR (shown on the watch) and RPE was.

4.4 Experimental Protocol

4.4.1 Participant Preparation

Once all consent forms and questionnaires were completed, participants were prepared with the reflective motion capture markers. Reflective markers were placed at the anatomical landmarks identified above in Figure 5 with hypoallergenic tape. Clusters were placed at their respective sites with Velcro straps. Once the participant was prepared and ready for motion capture, the participant was calibrated in the motion capture space. A static calibration trial was collected, then the anatomical markers were removed. Dynamic calibration trials were also performed that were used later to assist in data processing. The shelves were adjusted to be matched to the participant's shoulder height.

4.4.2 Lifting Protocol

Participants completed two-handed box lifts from floor-to-shoulder height at a self-selected pace until volitional fatigue, or until a maximum time limit of one hour was reached. Participants started behind a line 10 feet away from three boxes and three shelves that were adjacent to each other (Figure 6). Each shelf had its own respective box that was lifted from just above the floor onto a shoulder-height shelf. Each box contained a weight that corresponded to approximately 30% of their maximum lifting capacity as determined using Matheson's EPIC Lifting Capacity test (Matheson et al., 1995).

Participants completed trials that consisted of three consecutive lifts. Once given the signal to start, participants walked towards the shelf on the left, picked up the box from the floor and placed it on the shoulder-height shelf. They then moved from left to right, lifting each box onto the respective shelf. Once a trial of three lifts was completed, the participant walked back towards the starting line, ending that trial. Participants repeated the lifting procedure at a self-selected pace until volitional fatigue or until they reached the 60-minute time limit. HR and RPE were collected after every five trials (15 lifts) and participants were asked if they were able to continue the protocol. If they answered yes, the participant would continue the lifts. If they answered no, the protocol was stopped as the participant reached volitional fatigue.

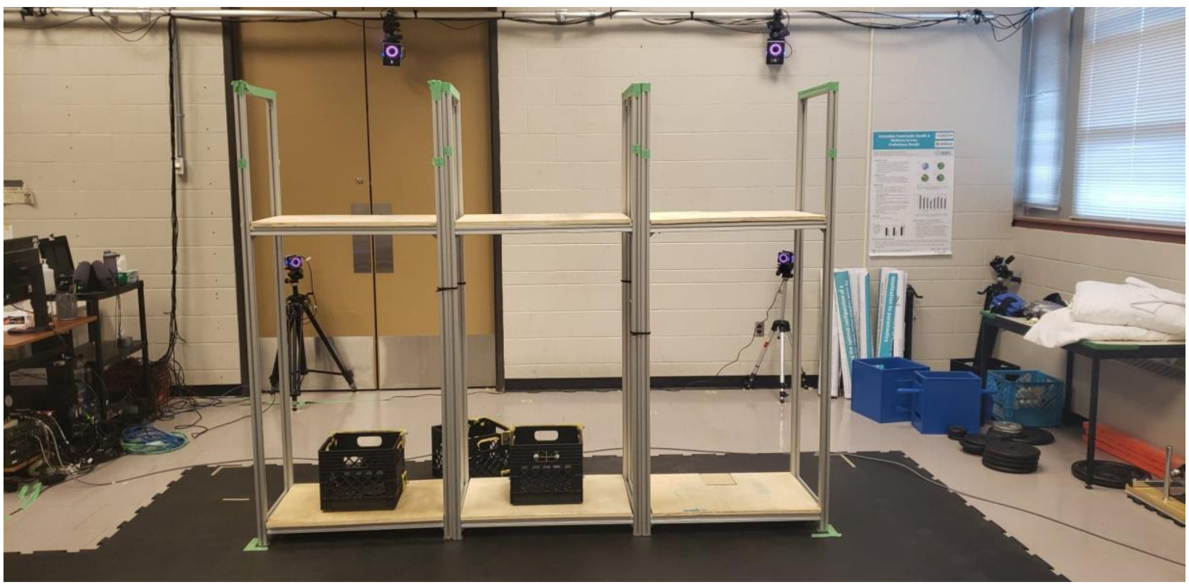

Figure 6. The shelf and box setup. There were three shelves that had its own respective box. The participants lifted the box from the shelf just above floor-height onto the shoulder-height shelf, moving from left to right. One trial consisted of three consecutive lifts.

4.5 Data Processing

Figure 7 illustrates the flow of data treatment and analysis. Methods are described in greater detail in the following sections.

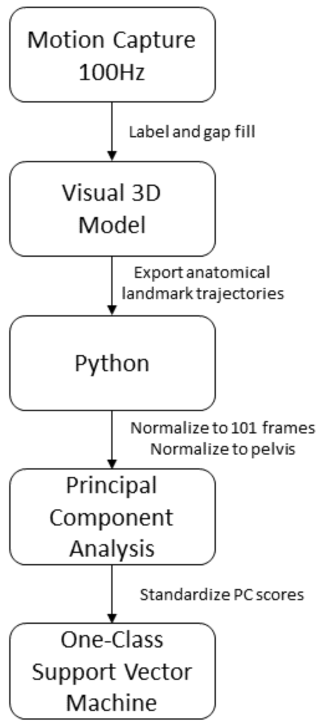

Figure 7. General overview of the data treatment and processing steps.

4.5.1 Data Treatment and Preparation

4.5.1.1 Vicon Nexus

All marker data were labelled, and gap filled in Vicon Nexus. For gaps less than 200ms in length, a cubic spline was used. If the gaps were greater than 200ms, either a pattern fill or rigid body fill was used (Howarth & Callaghan, 2010). Rigid body fills were used if there were three available markers on the rigid body cluster for the required time points. Otherwise, a pattern fill was completed using one other marker that was available from the same rigid cluster.

4.5.1.2 Visual3D

The position data were exported into Visual3D software (v6.01.03, C-Motion, Germantown, Maryland). Data were dual pass filtered in Visual3D through a low pass second order Butterworth filter with an effective cut off frequency of 6Hz (Winter, 2009). The initial cut off frequency was 7.5Hz to correct for the dual passing of the filter.

A whole-body kinematic model of each participant was created in Visual3D using the anatomical landmark position data from the static trial to define segments. The model consisted of a pelvis, thorax, head and bilateral foot, shank, thigh, upper arm, forearm and hand segments. Figure 8 shows a representation of a full-body model in Visual3D for a static standing trial. The pelvis segment was created using a CODA pelvis defined by the left and right ASIS and PSIS, and hip joint centres. The thorax segment was defined using the C7, suprasternal notch, left and right acromion, left and right ASIS and PSIS, left and right iliac crest and sacrum anatomical markers. The thigh segments were defined by the medial and lateral markers at the knee joint center, and an estimate of the hip joint centre as defined in (Bell et al., 1989). The upper arm segment was defined by the medial and lateral epicondyles distally. To define the glenohumeral joint centre, the acromion markers and upper and lower mid-torso landmarks were used. A point was projected 5cm inferiorly from the acromion along a vector from the upper and lower mid-torso (Nussbaum & Zhang, 2000). Medial and lateral markers on their proximal and distal endpoints were used to define the foot, shank, forearm, and hand segments. The head, pelvis and trunk centre of gravity, and the bilateral ankle, knee, hip, wrist, elbow, shoulder joint centres were all calculated for each frame of each lift and exported from Visual3D. These fifteen anatomical location three-dimensional data points were used to represent the whole-body trajectories during the lifts. A visual example of these 15 anatomical locations can be seen in Figure 18 (Section 6.3.1).

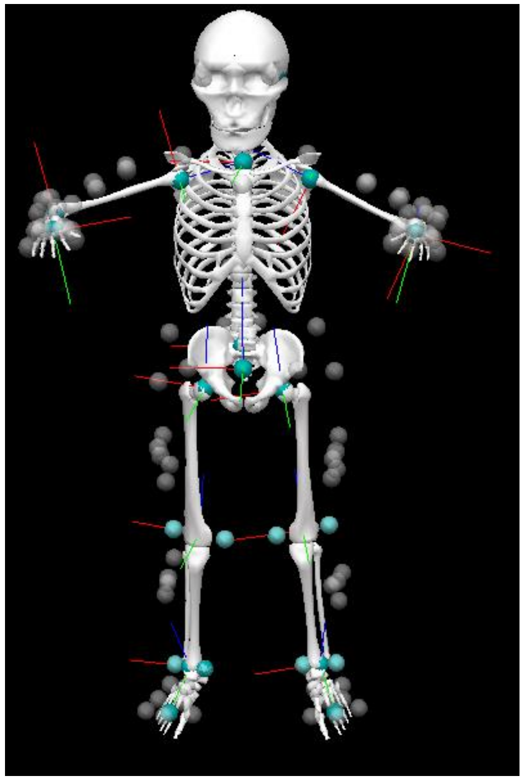

Figure 8. Full-body model of a static standing trial in Visual3D.

A trial collection consisted of three lifts, starting as the participant approached the first box, and ending once they finished the third box lift. Therefore, trial recordings needed to be segmented into individual lifts for analysis. Events in Visual3D were created to segment the trials at the appropriate frames to signify the start and end of each lift. Figure 9 shows an example of these events created in Visual3D. The start of a lift was defined as the moment a head marker reached 85% of the participants height, signifying when they were starting to bend down to pick up the box. The end of the lift was defined as the moment the box markers reached a maximum horizontal displacement along the global Y-axis (anterior to the direction that the participant was facing when lifting the box). These events were manually checked for each lift to ensure proper placement. The trajectory data for each of the fifteen anatomical locations mentioned above were exported for each lift for only the frames between the start and end events for each lift. A lift was not included if the collection of the trial started too late or ended too early (i.e. if the participant was already in the middle of a lift when the collection started, or the collection ended before the participant finished the lift).

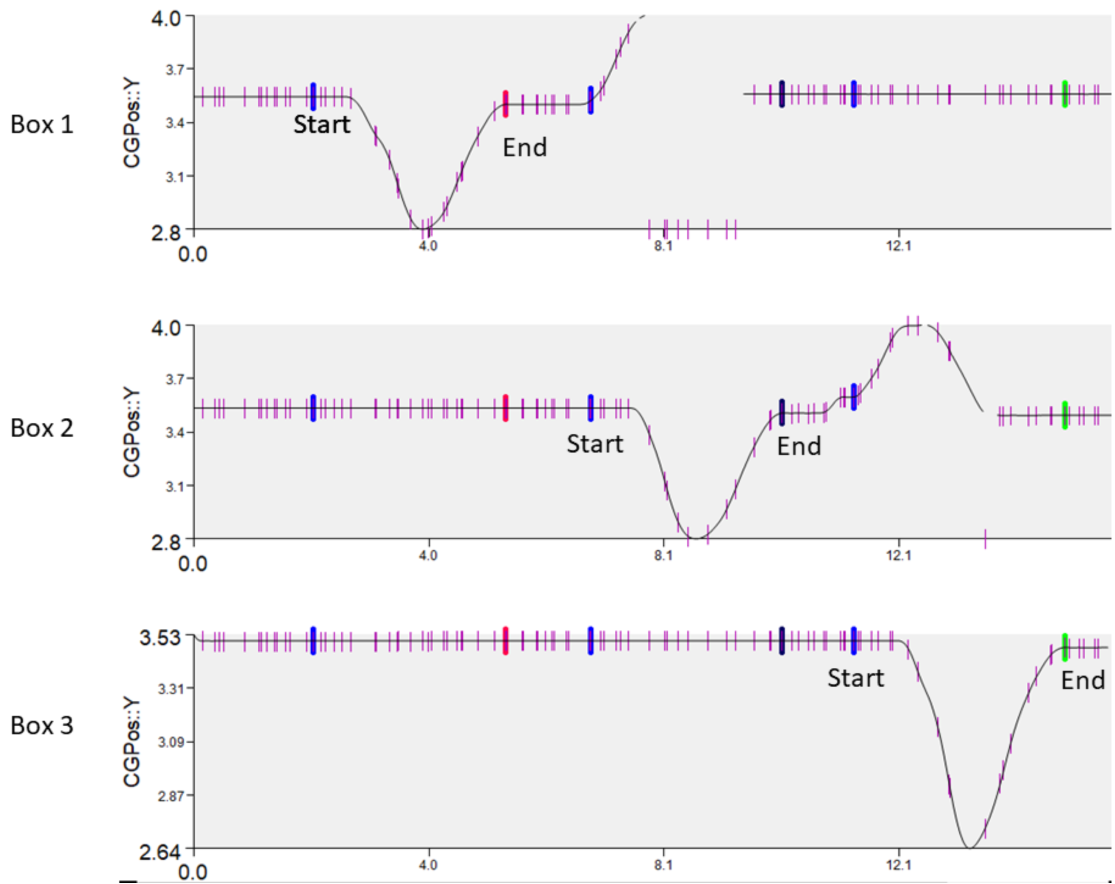

Figure 9. Example of events created in Visual3D to signal the start and end of a lift, graphed on the Y-axis (anterior-posterior) trajectory of each box. Only data in between the start and end of each lift was exported for analysis.

4.5.1.3 Python

The exported trajectory data from Visual 3D were read into Python (version 3.8) for further conditioning. The length of each lift was calculated from the total number of frames in each lift. Trajectory data for each lift was then normalized to 101 frames. Since the trajectory data were expressed in this global coordinate system, trial-to-trial variations in global positioning could induce variability in the trajectory data that could be captured by the PCA model. Therefore, these trajectory data were expressed in the local pelvis body-specific coordinate system so that trial-to-trial variability would represent local differences in body motion relative to the pelvis. Following conditioning, each frame of a lift was represented by the three-dimensional coordinates of the fifteen anatomical locations, defining an m = 45 by n = 101 frames matrix. Frame specific matrices were concatenated into a 1x4545 length vector, to represent the time normalized trajectory data as a single row vector.

The OCSVM requires a training set to define the decision boundary, and consequently the baseline lifting movement pattern. As a result, each participant's lifting data were separated into a training set and multiple test sets (presumed to represent motion at increasing levels of fatigue). The training sets were sectioned to include about 35% of the total number of lifts completed by each respective participant. This percentage was chosen to provide enough data to train the model to appropriately define the baseline movement of a participant. There are no examples to draw from the literature that use OCSVMs for classifying fatigued motion data to draw a specific conclusion about the size of the training set that should be used. However, Clermont, Benson, et al. (2019) used kilometers 4-14 from a marathon, or approximately 23.7% of the total marathon, to use as baseline "typical" running motion data in their study. For this study, it is important to try not to include lifts that may be when a person is starting to fatigue and possibly adapting away from their baseline lifting movement pattern. However, since each individual lift cannot be specifically labelled as either “fatigued” or “non-fatigued”, an assumption was made that approximately the first 35% of the lifting data was “non-fatigued.” Table 2 shows statistics about the lifts completed by each participant.

Table 2. Descriptive statistics for the number of lifts completed by each participant, the number of lifts used for analysis after omitting lifts that were not properly captured due to human or technological error, the number of training set lifts used in the OCSVM, and the percentage of the total lifts that were used in the training set.

Participant	# Lifts	# Lifts used	# Training Lifts	% Lifts used for Training
P1	465	455	157	34.5
P2	333	326	120	36.8
P3	288	285	105	36.8
P4	510	504	180	35.7
P5	618	575	204	35.5
P6	540	538	180	33.5
P7	495	490	165	33.7
P8	255	255	90	35.3
P9	489	489	165	33.7
P10	510	494	177	35.8
P11	420	420	150	35.7
P12	495	494	164	33.2
P13	345	345	120	34.8
P14	450	450	150	33.3
Group Mean ± SD	443.8±103.3	437.1±97.5	151.9±32.2	34.9±1.3

The remaining lifts were then separated into multiple test sets. Since HR and RPE were collected after every 15 lifts, a test set was sectioned to include one HR and one RPE measurement, and 15 lifts. Some test sets contained more or less than 15 lifts depending on collection errors that led to lifts being omitted, or errors with collecting the HR and RPE from the participant at the appropriate times (i.e. if HR and RPE was collected three lifts too late, 18 lifts would be included in that test set instead of 15). The separation of the data into training and test sets can be seen in Figure 10.

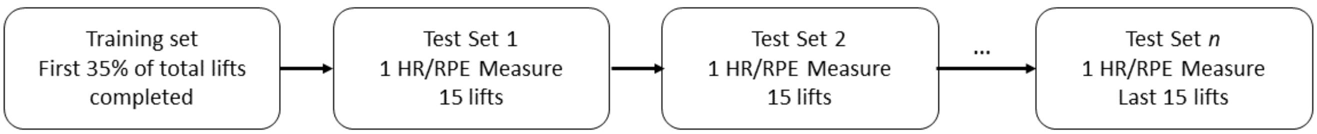

Figure 10. Separation of total lifts completed by a participant into one training and multiple test sets, where *n* represents the number of test sets for one individual. The progression can also be thought of as the start of the protocol (first lifts in training set) to the end of the protocol (last lifts in test set *n*).

4.5.2 Participant Fatigue Likeliness Classification

Participants were classified as either likely fatigued or unlikely fatigued to aid in analysis as a correlation between HR or RPE and outliers in those that did not fatigue would not be anticipated. Figure 11 shows a decision tree that led to their classification. If participants did not finish the full hour-long protocol due to volitional fatigue, they were labelled as likely to be fatigued. The remaining participants had their median RPE and HR for the training set and test sets calculated (i.e. a median RPE difference of two means their median RPE during the test sets was two higher than their median RPE during the training set). If participants that finished the hour-long protocol had a median RPE difference of two or greater, or a median HR difference of seven beats per minute (BPM) or greater, they were labelled as likely to be fatigued. If participants did finish the hour-long protocol but had a median RPE difference of less than two, or a median HR difference of less than seven BPM, they were labelled as unlikely to be fatigued. Participants were labelled twice for both RPE and HR. For example, a participant could have finished the hour-long protocol, had a median RPE difference of three and a median HR difference of five BPM. In this case, they would be labelled as likely to be fatigued based on the RPE data, and unlikely to be fatigued based on the HR data.

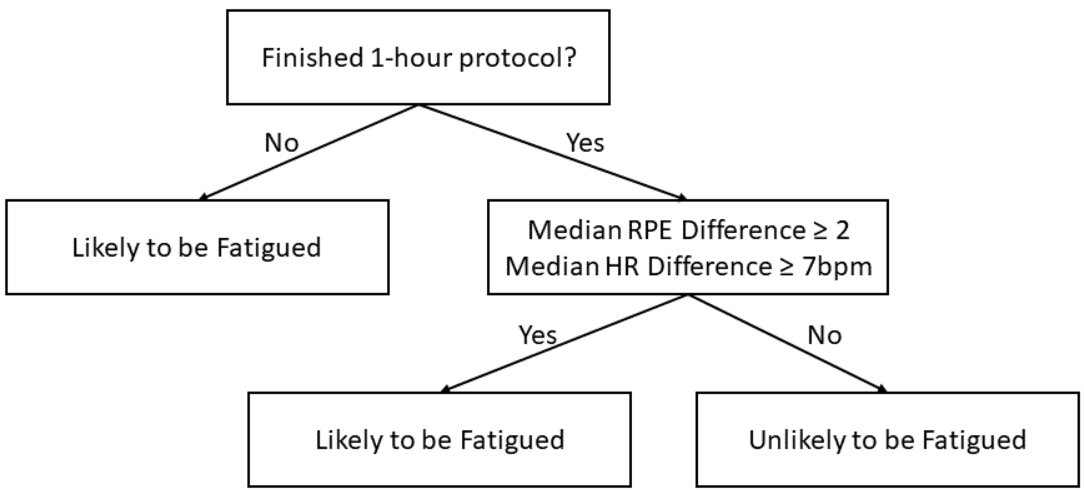

Figure 11. Decision tree leading to each participants' fatigue likeliness classification.

4.5.3 Feature Extraction using Principal Component Analysis

PCA was used as a feature extraction and data reduction method to identify features of movement to capture orthogonal modes of variability (Remedios et al., 2020). Each PC calculated from PCA describes a specific feature of movement (Armstrong et al., 2020). A 95% trace criterion was used, where PCs were retained that explained up to at least 95% of the overall variance in the data (Armstrong et al., 2020; Brandon et al., 2013; Sadler et al., 2011). PCA was first used on only the training set of lifts. The PC scores retained from the training set PCA were used as features to define the overall baseline lifting pattern for each participant. Since this is a subject-specific analysis, 14 PCAs, one unique to each participant, were conducted. The motion data from each test set was then projected back onto the PC feature space derived from the training set to generate PC scores for each test set using the *transform* function from the PCA toolkit. The PCA tool was used from the Scikit learn library in Python. PC scores from the training set and each test set were then used in the OCSVM model.

4.5.4 OCSVM Classification

As mentioned in the introduction, the OCSVM defines a decision boundary from “typical” data and classifies new data as fitting either within or outside of that decision boundary (Kobsar & Ferber, 2018), serving as an outlier detection algorithm. The OCSVM tool from scikit was used and a Gaussian kernel function was used to train the boundary.

The OCSVM has two hyperparameters that can be altered to modify how the decision boundary is made. *Gamma* is a hyperparameter of the Gaussian kernel, controlling the influence of individual training examples (set as a number > 0). Essentially, *gamma* effects the ‘smoothness’ of the model, where a higher value will lead to a tighter fitted model, while a lower value will lead to a more generalizable model. The default value, *scaled*, was used for each model, and calculated using Equation 1. Matrix X refers to the matrix being input into the OCSVM. Figure 12 shows how the *gamma* value influences the decision boundary.

The other hyperparameter, *nu* (set between 0 and 1), is defined as the upper bound on the fraction of training errors and a lower bound on the fraction of support vectors. In other words, *nu* represents the proportion of outliers that is expected in the data. For this model, *nu* was set to $nu = 0.01$. Figure 13 shows how the value of *nu* influences the decision boundary. These two hyperparameters had to be set so the training set boundary included a high percentage of the training set lifts as inliers so that the overall lifting movement pattern was defined, but not improperly set to make the decision boundary too large, where outliers would not be found in the test set data.

Equation 1. Calculating the hyperparameter value *gamma*. The number of features refers to the number of PCs retained to use in each respective OCSVM, and matrix X is the matrix of PC scores defining each participant's movement pattern.

$$1 \div (\# \ of \ features * variance \ of \ matrix \ X)$$

The PC scores in both the training set and test set were standardized to the mean and standard deviation of the training set using the *StandardScaler* function from Scikit learn. The PC scores from the training set were then used to fit a OCSVM decision boundary to the data. The percentage of outliers in the training set data were then calculated using Equation 2. The test set PC scores were then tested against this decision boundary using the *predict* function from the OCSVM tool in Scikit learn. The number and percentage of outlier lifts were calculated for each test set. The percentage of outliers was calculated using Equation 2.

Equation 2. Calculating the percentage of outliers in the training and test sets.

$$\left(\frac{\# \ outlier \ lifts \ in \ training/test \ set}{Total \ \# \ lifts \ in \ training/test \ set}\right) \text{ x } 100$$

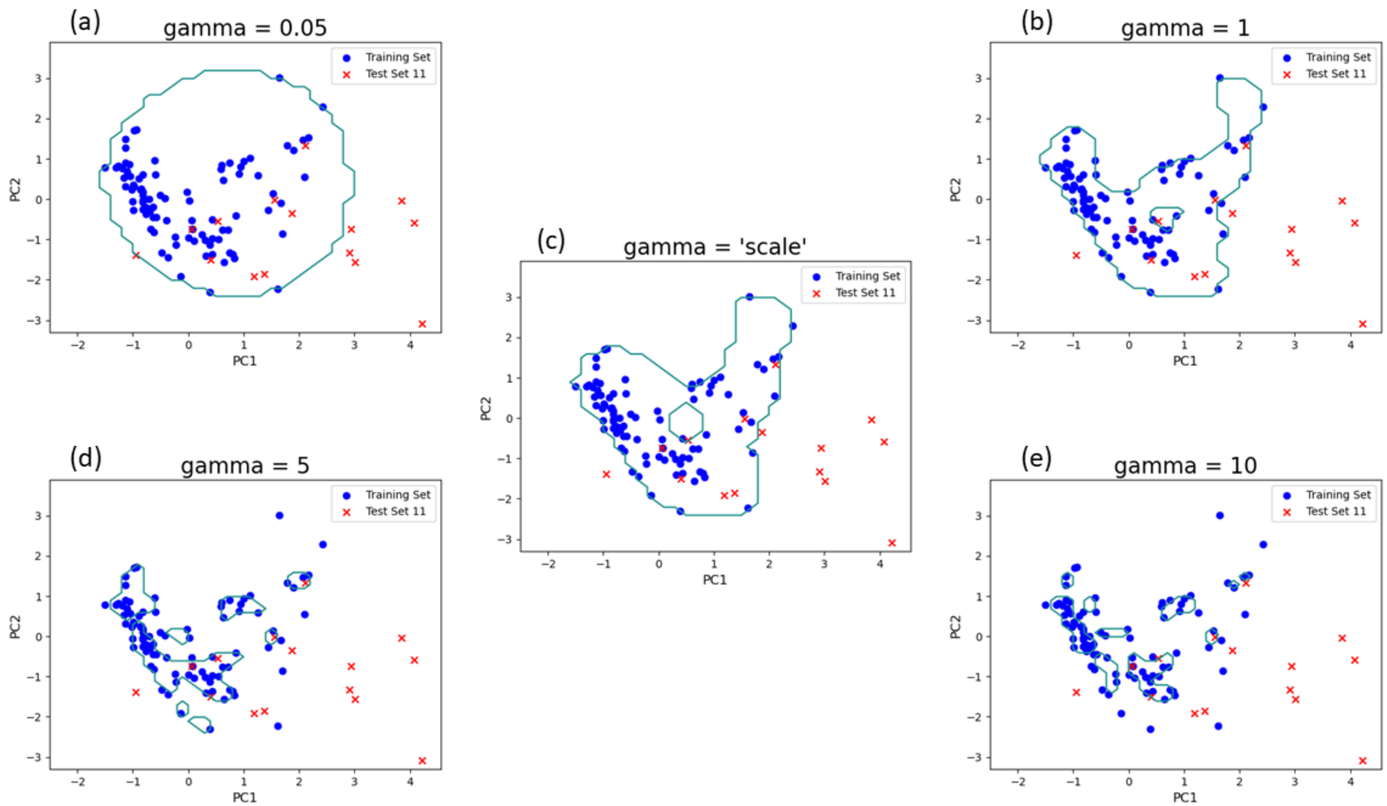

Figure 12. Influence of the hyperparameter gamma on the smoothness or generalizability of the decision boundary. A lower value, such as 0.05, will result in a wide, more generalizable model (a). A higher value, such as 10 (e), results in a tighter fit model. The default value 'scale' was used for the models in this study (c). Equation 1 shows how this value was calculated. Two PCs (features) were used for visualization purposes.

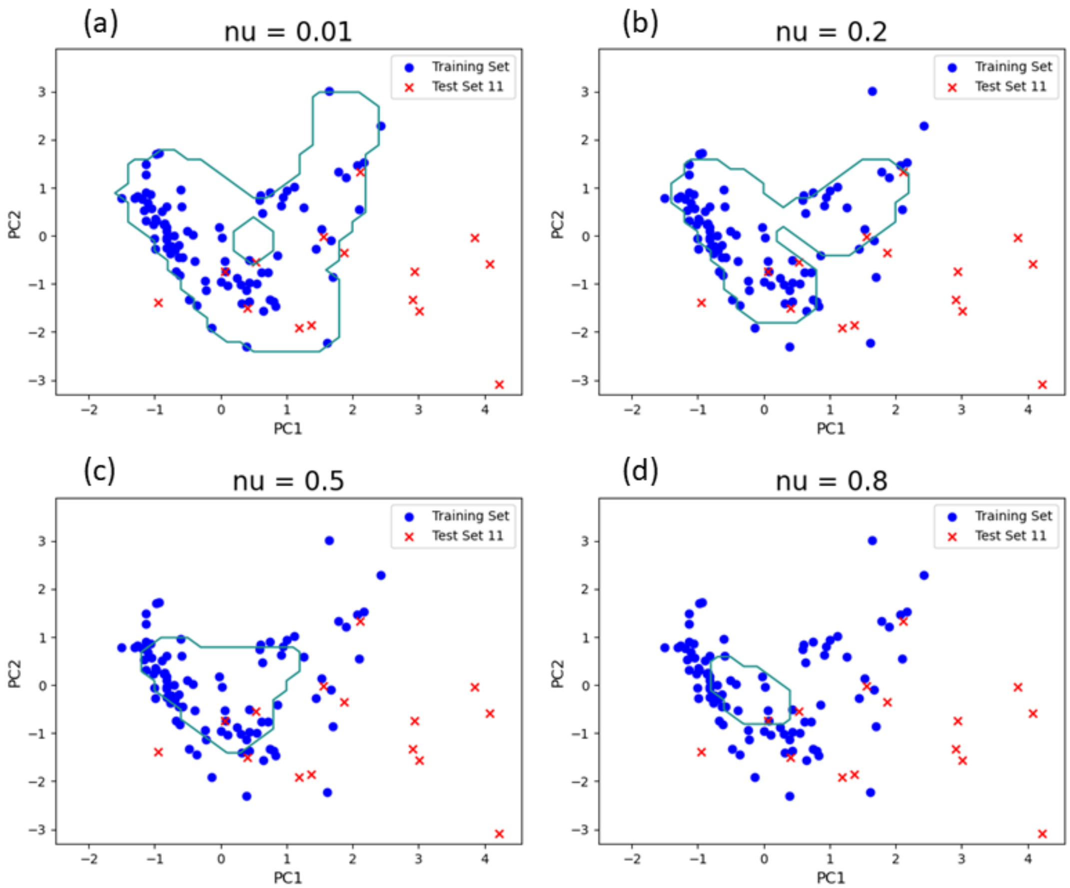

Figure 13. Influence of the hyperparameter nu on the amount of training errors the model includes. A lower value, such as 0.01, will have less training data outliers (a). A higher value, such as 0.8, will have more training data outliers (d). For example, a value of 0.2 would expect to include about 20% of the training data as outliers in the model. A nu value of 0.01 was used for the models in this study. Two PCs (features) were used for visualization purposes.

4.6 Statistical Analysis

To test the hypotheses, the recorded HR, RPE and percentage of outliers for each test set was tested for association. Spearman's rank-order correlation (ρ), a non-parametric test, was used to assess the association between the percentage of outliers, and both RPE and HR separately. Spearman's rank-order correlations were applied to each individual participant's data. Each individuals' data were combined and assessed for associations to also explore the generalizability of the findings across the sample. A correlation of 0.10-0.29, 0.30-0.49 and 0.5+ was interpreted as small, medium and large, respectively (Cohen, 1988). A p-value of < 0.05 indicated significance. The *spearmanr* function from SciPy was used in Python for the statistical testing. Descriptive statistics were also calculated for the recorded HR, RPE and percentage of outlier measures.

5. Results

5.1 Fatigue Likeliness Classification

The descriptive statistics of each participant's RPE measurements, their total time spent completing the repetitive lifting protocol and the fatigue likeliness classification are presented in Table 3. There were three participants that were classified as likely to be fatigued due to not finishing the hour-long protocol from volitional fatigue. Seven participants were classified as likely to be fatigued due to having a difference in RPE medians between the test and training sets of two or greater. Four participants were classified as unlikely to be fatigued due to not having a difference of two or greater.

The descriptive statistics of each participant's HR measurements, their total time spent completing the repetitive lifting protocol and the fatigue likeliness classification are presented in Table 4. There were three participants that were classified as likely to be fatigued due to not finishing the hour-long protocol. Five participants were classified as likely to be fatigued due to having a difference in HR medians between the test and training sets of seven BPM or greater. Six participants were classified as unlikely to be fatigued due to having a difference in median HR of less than seven.

Two of the fourteen participants had a different classification of fatigue likeliness between RPE and HR. Both P4 and P12 were classified as likely fatigued based on their RPE median difference but classified as unlikely fatigued based on their HR median difference.

Table 3. Descriptive statistics for each participant's RPE measurements for the training and test sets, their total time spent completing the repetitive lifting protocol, and the likelihood of fatigue classification based on the decision tree in Figure 11. Participants are ordered by their likelihood of fatigue classification.

Participant	Training RPE Median	Training RPE Range (Min:Max)	Test RPE Median	Test RPE Range (Min:Max)	Difference in Medians (Test - Training)	Total Time, 1hr max (hh:mm)	Likelihood of Fatigue
P2	15.5	14:16	18	17:19	2.5	0:38	Fatigue likely - volitional fatigue
P3	16	15:17	18	17:20	2.0	0:35	Fatigue likely - volitional fatigue
P8	11.5	9:14	18	14:20	6.5	0:27	Fatigue likely - volitional fatigue
P1	11	9:13	15	13:17	4.0	1:01	Fatigue likely - increased RPE
P4	13	12:14	15	14:16	2.0	1:02	Fatigue likely - increased RPE
P6	14	11:15	16	15:18	2.0	1:03	Fatigue likely - increased RPE
P10	13.5	12:15	17	15:18	3.5	1:05	Fatigue likely - increased RPE
P11	13	12:15	15	14:15	2.0	1:03	Fatigue likely - increased RPE
P12	13	12:14	16	14:18	3.0	1:01	Fatigue likely - increased RPE
P13	12.5	9:14	15	14:20	2.5	0:57	Fatigue likely - increased RPE
P5	17	13:17	18	17:19	1.0	0:59	Fatigue unlikely - small RPE change
P7	16	15:16	17	16:17	1.0	1:04	Fatigue unlikely - small RPE change
P9	9	9	9	9	0.0	1:01	Fatigue unlikely - no RPE change
P14	13	13	13	13	0.0	1:03	Fatigue unlikely - no RPE change

Table 4. Descriptive statistics for each participant's HR measurements for the training and test sets, their total time spent completing the repetitive lifting protocol, and the likelihood of fatigue classification based on the decision tree in Figure 11. Participants are ordered by their likelihood of fatigue classification.

Participant	Training HR Median (BPM)	Training HR Range (Min:Max) (BPM)	Test HR Median (BPM)	Test HR Range (Min:Max) (BPM)	Difference in Medians (Test - Training)	Total Time, 1hr max (hh:mm)	Likelihood of Fatigue
P2	148.5	129:156	147	143:149	-1.5	0:38	Fatigue likely - volitional fatigue
P3	138	131:145	146.5	135:157	8.5	0:35	Fatigue likely - volitional fatigue
P8	139	125:142	154	137:159	15.0	0:27	Fatigue likely - volitional fatigue
P1	93	89:100	104.5	92:111	11.5	1:01	Fatigue likely - increased HR
P6	133	122:138	141	131:148	8.0	1:03	Fatigue likely - increased HR
P10	174.5	150:180	182	179:187	7.5	1:05	Fatigue likely - increased HR
P11	133	116:142	145.5	138:157	12.5	1:03	Fatigue likely - increased HR
P13	131.5	124:146	147	135:154	15.5	0:57	Fatigue likely - increased HR
P4	129	119:134	134	127:139	5.0	1:02	Fatigue unlikely - small increase in HR
P5	141.5	139:146	136	124:146	-5.5	0:59	Fatigue unlikely - decreased HR
P7	106	103:110	112.5	106:118	6.5	1:04	Fatigue unlikely - small increase in HR
P9	107	102:112	106.5	102:111	-0.5	1:01	Fatigue unlikely - decreased HR
P12	130	124:139	133.5	131:138	3.5	1:01	Fatigue unlikely - small increase in HR
P14	114	112:118	116.5	112:123	2.5	1:03	Fatigue unlikely - small increase in HR

5.2 Feature Extraction Results

There was an average of 20.4 ± 8.8 PCs retained from the PCAs (see Table 5). The first PC explained an average of 56.5 ± 18.1% variance in each participants' baseline training set lifts. The PC scores for these PCs were input as features into each participants' respective OCSVM model to define their baseline lifting movement pattern.

Table 5. PCA results. The number of PCs that were retained, and the percentage of variance that the first PC explains from the training set data.

Participant	# PCs Retained	First PC % Explained
P1	17	69.8
P2	39	28.1
P3	21	33.1
P4	28	36.9
P5	17	68.3
P6	13	62.1
P7	27	42.0
P8	18	54.2
P9	10	81.2
P10	12	76.5
P11	32	38.0
P12	20	59.6
P13	23	60.9
P14	8	80.8
Group Mean ± SD	20.4 ± 8.8	56.25± 18.1

5.3 OCSVM Classification Results

After using the training set lifts to define the OCSVM, the same lifts were classified against the decision boundary to get a measure of how many training lifts were outliers. There was an average of 16.0 ± 2.9% of training lifts that were classified as outliers from the training set decision boundary (see Table 6).

The results of the test set classification for each participant is also shown in Table 6. Participants have been separated by their likeliness of fatigue classifications that was based on RPE. Seven of the ten participants who were likely fatigued have a higher test set median outlier percentage than those participants who were likely not fatigued. Line graphs of the percentage of outliers for each test set for each participant are presented in Figure 14 and 15, where each plot separates participants based on their likeliness of fatigue. Most participants who finished with a higher percentage of outliers were classified as likely fatigued based on their RPE (see Figure 14), but less so based on HR (see Figure 15).

Table 6. Training set and test set OCSVM classification results. Included for each participant are the percentage of training lifts that were classified as outliers based on the boundary definition resulting from the training set, the number of test sets, the test set outliers percentage median, and the test set outliers percentage range. Participants are separated by their likeliness of fatigue classification from RPE.

Participant	Training Set Outliers (%)	# Test Sets	Test Set Outliers Median (%)	Test Set Outliers Range (%) (Min:Max)
Participants likely to have fatigued (based on RPE)				
P1	12	20	14	7:40
P2	22	14	90	50:100
P3	18	12	67	47:93
P4	12	22	41	27:73
P6	14	23	100	80:100
P8	19	11	80	47:100
P10	15	21	53	7:100
P11	16	18	80	27:100
P12	17	22	87	27:100
P13	19	15	87	60:100
Participants unlikely to have fatigued (based on RPE)				
P5	16	26	47	21:33
P7	17	22	53	21:27
P9	14	22	33	7:80
P14	13	20	40	13:73
Group Mean ± SD	16.0 ± 2.9	19.1 ± 4.5	---	---

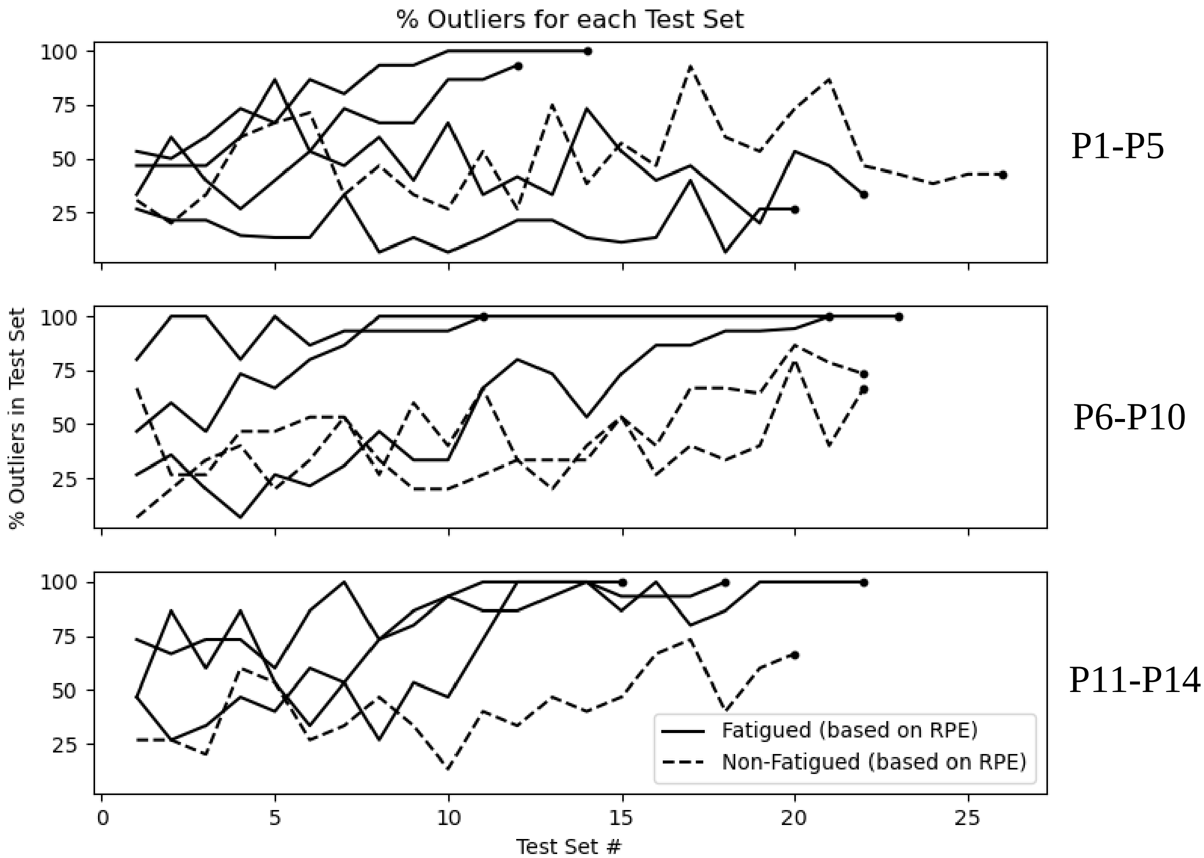

Figure 14. Each participants' percentage of outliers graphed over the increasing test set number. The left side of the graph represents the first test set after the training set and moving to the right indicates an increasing test set number (increasing time spent completing the repetitive lifting protocol). The subplots are used to improve the ease of visualization. Solid lines are participants who were likely fatigued based on RPE, and dashed lines are participants who were not likely fatigued based on RPE.

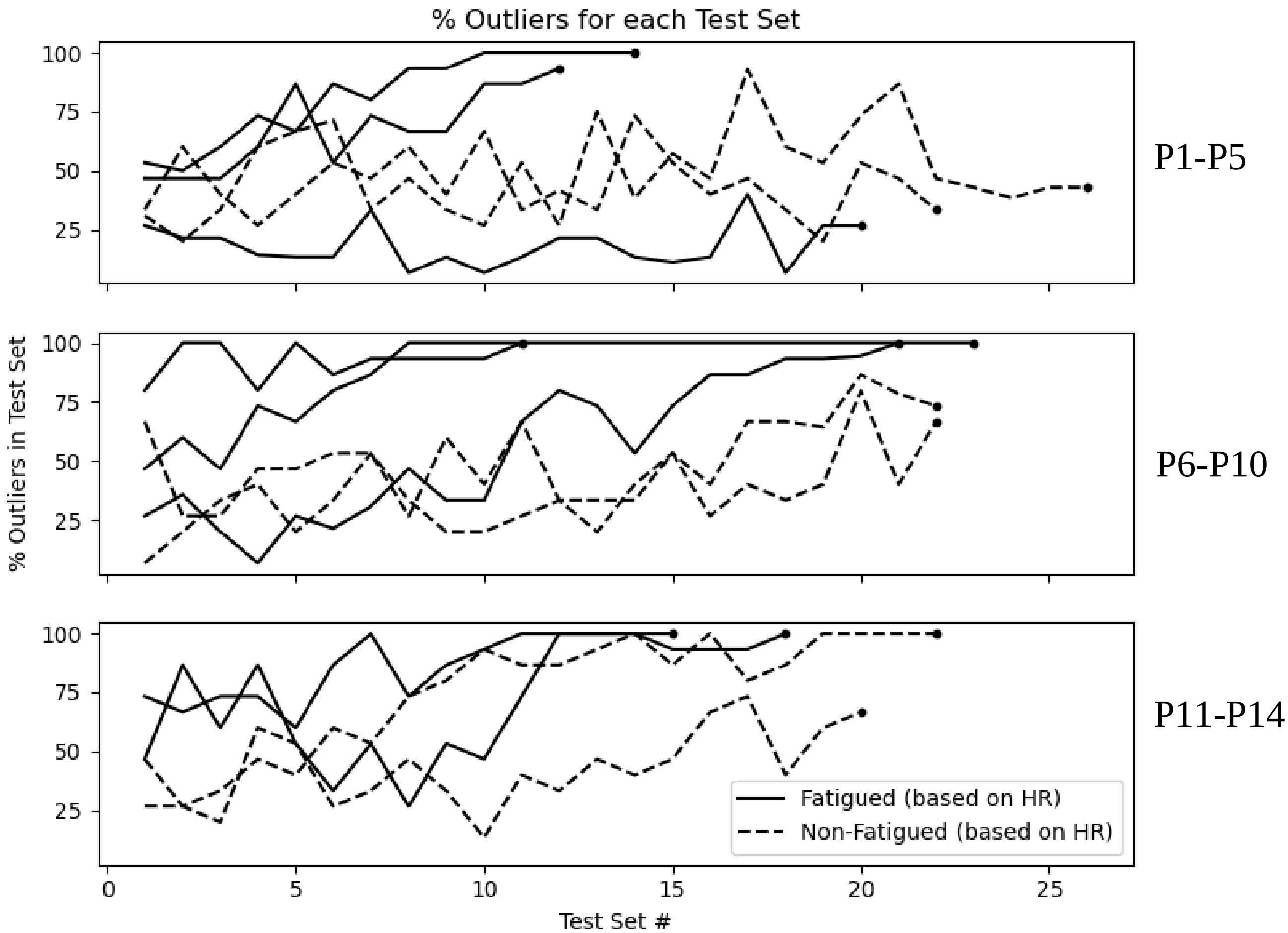

Figure 15. Each participants' percentage of outliers graphed over the increasing test set number. The left side of the graph represents the first test set after the training set and moving to the right indicates an increasing test set number (increasing time spent completing the repetitive lifting protocol). The subplots are used to improve the ease of visualization. Solid lines are participants who were likely fatigued based on HR, and dashed lines are participants who were not likely fatigued based on HR.

5.4 Assessing Association Between Fatigue and Changes in Movement

5.4.1 RPE Spearman Results

Seven of the ten participants who were likely fatigued had a significant large positive association between RPE and the percentage of outliers, while one participant had a significant large negative association (see Table 7). All four participants that were labelled as unlikely to be fatigued had no significant association. A scatterplot of all the participants' data, as well as lines of best fit, is shown in Figure 16.

Table 7. Spearman rank order correlation results for RPE and the percentage of outliers in the test sets. (*) indicates significance. Participants are ordered by their RPE fatigue likeliness classification. Note that P9 and P14 gave RPE values that did not change over the course of the protocol, therefore no association can be calculated.

Participant	RPE rho	RPE p-value	Likeliness of Fatigue
P1	-0.0527	0.825	Fatigue likely - increased RPE
P2	0.831	<0.001*	Fatigue likely - volitional fatigue
P3	0.775	0.00306*	Fatigue likely - volitional fatigue
P4	-0.0925	0.682	Fatigue likely - increased RPE
P6	0.644	<0.001*	Fatigue likely - increased RPE
P8	0.936	<0.001*	Fatigue likely - volitional fatigue
P10	0.880	<0.001*	Fatigue likely - increased RPE
P11	-0.585	0.0107*	Fatigue likely - increased RPE
P12	0.831	<0.001*	Fatigue likely - increased RPE
P13	0.816	<0.001*	Fatigue likely - increased RPE
P5	0.195	0.340	Fatigue unlikely - small RPE change
P7	0.337	0.125	Fatigue unlikely - small RPE change
P9	N/A	N/A	Fatigue unlikely - no RPE change
P14	N/A	N/A	Fatigue unlikely - no RPE change

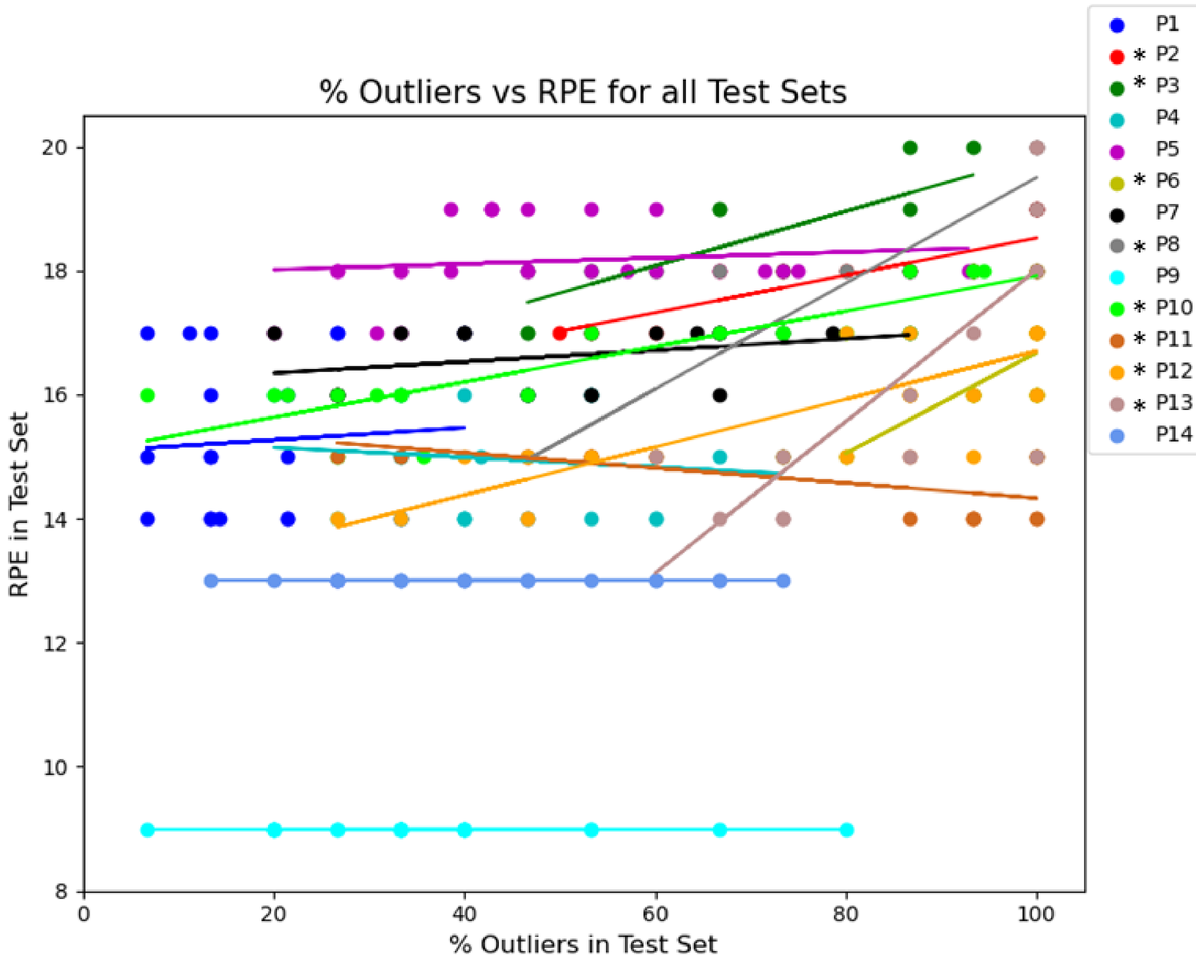

Figure 16. Scatterplot showing each participants' RPE and percentage of outliers for each test set. Each participant is represented by a different color, and participants with a significant association are denoted in the legend with a (*). Lines of best fit are included for ease of visualization. Note that some data points may not be shown due to overlap where some participants may share the same data point values.

5.4.2 HR Spearman Results

Two participants had a significant large positive association between the percentage of outliers and HR during the test sets (see Table 8). Nine participants had either a small or medium negative, non-significant association, indicating HR and the percentage of outliers tended towards a negative association at the individual level. Of the eight participants that were labelled as likely to be fatigued based on HR, two had a significant positive association. All six participants that were labelled as unlikely to be fatigued had no significant association. A scatterplot of all the participants' data, as well as lines of best fit, is shown in Figure 17.

Table 8. Spearman rank order correlation results for HR and the percentage of outliers in the test sets. (*) indicates significance. Participants are ordered by their HR fatigue likeliness classification.

Participant	HR rho	HR p-value	Likeliness of Fatigue
P1	-0.236	0.317	Fatigue likely - increased HR
P2	-0.0412	0.889	Fatigue likely - volitional fatigue
P3	-0.404	0.192	Fatigue likely - volitional fatigue
P6	-0.163	0.457	Fatigue likely - increased HR
P8	0.633	0.0366*	Fatigue likely - volitional fatigue
P10	0.509	0.0186*	Fatigue likely - increased HR
P11	-0.376	0.124	Fatigue likely - increased HR
P13	0.0375	0.894	Fatigue likely - increased HR
P4	0.107	0.636	Fatigue unlikely - small increase in HR
P5	-0.383	0.0531	Fatigue unlikely - decreased HR
P7	0.397	0.0673	Fatigue unlikely - small increase in HR
P9	-0.259	0.245	Fatigue unlikely - decreased HR
P12	-0.310	0.161	Fatigue unlikely - small increase in HR
P14	-0.334	0.150	Fatigue unlikely - small increase in HR

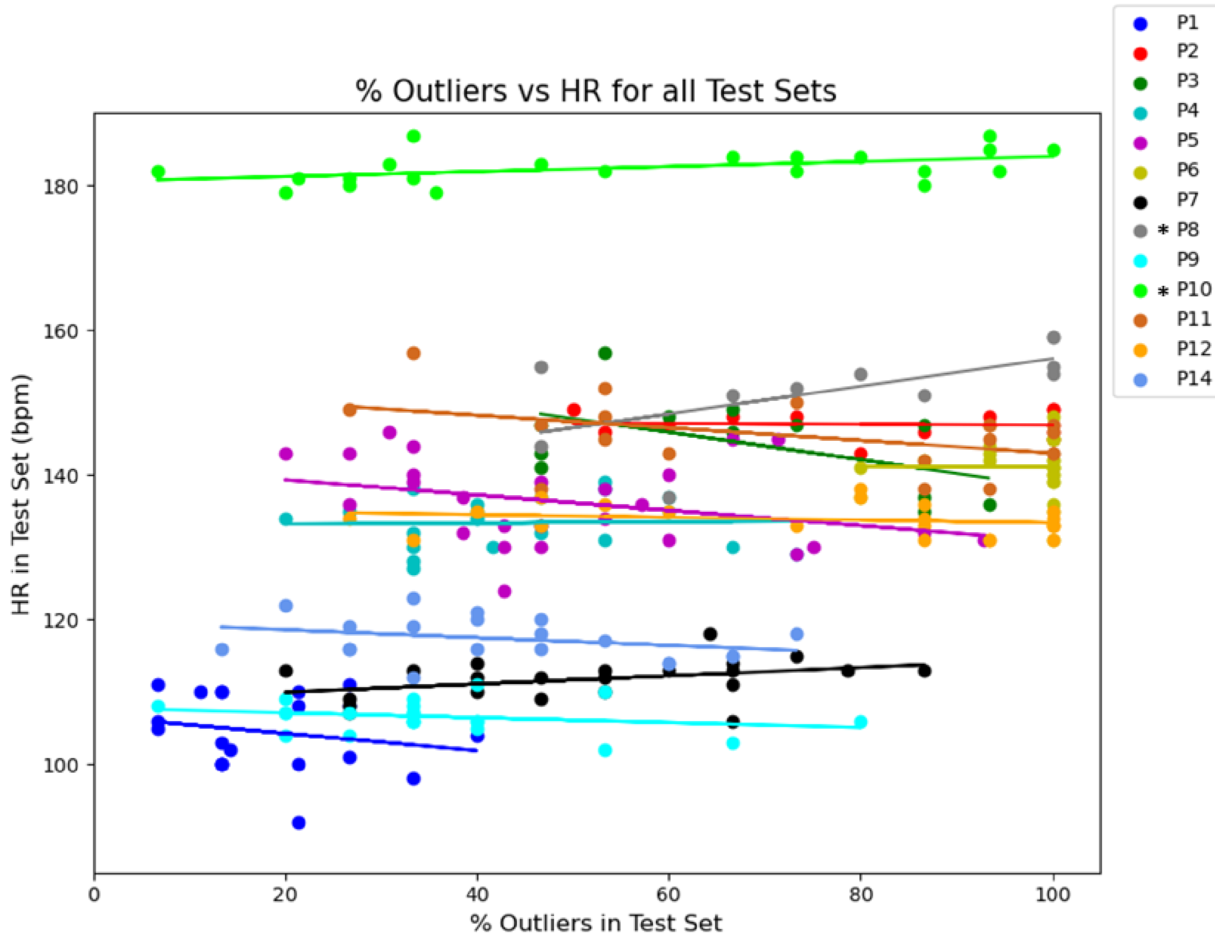

Figure 17. Scatterplot showing each participants' HR and percentage of outliers for each test set. Each participant is represented by a different color, and participants with a significant association are denoted in the legend with a (*). Lines of best fit are included for ease of visualization. Note that some data points may not be shown due to overlap where some participants may share the same data point values.

6. Discussion

The use of OCSVM as an outlier detection algorithm for identifying fatigue related changes in whole-body movement patterns over the course of a repetitive lifting protocol was investigated in this thesis. Based on the results of the fatigue likeliness classification in Tables 3 and 4, not every participant fatigued to the same degree from the protocol. Most of the participants (except for two) were labelled the same when classified based on either their RPE or HR measurements, indicating consistency between the two measures for identifying potential fatigue. Significant positive associations were found between RPE and the percentage of outliers in those participants that were likely fatigued from the repetitive lifting protocol. However, participants exhibited a more negative trend between HR and the percentage of outliers, and less significant positive associations in those that were likely fatigued. Overall, the results support the first hypothesis that there would be an association between RPE and the percentage of outliers in those who were likely fatigued. However, the second hypothesis is not supported that there would be an association between HR and percentage of outliers in those who were likely fatigued.

6.1 Movement Pattern Changes When Fatigued

6.1.1 RPE Outcomes

Most participants (seven out of ten) labelled as likely fatigued had significant positive associations between the percentage of outliers and RPE scores, partially supporting the first hypothesis. No significant associations between the percentage of outliers and RPE were found in participants who were unlikely to be fatigued, which would also be a logical outcome given the stated hypothesis. These results suggest that those participants who were more fatigued from the repetitive lifting task were indeed more likely to change their movement patterns relative to their baseline, and those who were not fatigued were less likely to change their movement patterns relative to baseline. These results are similar to past research, where participants who fatigued during a repetitive lifting protocol (measured as a significant decrease in instantaneous median frequency in their electromyography signal) also had significant changes in their knee and elbow ranges of motion (Bonato et al., 2002). However, participants that did not have a significant decrease in the instantaneous median frequency in their electromyography signals also did not have significant changes in their lifting kinematics. Although the length and type of lifting protocols were different to the current study, the results of Bonato et al. (2002) help explain and reinforce why fewer outliers were detected in those that were not likely fatigued from the prolonged lifting protocol, relative to those that likely fatigued.

Repetitive lifting literature indicates that fatigue related kinematic changes occur (Bonato et al., 2003; Fischer et al., 2015; Mehta et al., 2014; Sparto et al., 1997). However, inconsistencies in the fatigue related kinematic changes that are observed during repetitive lifting exist. For example, Bonato et al. (2003) found that participants who started with a stoop style lift switched to more of a squat style lift, seen as increased hip and trunk range of motion, but those who started with a squat style did not change their lifting technique. In contrast, a decrease in knee and hip range of motion was seen at the end of a repetitive lifting protocol compared to the start, suggesting a switch to a stoop style lift (Sparto et al., 1997).

The differences in past study results suggest that fatigue related kinematic changes may be subject-specific. It may be possible that kinematic differences exist from the studies due to differences in protocols or weights used, but the overall task was similar. The inherent variability of movement, where individuals can select movement pattern strategies to complete a task on their own accord, may not make measuring group level fatigue related changes in movement patterns useful. Due to these inconsistencies in movement pattern changes, the use of a subject-specific pattern recognition feature extraction method such as PCA was justified in this study.

Circling back to the main issue with FCEs guiding this thesis is the subjective and vague biomechanical criteria that evaluators use to determine test endpoints during simulated work tasks. Evaluators currently have pre-determined biomechanical criteria that they observe, but individuals may not respond homogenously with respect to those pre-determined criteria. Fatigue related changes across individuals in lifting are inconsistent as shown by the literature, showing the need for a subject-specific approach. The results from this thesis show support for the use of the outlier detection approach, where lifting pattern outliers were classified based on a comparison to their own initial lifting patterns in individuals who were likely fatigued based on their RPE. It is not known the magnitude of differences in movement patterns used across participants, but there was success with the subject-specific approach. The use of an outlier detection model may be considered for use in clinical applications, such as FCEs, for possible identification of heterogenous variability in movement patterns and adaptations to fatigue among individual patients.

6.1.2 HR Outcomes

In contrast to the RPE results and the first hypothesis, the second hypothesis that those who were likely fatigued would exhibit a significant positive association between the percentage of outliers and HR during the test sets was not supported. Only two of the eight participants that were likely fatigued had significant positive associations between the percentage of outlier lifts and their increase in HR. Similar to RPE, no significant associations were found in participants who had no difference or a decrease in HR. Nine participants actually had a negative non-significant correlation where HR decreased as the percentage of outliers increased. These results suggest that HR may not be a suitable measure of fatigue during a repetitive lifting protocol. It is likely that HR can be used as a measure of how hard the body is working (more intensity would have a higher heart rate), but HR increases may or may not be associated with overall fatigue when considering steady-state prolonged work.

Past literature has recorded HR measurements during repetitive lifting protocols. In a similar study that had participants undergo an hour-long repetitive lifting protocol, no significant increase in HR were observed when the work rate was below 40% of their VO_2 maximum for all different weights used (Petrofsky & Lind, 1978). Because of this, it is possible that the weight and rate of lifting used by participants in the current study may not have been enough to cause an increase in heart rate, suggesting they were working below the 40% of their VO_2 maximum. Only eight of the fourteen participants were likely fatigued based on their HR, and only two had significant associations between the percentage of outliers and HR. A higher load or work rate may have led to more increases in HR. Also, in a study of 500 participants undergoing an FCE progressive loading lifting test, there was an average increase in HR of 18.1% from resting to peak (85.7 to 99.8 bpm) in a sub-maximal floor to waist lift, and an average of 16.8% (86.4 to 100.5 bpm) increase in a sub-maximal waist to shoulder height lift (Morgan et al., 2012). Since HRs only went as high as 100bpm on average, this suggests that a submaximal progressive lifting test, although different from the submaximal repetitive lifting test used in this thesis, may not produce large increases in HR. As seen in the this thesis and the above studies, there will undoubtedly be increases in HR when a person is doing sub-maximal work when compared to their resting HR. However, these increases may not be correlated with fatigue that would result in movement pattern changes, but may be due to overall increase in workload compared to rest, where the cardiovascular system must supply the needs to the working muscles appropriately.

Another possible explanation for the difference between the RPE and HR findings in the current study is that HR can vary much more over the course of a prolonged lifting protocol. If the participant ever stopped to take a water break, or had time in between sets to give their RPE and HR measurement, the HR may have dropped quickly due to the stoppage of work. RPE is more subjective and may be more susceptible to the participant's overall physical and mental fatigue. Most participants saw episodic rises and falls in their HR data during the protocol but maintained an overall steady state. Conversely, RPE consistently increased for most participants during the protocol without reporting a decrease as they progressed. HR will increase in response to an increased workload to meet the working muscle's oxygen requirements but may not increase further when there is localized muscular fatigue. However, RPE may be more sensitive to fatigue, where participants will be aware of the feeling of their fatiguing muscles, contributing to their perception of their subjective workload and fatigue. RPE was therefore more likely to see a positive significant association with the percentage of outliers if a participant's movement pattern was in fact changing from fatigue as they neared the end of the protocol. The use of HR in FCEs for physiological monitoring is common, but its use may be focused more on safety (e.g. making sure participants stay under 85% of their maximum HR) and tracking effort level (Allison et al., 2018), rather than as a specific measure of fatigue. A reliable method of detecting fatigue may be needed in order to appropriately assess fatigue related changes in movement patterns during an FCE using an outlier detection approach.

6.2 Movement Pattern Changes When Not Fatigued

Although a portion of the participants were likely not fatigued from the repetitive lifting protocol based on their RPE and HR measurements, some still exhibited an increase in their percentage of outlier lifts during the test sets (see Figures 14 and 15). For example, the OCSVM boundary for P9 was determined such that 14% of the training lifts were considered outliers. In the first test set, 7% of lifts were identified as outliers. In the third from last test set, 80% of the lifts were outliers. However, P9 gave a consistent RPE rating of 9 throughout the entire protocol, did not have a difference in median HR greater than seven between training and test sets, and completed the entire 60-minute protocol suggesting they were not fatigued from the protocol. These results indicate that although they were not fatigued on the basis of the criteria used to define fatigue in this study, they were still changing their movement pattern compared to their baseline.

There may be different reasons for these findings that should be considered. First, the nature of the repetitive lifting protocol and the large number of degrees of freedom in whole-body movement allows participants the freedom to vary their movement from one lift to the next. When considering the amount of variability exhibited at the start versus the end of a repetitive fatiguing protocol, there is evidence that variability can both increase and decrease, depending on the movement variable examined (N. Cortes et al., 2014). In the context of lifting, it was found that there were significant increases in variability of whole-body centre of mass kinematics during a repetitive lifting and lowering task as participants started to fatigue (Sedighi & Nussbaum, 2017). Increasing movement variability may help to limit fatigue development, or relieve loading on tissues that may be fatiguing during repetitive movements (Srinivasan & Mathiassen, 2012). Combining all of this information, it may be possible that participants such as P9 have more variable movement in the later stages of a prolonged lifting protocol in order to prevent possible fatigue and injury, all while maintaining their ability to perform the task consistently.

A secondary explanation is that HR and RPE may not be reliable fatigue measure for all participants. The subjectivity of RPE may limit its potential to reliably measure a participant's whole-body fatigue state. Although P9 gave a consistent RPE measurement that did not change during the hour-long lifting protocol, it might be hard to argue that they did not fatigue at all. In the original data collection that this data was used from, participants were not specifically told to use RPE as a measure of their fatigue. Instead, it was presented as a measure of exertion and fatigue. Therefore, differences in interpretation could have led to some participants not considering how tired they felt, but rather how hard they were working.

A more objective and sensitive measure of whole-body fatigue may have been able to identify, at the minimum, some changes in participants' fatigue states. A more dedicated subjective fatigue measure, such as a fatigue visual analog scale used by Chan et al. (2020), may provide a better self-reported measure of fatigue. Also, Chan et al. (2020) used an isometric lift strength assessment to measure the maximum tensile force the participant could exert on a load cell. Similarly, a 31% reduction in lifting power has previously been used as fatigue criteria in a repetitive lifting protocol (Sparto et al., 1997). A whole-body isometric maximal exertion, such as a mid-thigh isometric pull, may be a more reliable objective measure of whole-body fatigue (Stone et al., 2019). Considering the first reason mentioned above that adaptations to movement strategy may occur to prevent potential fatigue or injury, and the possibility for fatigue to not appropriately be quantified, the results from participants like P9 can be supported. The capability to measure an association between the percentage of outlier lifts detected by a OCSVM and the level of fatigue exhibited by the participants may rely on the sensitivity of the fatigue measure, but inherent movement variability while completing repetitive tasks should also be considered. Although there were no significant associations with RPE or HR in participants who were likely not fatigued, the OCSVM was still able to identify outlier lifts. When considering outlier detection use in FCEs, importance should be placed on having a reliable measure of fatigue to make accurate conclusions about possible fatigue related movement pattern changes from baseline when completing a repetitive protocol. If a reliable association between fatigue and an increase in outliers is established, inherent movement variability may become less of a potential factor for outliers to be observed.

6.3 OCSVM Considerations

6.3.1 Feature Selection

The feature selection process is important for appropriately defining the overall movement pattern of an action such as lifting. In this case, PCA was used as a data reduction and feature extraction method based on pattern recognition. The OCSVM was able to detect when participants were adapting their lifting movement pattern relative to baseline but understanding how they were adapting is also important to know. Although PCA is useful for extracting PCs that explain the variance in the data, interpreting what these PCs represent biomechanically in motion data can be difficult. A method to interpret PCs is to add and subtract a scalar multiple of the particular PC to the overall group mean, known as single component reconstruction (Brandon et al., 2013). An example of a scalar multiple to use is ± 2, to represent roughly two standard deviations above or below the mean. This would give visualization of the representative extremes, or the 5^{th} (low) and 95^{th} (high) percentile, assuming the PC scores are normally distributed (Brandon et al., 2013). Upon visualization of the representative extremes compared to the group mean for a particular kinematic variable, inferences can be made about the movement pattern that particular PC represents. Since the success of this method relies on visual inspection of the waveforms, an untrained person without prior knowledge or expertise may not be able to properly interpret the meaning. If interpreting the PCs is not a priority, PCA can still be a useful feature selection method for the OCSVM approach used in this thesis.

Another potential problem of applying PCA in this approach is that the PCs retained as features may not be associated with fatigue. Single component reconstruction was applied to a small portion of the data from this thesis to get a sense of how one of the participants in this study was varying their movement. An example of a single component reconstruction of PC1 for P1 is shown in Figure 18. In short, the first PC representing the largest degree of variability of one of the participants seemed to be which foot they placed forward while lifting the box. This PC may or may not be associated with fatigue but was more likely just natural variance in their lifting pattern over the repetitive protocol. Without doing an analysis of each individual PC retained as a feature to assess its association with fatigue, the PCs input into the OCSVM may not be relevant for detecting fatigue related changes. Using pre-selected task-relevant variables (Chan et al., 2020; Clermont, Benson, et al., 2019) that are known to be associated with fatigue is an alternative option, but may ignore subject-specific differences. In the future, performance of outlier detection algorithms for detecting fatigue related changes in MMH tasks should be evaluated using pre-selected task-relevant features and compared to pattern recognition methods such as PCA.

PCA does have strengths however, where the whole-body motion is considered in the analysis. The large number of degrees of freedom of movement in the kinetic chain creates many possible movement combinations when completing a task such as lifting (Scott, 2004). Selecting pre-defined variables may lead to situations where possibly important information is not being considered. Using a pattern recognition method such as PCA is also in line with the goal of removing subjective appraisal from FCEs, where pre-selected variables may be more similar to the biomechanical criteria already used by evaluators, and may not be homogenous for every individual.

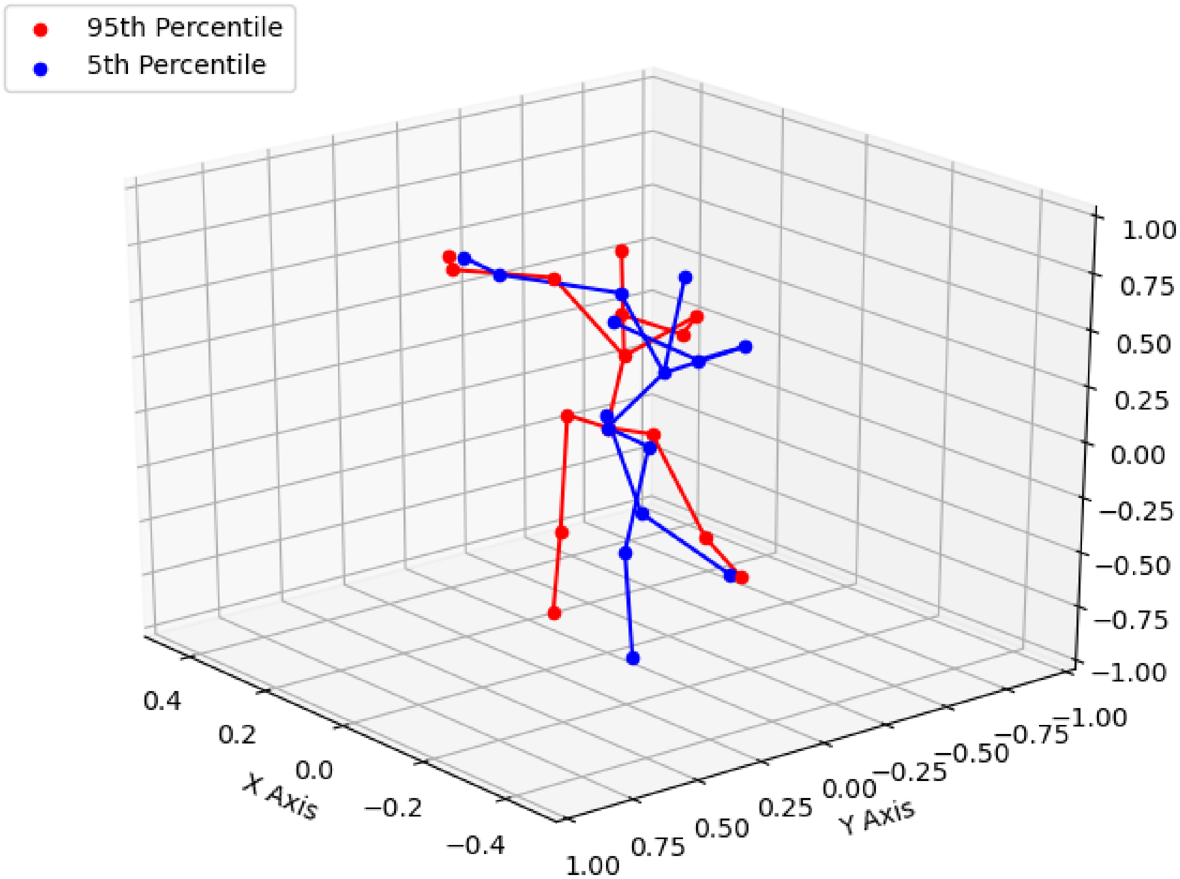

Figure 18. Example of PCA single component reconstruction of the first PC of P1, which explains 70% of their variance. One of the main differences between the 5th and 95th percentile is which foot is forward during the lift.

6.3.2 Hyperparameters

The overall performance of the OCSVM model for identifying outliers in data relies heavily on the setting of the hyperparameters *gamma* and *nu*. Improper selection of these values in the models can lead to overfitting or underfitting of data, which can affect the ability of the model to detect outliers effectively (Wang et al., 2018). If *gamma* is set too large, the training data will be overfit with a very tight decision boundary, thus the model will classify too many data points as outliers (see Figure 12(e)). If *gamma* is set too small, the data will be underfit with a very smooth and generalizable decision boundary, making it hard to detect outliers (see Figure 12(a)). Similarly, if *nu* is set too large there will be too many training set outliers (see Figure 13(d)). A small *nu* may result in a decision boundary distorted by noisy training data (see Figure 13(a)) (Wang et al., 2018). Appropriately setting these two parameters are crucial for the OSCVM performance.

Hyperparameters are typically optimized in machine learning models using an automated approach that aims to maximize the accuracy of the model (Wu et al., 2019). For example, consider if each lift in the training data being input into the OCSVM models in this thesis were already pre-classified as similar or different to the baseline lifting movement pattern used by participants. The hyperparameters would be optimized to maximize the accuracy of the model to only include lifts classified as differing from baseline as outliers from the decision boundary. However, OCSVMs are an unsupervised approach where the data do not have labels to indicate an outlier class (in this case, individual lifts cannot be reliably pre-labelled as the same or different movement patterns). Therefore, traditional hyperparameter optimization approaches such as cross-validation grid searches cannot be used (Wang et al., 2018).

A manual optimization search was instead used for this thesis, with the definition of the hyperparameters used as a guideline. Since *nu* is defined as the upper bound on the fraction of training errors, *nu* was set to 0.01 based on the assumption that about 1% of the training lifts might be classified as an outlier. This assumption was made so the model could form a decision boundary around most of the training data to define the baseline movement pattern. *Gamma* was set to the default *scaled* value that was calculated using Equation 1. Through the manual search, the values used were found to have consistent performance for the training sets (see Table 6), where an average of 16% of training lifts were classified as outliers. These hyperparameter values were also able to identify a wide range of test set outlier percentages (see Table 6). Other values (*gamma* >0, *nu* in range of 0 to 1) of these hyperparameters were examined to test model performance for both the training and test sets outlier percentage. These other values tested in the manual search would result in narrower ranges of outlier percentages, or in some cases, 100% or 0% of the test set lifts being classified as outliers (i.e. an overfit or underfit model). The manual search used to settle on the values used in this thesis may have led to overfitting or underfitting of the data, but the nature of the data and OCSVM data made it difficult to measure. However, consistency and transparency were regarded as crucial for the approach used.

The presence of similar applications of OCSVM for movement pattern outlier detection in the literature is limited in order to draw conclusions about the hyperparameter optimization approach. Kobsar and Ferber (2018) used a method where the hyperparameter value chosen was able to achieve less than 1% of outliers in a randomly selected 20% cross-validation set from the baseline data. However, they had pre-defined gait movement data where the baseline data were from a pre-exercise intervention, and the test data were from the post-exercise intervention. Having these separate time points for data collection may have allowed them to fit the decision boundary more loosely around their baseline data than in this thesis. Compared to the average 16% of training data being classified as outliers in this thesis, Kobsar and Ferber (2018) had an average of 0.5% of the 20% cross-validation set classified as outliers. Also, this thesis had a range of 13.8 to 100 median percentage of test set outliers, where the majority of participants had a median of 40% or higher (see Table 6). In contrast, Kobsar and Ferber (2018) had an average of 17.7% of outliers in their post-intervention data. These results signify that the decision boundaries in this thesis may have been overfit since there is a higher percentage of outliers in both the training and test set data. However, it is hard to determine how well the model fit the data due to the high dimensionality of the data, and the lack of ability to calculate classification accuracy.

In the future, emerging methods for OCSVM hyperparameter optimization should be investigated to examine performance. One such method is described in Wang et al. (2018), where they use a self-adaptive data shifting method. This new proposed method outperformed seven state-of-the-art OCSVM hyperparameter optimization methods. Nevertheless, hyperparameters should be optimized for each individual model and special attention should be paid to model performance based on these hyperparameters. Precise reporting of hyperparameter values or optimization methods are also crucial for reproducibility and understanding of research studies.

6.4 Methodological Limitations

The use of an outlier detection machine learning model was able to detect fatigue related changes in some participants during a repetitive lifting protocol, however, limitations should be considered. First, a sample size of 14 participants recruited from the local university population was used. Although this was a subject-specific approach, only a small sample was used which may limit the generalizability of the results. Also, the demographics of the population could limit generalizability. The main objective guiding the thesis was the improvement of FCEs, where a younger population may not be representative of the people who may undergo FCE testing. In this study, only 50% of participants had MMH experience, and only three had experience of one year or longer. Results may differ based on the age and work experience of the participants included.

Another limitation to consider are the fatigue measures used during the protocol. As mentioned in Sections 6.1.2 and 6.2, the fatigue measures used may not reliably detect whole-body fatigue. Nonetheless, RPE is a common measurement used to get an idea of training load (Roos et al., 2013), and has been shown to have associations with other physiological measurements during workplace lifting tasks (Jakobsen et al., 2014). RPE is also a common monitoring method used in FCEs, but is usually combined with other physiological methods of monitoring individual's responses to simulated work-tasks (Allison et al., 2018).

Lastly, the use of PCA as a data reduction and feature extraction method distorts the time domain of the movements since each lift was normalized to 101 frames. The OCSVM model could therefore not discriminate lifts based on any measures that may have varied with time. For example, time to complete a lift has been shown to both increase (Fischer et al., 2015) and decrease (Mehta et al., 2014) over the duration of a repetitive lifting protocol. Biomechanical variables such as joint velocities and accelerations would not be considered due to the time distortion. However, PCA was useful to account for at least 95% of the variance in participants' baseline whole-body movement patterns based on pattern recognition, creating a truly unique subject-specific feature space for each individual model.

6.5 Potential FCE Applications and Future Directions

The central objective of this thesis was to examine the use of an outlier detection machine learning algorithm as a potential alternative approach to move away from the subjective visual appraisal approach currently used by evaluators during FCEs. The results support that when individuals were fatigued, there was an associated change in their lifting movement strategy away from their baseline. However, laboratory-grade motion capture systems are not available for use in day-to-day clinical applications, like an FCE, to make tracking of whole-body kinematics possible. Therefore, alternative approaches should be considered. Similar to other applications of subject-specific approaches to detecting changes in movement (Clermont, Benson, et al., 2019; Conforti et al., 2020; Kobsar & Ferber, 2018), IMUs are a potential wearable sensor alternative to track various kinematic variables. IMUs are an affordable, easy to use sensor that could be used in conjunction with real-time monitoring systems to get quick feedback and measurements, even with potential to utilize machine learning algorithms (Graham & Josan, 2017). An additional potential approach to collecting whole-body kinematic data is through the use of marker-less motion capture and convolutional pose estimation (Cao et al., 2019; Wei et al., 2016). Again, real-time motion data could be captured and analyzed (Van Den Bogert et al., 2013) in conjunction with OCSVMs to give quick information to evaluators in FCEs to make more objective decisions.

A second future-direction is associated with the OCSVM considerations mentioned previously (see Section 6.3). Future research should explore the use of various kinematic or kinetic variables during MMH tasks for use as features in OCSVMs, where pre-selected task-relevant variables (Chan et al., 2020) may give more direct insight for FCE evaluators and patients. Although PCA is useful for extracting movement features in an unsupervised manner, interpreting the relevance of what those features may represent is less straight-forward. However, PCA is useful for the subject-specific approach described in this thesis, where some pre-selected variables may not be homogenous in every individual if that alternative approach was used. Secondly, the hyperparameter selection process should be examined further in future research. Since this is a novel study where whole-body kinematics from a lifting task were input into a OCSVM, there is not currently a basis to follow. The lack of labels of lifts as specifically fatigued or non-fatigued does not allow for the testing of model accuracy, so under-fitting and over-fitting of the model may be an issue. However, steps were taken to ensure consistency and transparency in this thesis. Using alternative approaches for hyperparameter selection should be investigated, such as the self-adaptive shifting method (Wang et al., 2018).

Lastly, the amount of repetitive lifts needed to accurately define a person's whole-body baseline lifting movement pattern should be explored. In this study, about 35% of the lifts completed in the hour-long protocol was used as the training set to define their baseline pattern. This threshold was chosen to allow for the remaining lifts in the test sets to have a potential adequate observable progression from non-fatigued to fatigued. Since this was one continuous data collection, there was no definitive moment that could be used to separate training and test lifts. For example, a defined separation would be using pre-intervention data as the training set and post intervention data as the test set (Kobsar & Ferber, 2018). Future research could utilize an approach where lifting patterns are recorded before a fatiguing activity, and then again after the fatiguing activity to have a defined separation between training and test data. Another issue with defining a baseline movement pattern is that movement is innately variable, especially during prolonged repetitive tasks like the one used in this thesis (Sedighi & Nussbaum, 2017). Knowing how many lifts are required to reliably define a baseline movement pattern would also be important in FCEs, where time may be limited. Lifting is not the only MMH task completed during an FCE, so baseline movement patterns may need to be measured for multiple tasks. Future research should investigate the influence on number of training lifts needed, and how changing the amount may influence the results from an outlier detection algorithm.

7. Conclusion

The use of an outlier detection machine learning algorithm to identify subject-specific fatigue related changes in lifting movement patterns during a prolonged repetitive protocol was examined. Spearman's rank order correlation was used to assess the level of association between participant's fatigue level and their percentage of outlier lifts in each test set as classified by the OCSVM. Results showed that for participants who were likely to be fatigued, there were significant deviations in their lifting movement pattern observed as more outlier movements from baseline, as they became more fatigued. For those who were likely not fatigued, there were fewer outliers from their baseline when associated with the fatigue measures. However, the OCSVM was able to discern changes in their movement pattern even though they may not have been fatigued.

The results from this thesis support the use of OCSVMs to detect subject-specific changes in whole-body movement when completing MMH tasks. A combination of more objective fatigue measures, and examination of different feature selection processes could further improve this method. Considerations for application of this method to the FCE space are discussed, where the combination of wearable sensors and real-time analyses could help detect fatigue related changes in movement in a clinical setting. Overall, this objective, subject-specific machine learning method may help eliminate the subjectivity of FCE to support improved RTW decision making.

References

Allison, S., Galper, J., Hoyle, D., & Mecham, J. (2018). *Current concepts in functional capacity evaluation: a best practices guideline.*

Armstrong, D. P., Budarick, A. R., Pegg, C. E. E., Graham, R. B., & Fischer, S. L. (2020). Feature detection and biomechanical analysis to objectively identify high exposure movement strategies when performing the EPIC lift capacity test. *Journal of Occupational Rehabilitation*. https://doi.org/10.1007/s10926-020-09890-2

Association of Workers' Compensation Boards of Canada (2019). 2017 Lost Time Claims in Canada. Retrieved from http://awcbc.org/?page_id=14

Bell, A. L., Brand, R. A., & Pedersen, D. R. (1989). Prediction of hip joint centre location from external landmarks. *Human Movement Science*, *8*(1), 3–16. https://doi.org/10.1016/0167-9457(89)90020-1

Bieniek, S., & Bethge, M. (2014). The reliability of WorkWell Systems functional capacity evaluation : a systematic review. *BMC Musculoskeletal Disorders*, *15*(106), 1–13.

Bonato, P., Boissy, P., Della Croce, U., & Roy, S. H. (2002). Changes in the surface EMG signal and the biomechanics of motion during a repetitive lifting task. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *10*(1), 38–47. https://doi.org/10.1109/TNSRE.2002.1021585

Bonato, P., Ebenbichler, G. R., Roy, S. H., Lehr, S., Posch, M., Kollmitzer, J., & Della Croce, U. (2003). Muscle fatigue and fatigue-related biomechanical changes during a cyclic lifting task. *Spine*, *28*(16), 1810–1820. https://doi.org/10.1097/01.BRS.0000087500.70575.45

Borg, G. (1982). Psychophysical bases of perceived exertion. *Medicine and Science in Sports and Exercise*, *14*, 377–381.

Brandon, S. C. E., Graham, R. B., Almosnino, S., Sadler, E. M., Stevenson, J. M., & Deluzio, K. J. (2013). Interpreting principal components in biomechanics: Representative extremes and single component reconstruction. *Journal of Electromyography and*

Kinesiology, *23*(6), 1304–1310. https://doi.org/10.1016/j.jelekin.2013.09.010

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*, 121–167.

Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2019). Realtime multi-person 2D pose estimation using part affinity fields. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. https://doi.org/10.1109/CVPR.2017.143

Chan, V. C. H., Beaudette, S. M., Smale, K. B., Beange, K. H. E., & Graham, R. B. (2020). A subject-specific approach to detect fatigue-related changes in spine motion using wearable sensors. *Sensors*, *20*(9), 2646.

Clermont, C. A., Benson, L. C., Edwards, W. B., Hettinga, B. A., & Ferber, R. (2019). New considerations for wearable technology data: Changes in running biomechanics during a marathon. *Journal of Applied Biomechanics*, *35*(6), 401–409. https://doi.org/10.1123/jab.2018-0453

Clermont, C. A., Phinyomark, A., Osis, S. T., & Ferber, R. (2019). Classification of higher- and lower-mileage runners based on running kinematics. *Journal of Sport and Health Science*, *8*(3), 249–257. https://doi.org/10.1016/j.jshs.2017.08.003

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Conforti, I., Mileti, I., Del Prete, Z., & Palermo, E. (2020). Measuring biomechanical risk in lifting load tasks through wearable system and machine-learning approach. *Sensors (Switzerland)*, *20*(6). https://doi.org/10.3390/s20061557

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297. https://doi.org/10.1109/64.163674

Cortes, N., Onate, J., & Morrison, S. (2014). Differential effects of fatigue on movement variability. *Gait and Posture*, *39*(3), 888–893. https://doi.org/10.1016/j.gaitpost.2013.11.020

Daffertshofer, A., Lamoth, C. J. C., Meijer, O. G., & Beek, P. J. (2004). PCA in studying coordination and variability: A tutorial. *Clinical Biomechanics, 19*(4), 415–428. https://doi.org/10.1016/j.clinbiomech.2004.01.005

De Baets, S., Calders, P., Schalley, N., Vermeulen, K., Vertriest, S., Van Peteghem, L., Coussens, M., Malfait, F., Vanderstraeten, G., Van Hove, G., & Van de Velde, D. (2018). Updating the evidence on functional capacity evaluation methods: a systematic review. *Journal of Occupational Rehabilitation*, *28*(3), 418–428. https://doi.org/10.1007/s12498-018-0238-1

Deluzio, K. J., & Astephen, J. L. (2007). Biomechanical features of gait waveform data associated with knee osteoarthritis. An application of principal component analysis. *Gait and Posture*, *25*(1), 86–93. https://doi.org/10.1016/j.gaitpost.2006.01.007

Fischer, S. L., Greene, H. P., Hampton, R. H., Cochran, M. G., & Albert, W. J. (2015). Gender-based differences in trunk and shoulder biomechanical changes caused by prolonged repetitive symmetrical lifting. *IIE Transactions on Occupational Ergonomics and Human Factors*, *3*(3–4), 165–176. https://doi.org/10.1080/21577323.2015.1034382

Fong, J., Ocampo, R., Gross, D. P., & Tavakoli, M. (2020). Intelligent robotics incorporating machine learning algorithms for improving functional capacity evaluation and occupational rehabilitation. *Journal of Occupational Rehabilitation*, *30*(3), 362–370. https://doi.org/10.1007/s10926-020-09888-w

Fukuchi, R. K., Eskofier, B. M., Duarte, M., & Ferber, R. (2011). Support vector machines for detecting age-related changes in running kinematics. *Journal of Biomechanics*, *44*(3), 540–542. https://doi.org/10.1016/j.jbiomech.2010.09.031

Geisser, M. E., Robinson, M. E., Miller, Q. L., & Bade, S. M. (2003). Psychosocial factors and functional capacity evaluation among persons with chronic pain. *Journal of Occupational Rehabilitation*, *13*(4), 259–276.

Gibson, L., & Strong, J. (2003). A conceptual framework of functional capacity evaluation for occupational therapy in work rehabilitation. *Australian Occupational Therapy*

Journal, *50*(2), 64–71. https://doi.org/10.1046/j.1440-1630.2003.00323.x

Gouttebarge, V., Wind, H., Kuijer, P. P., & Frings-Dresen, M. H. (2004). Reliability and validity of functional capacity evaluation methods: a systematic review with reference to Blankenship system, Ergos work simulator, Ergo-Kit and Isernhagen work system. *International Archives of Occupational Environmental Health*, *77*, 527–537. https://doi.org/10.1007/s00420-004-0549-7

Gouttebarge, V., Wind, H., Kuijer, P. P., Sluiter, J. K., & Frings-Dresen, M. H. (2006). Reliability and agreement of 5 Ergo-Kit functional capacity evaluation lifting test in subjects with low back pain. *Archives of Physical Medicine and Rehabilitation*, *87*(10), 1365–1370. https://doi.org/10.1016/j.apmr.2006.05.028

Graham, R. B., & Josan, G. P. K. (2017). Development of a novel wearable system for the clinical assessment of movement quality and control in low back pain. *Proceedings of the XXVI Congress of the International Society of Biomechanics, Brisbane, Australia, 23-27 July 2017*, 407.

Gross, D. (2004). Measurement properties of performance-based assessment of functional capacity. *Journal of Occupational Rehabilitation*, *14*(3), 165–174.

Gross, D. (2006). Are functional capacity evaluations affected by the patient's pain. *Current Pain and Headache Reports*, *10*, 107–113.

Gross, D., Asante, A. K., Miciak, M., Battié, M. C., Carroll, L. J., Sun, A., Mikalsky, M., Huellstrung, R., & Niemeläinen, R. (2014). A cluster randomized clinical trial comparing functional capacity evaluation and functional interviewing as components of occupational rehabilitation programs. *Journal of Occupational Rehabilitation*, *24*(4), 617–630. https://doi.org/10.1007/s10926-013-9491-4

Gross, D., & Battié, M. C. (2005). Functional capacity evaluation performance does not predict sustained return to work in claimants with chronic back pain. *Journal of Occupational Rehabilitation*, *15*(3). https://doi.org/10.1007/s10926-005-5937-7

Halilaj, E., Rajagopal, A., Fiterau, M., Hicks, J. L., Hastie, T. J., & Delp, S. L. (2018).

Machine learning in human movement biomechanics: Best practices, common pitfalls, and new opportunities. *Journal of Biomechanics*, *81*, 1–11. https://doi.org/10.1016/j.jbiomech.2018.09.009

Howarth, S. J., & Callaghan, J. P. (2010). Quantitative assessment of the accuracy for three interpolation techniques in kinematic analysis of human movement. *Computer Methods in Biomechanics and Biomedical Engineering*, *13*(6), 847–855. https://doi.org/10.1080/10255841003664701

Hsu, C.-W., & Lin, C.-J. (2002). A comparison of model selection methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, *13*(2), 415–425. https://doi.org/10.1007/11424925_119

Innes, E., & Straker, L. (1998a). A clinician's guide to work-related assessments: 1 - Purposes and problems. *Work*, *11*(2), 183–189.

Innes, E., & Straker, L. (1998b). A clinician's guide to work-related assessments: 3 - administration and interpretation problems. *Work*, *11*(2), 207–219.

Innes, E., & Straker, L. (1999). Validity of work-related assessments. *Work*, *13*(2), 125–152.

Isernhagen, S. J. (1992). Functional capacity evaluation: rationale, procedure, utility of the kinesiophysical approach. *Journal of Occupational Rehabilitation*, *2*(3), 157–168.

Jakobsen, M. D., Sundstrup, E., Persson, R., Andersen, C. H., & Andersen, L. L. (2014). Is Borg's perceived exertion scale a useful indicator of muscular and cardiovascular load in blue-collar workers with lifting tasks? A cross-sectional workplace study. *European Journal of Applied Physiology*, *114*(2), 425–434. https://doi.org/10.1007/s00421-013-2782-9

Jay, M. A., Lamb, J. M., Watson, R. L., Young, I. A., Fearon, F. J., Alday, J. M., & Tindall, A. G. (2000). Sensitivity and specificity of the indicators of sincere effort of the EPIC lift capacity test on a previously injured population. *Spine*, *25*(11), 1405–1412. https://doi.org/10.1097/00007632-200006010-00013

Kendrick, D., O'Brien, C., Christie, N., Coupland, C., Quinn, C., Avis, M., Barker, M., Barnes, J., Coffey, F., Joseph, S., Morris, A., Morriss, R., Rowley, E., Sleney, J., & Towner, E. (2011). The impact of injuries study. Multicentre study assessing physical, psychological, social and occupational functioning post injury - a protocol. *BMC Public Health*, *11*(963), 1–7. https://doi.org/10.1186/1471-2458-11-963

King, P. M., Tuckwell, N., & Barrett, T. E. (1998). A critical review of functional capacity evaluations. *Physical Therapy*, *78*(8), 852–866.

Kobsar, D., & Ferber, R. (2018). Wearable sensor data to track subject-specific movement patterns related to clinical outcomes using a machine learning approach. *Sensors (Switzerland)*, *18*(9). https://doi.org/10.3390/s18092828

Lebeau, M., & Duguay, P. (2013). The costs of occupational injuries. A review of the literature. In *IRSST*.

Matheson, L. N., Isernhagen, S. J., & Hart, D. L. (2002). Relationships among lifting ability, grip force, and return to work. *Physical Therapy*, *82*(3), 249–256. https://doi.org/10.1093/ptj/82.3.249

Matheson, L. N., Mooney, V., Grant, J. E., Affleck, M., Hall, H., Melles, T., Lichter, R. L., & McIntosh, G. (1995). A test to measure lift capacity of physically impaired adults: Part 1-development and reliability testing. *Spine*, *20*(19), 2119–2129. https://doi.org/10.1097/00007632-199510000-00010

Mehta, J. P., Lavender, S. A., & Jagacinski, R. J. (2014). Physiological and biomechanical responses to a prolonged repetitive asymmetric lifting activity. In *Ergonomics* (Vol. 57, Issue 4, pp. 575–588). Taylor & Francis. https://doi.org/10.1080/00140139.2014.887788

Morgan, M. V., Allison, S., & Duhon, D. (2012). Heart rate changes in functional capacity evaluations in a workers' compensation population. *Work*, *42*(2), 253–257. https://doi.org/10.3233/WOR-2012-1348

Mourão-Miranda, J., Hardoon, D. R., Hahn, T., Marquand, A. F., Williams, S. C. R., Shawe-Taylor, J., & Brammer, M. (2011). Patient classification as an outlier detection problem: An application of the one-class support vector machine. *NeuroImage*, *58*(3), 793–804.

https://doi.org/10.1016/j.neuroimage.2011.06.042

Nussbaum, M. A., & Zhang, X. (2000). Heuristics for locating upper extremity joint centres from a reduced set of surface markers. *Human Movement Science*, *19*(5), 797–816. https://doi.org/10.1016/S0167-9457(00)00020-8

Parachute. (2015). *The cost of injury in Canada*.

Petrofsky, J. S., & Lind, A. R. (1978). Metabolic, cardiovascular, and respiratory factors in the development of fatigue in lifting tasks. *Journal of Applied Physiology Respiratory Environmental and Exercise Physiology*, *45*(1), 64–68. https://doi.org/10.1152/jappl.1978.45.1.64

Phinyomark, A., Hettinga, B. A., Osis, S. T., & Ferber, R. (2014). Gender and age-related differences in bilateral lower extremity mechanics during treadmill running. *PLoS ONE*, *9*(8). https://doi.org/10.1371/journal.pone.0105246

Phinyomark, A., Petri, G., Ibáñez-Marcelo, E., Osis, S. T., & Ferber, R. (2018). Analysis of Big Data in Gait Biomechanics: Current Trends and Future Directions. *Journal of Medical and Biological Engineering*, *38*(2), 244–260. https://doi.org/10.1007/s40846-017-0297-2

Pransky, G. S., & Dempsey, P. G. (2004). Practical aspects of functional capacity evaluations. *Journal of Occupational Rehabilitation*, *14*(3), 217–229.

Remedios, S. M., Armstrong, D. P., Graham, R. B., & Fischer, S. L. (2020). Exploring the application of pattern recognition and machine learning for identifying movement phenotypes during deep squat and hurdle step movements. *Frontiers in Bioengineering and Biotechnology*, *8*(364), 1–15.

Reneman, M. F. (2003). Introduction to the special issue on functional capacity evaluations: from expert based to evidence based. *Journal of Occupational Rehabilitation*, *13*(4), 203–206.

Reneman, M. F., Dijkstra, P., Westmaas, M., & Göeken, L. (2002). Test-retest reliability of lifting and carrying in a 2-day functional capacity evaluation. *Journal of Occupational*

Rehabilitation, *12*(4), 269–275.

Roos, L., Taube, W., Brandt, M., Heyer, L., & Wyss, T. (2013). Monitoring of daily training load and training load responses in endurance sports: What do coaches want? *Schweizerische Zeitschrift Fur Sportmedizin Und Sporttraumatologie*, *61*(4), 30–36.

Sadler, E. M., Graham, R. B., & Stevenson, J. M. (2011). The personal lift-assist device and lifting technique: A principal component analysis. *Ergonomics*, *54*(4), 392–402. https://doi.org/10.1080/00140139.2011.556259

Scott, S. H. (2004). Optimal feedback control and the neural basis of volitional motor control. *Nature Reviews Neuroscience*, *5*(7), 532–544. https://doi.org/10.1038/nrn1427

Sedighi, A., & Nussbaum, M. A. (2017). Temporal changes in motor variability during prolonged lifting/lowering and the influence of work experience. *Journal of Electromyography and Kinesiology*, *37*(September), 61–67. https://doi.org/10.1016/j.jelekin.2017.09.001

Sedighi, A., & Nussbaum, M. A. (2019). Exploration of different classes of metrics to characterize motor variability during repetitive symmetric and asymmetric lifting tasks. *Scientific Reports*, *9*(1), 1–9. https://doi.org/10.1038/s41598-019-46297-3

Sinden, K. E., McGillivary, T. L., Chapman, E., & Fischer, S. L. (2017). Survey of kinesiologists' functional capacity evaluation practice in Canada. *Work*, *56*(4), 571–580. https://doi.org/10.3233/WOR-172519

Sparto, P. J., Parnianpour, M., Reinsel, T. E., & Simon, S. (1997). The effect of fatigue on multijoint kinematics, coordination, and postural stability during a repetitive lifting test. *Spine*, *22*(22), 2647–2654. https://doi.org/10.1097/00007632-199711150-00013

Srinivasan, D., & Mathiassen, S. E. (2012). Motor variability in occupational health and performance. *Clinical Biomechanics*, *27*(10), 979–993. https://doi.org/10.1016/j.clinbiomech.2012.08.007

Stergiou, N., & Decker, L. M. (2011). Human movement variability, nonlinear dynamics, and pathology: Is there a connection? *Human Movement Science*, *30*(5), 869–888.

https://doi.org/10.1016/j.humov.2011.06.002

Stone, M. H., O'Bryant, H. S., Hornsby, G., Cunanan, A., Mizuguchi, S., Suarez, D. G., South, M., Marsh, D., Haff, G. G., Ramsey, M. W., Beckham, G. K., Santana, H. A., Wagle, J. P., Stone, M. E., & Pierce, K. C. (2019). Using the isometric mid-thigh pull in the monitoring of weightlifters: 25+ years of experience. *Professional Strength & Conditioning*, *54*, 19–26.

Van Den Bogert, A. J., Geijtenbeek, T., Even-Zohar, O., Steenbrink, F., & Hardin, E. C. (2013). A real-time system for biomechanical analysis of human movement and muscle function. *Medical and Biological Engineering and Computing*, *51*(10), 1069–1077. https://doi.org/10.1007/s11517-013-1076-z

Wang, S., Liu, Q., Zhu, E., Porikli, F., & Yin, J. (2018). Hyperparameter selection of one-class support vector machine by self-adaptive data shifting. *Pattern Recognition*, *74*, 198–211. https://doi.org/10.1016/j.patcog.2017.09.012

Wei, S. E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional pose machines. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 4724–4732. https://doi.org/10.1109/CVPR.2016.511

Winter, D. A. (2009). *Biomechanics and Motor Control of Human Movement, Fourth Edition*. John Wiley & Sons, Inc.

Wu, J., Chen, X. Y., Zhang, H., Xiong, L. D., Lei, H., & Deng, S. H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology*, *17*(1), 26–40. https://doi.org/10.11989/JEST.1674-862X.80904120